# ACCELERATING VIOLENCE DETECTION: A COMPARATIVE STUDY OF FRAME DIFFERENCE AND TRADITIONAL METHODS ON UCF CRIME DATASET

**SANIUL ISLAM SANI[1], RASHEDUR RAHMAN[1], ZAHIDUL ISLAM[1], SIAM TAHSIN BHUIYAN[1], MAHMUDUL HAQUE[1], SEFATUL WASI [2], SAJED IMTENANUL HAQUE[2], SAADIA BINTE ALAM [1]**

[1]Center for Computational & Data Sciences, Independent University, Bangladesh, Dhaka 1299, Bangladesh
[2]Department of Computer Science and Engineering, Independent University, Bangladesh, Dhaka 1229, Bangladesh
E-MAIL: saniulislam58@gmail.com, rashed.riyadh14@gmail.com, zahid.imx@gmail.com, siamtbhuiyan@gmail.com, mahmud.eece@gmail.com, sefatulwasi@gmail.com, imtenanul@iub.edu.bd, saadiabinte@iub.edu.bd

## Abstract:

**Effective violence detection is crucial for enhancing public safety and preventing harmful events. This paper evaluates two approaches using a ResNet50-LSTM model: normal frame input and frame difference input. Data preprocessing involved using normal frames for one method and computing differences between consecutive frames for the other. Performance was assessed using ROC curves and AUROC metrics. The findings indicate that the frame difference method improves AUROC scores and reduces computational time, making it a more efficient and accurate solution for real-time violence detection.**

## Keywords:

**Violence Detection; Frame Difference; ResNet50; LSTM; UCF Crime Dataset; Deep Learning; Frame Difference; AUROC; Computational Efficiency; Video Analysis.**

## 1. Introduction

Violence detection in video surveillance is crucial for ensuring public safety, supporting law enforcement, and preventing crime. The World Health Organization (WHO) reports that violence is responsible for over 5 million deaths every year [1], creating a significant global economic burden exceeding $14.1 trillion [2]. Accurate and effective violence detection systems are essential for addressing these severe impacts by allowing for early intervention, identifying threats in real-time, and enabling proactive measures in high-risk areas. By developing and implementing these systems, we can work towards reducing violence, enhancing public safety, and creating a more secure and peaceful society.

The field of violence detection in video surveillance has gained significant attention due to rapid advancements in technology and increasing concerns about safety [3]. Traditional methods of surveillance depend heavily on human operators who manually monitor video feeds [4]. This approach is often labor-intensive and prone to errors due to operator fatigue and the enormous volume of data that needs to be reviewed. To overcome these limitations, there has been a shift towards developing automated systems that leverage advanced technology like deep learning and machine learning algorithms.

Modern violence detection systems use sophisticated tools such as convolutional neural networks (CNNs) to analyze video footage and identify violent activities in real-time [10]. CNNs are particularly effective at extracting features from images, which helps in recognizing patterns associated with violent behavior [4] [14]. In addition to CNNs, other technologies such as edge computing and cloud analytics have significantly improved the performance of these systems. Edge computing allows for the processing of video data directly on the surveillance device, reducing latency and the need for extensive data transmission. Cloud analytics, on the other hand, enables the handling of large-scale data and supports more complex analysis.

These advancements are particularly important for various settings, including public spaces like parks and malls, educational institutions such as schools, and online platforms where monitoring and preventing violent incidents are crucial. The integration of artificial intelligence (AI) and machine learning into surveillance systems provides a more effective solution compared to traditional manual monitoring. AI-driven systems can analyze video feeds much faster and with greater accuracy, which helps in swiftly responding to incidents and preventing potential violence [6] [13].

Previous research in violence detection explored several approaches. Traditional frame-based methods involved analyzing individual video frames to detect violent activities. More advanced models used deep learning techniques such as ResNet-LSTM [11], which combined a Residual Network (ResNet) for extracting features from video frames with a Long Short-Term Memory (LSTM) network for understanding temporal patterns in video sequences [7] [15]. Other methods included 3D Convolutional Neural Networks (3D CNNs) and Recurrent Neural Networks (RNNs), which analyzed video sequences to detect violent content [5] [16]. While these methods had shown promise, they also came with certain limitations. For example, they might have lacked sufficient temporal context, had high false positive rates, or were vulnerable to overfitting and adversarial attacks. These issues highlighted the need for more robust and contextaware models that could provide reliable and accurate violence detection.

One technique that has shown potential is the frame difference method. This approach involves calculating the difference between consecutive frames in a video to detect motion and identify violent activities. It is computationally light and requires minimal training data, making it suitable for real-time applications. Additionally, the frame difference method integrates well with other computer vision techniques, which enhances its effectiveness by utilizing the UCF Crime dataset, a benchmark comprising 1900 videos depicting various crime scenarios, which is often used to evaluate such methods [8]. This data set provides a diverse range of scenarios that are valuable for testing the performance of violence detection systems.

This study aims to compare the performance of the frame difference technique with traditional frame-based approaches using ResNet-LSTM, for violence detection [12]. The goal is to demonstrate that the frame difference method offers advantages in terms of accuracy and computational efficiency. By investigating this technique, we hope to find a more effective and efficient method for detecting violence in videos. This could lead to improved public safety and more effective surveillance systems.

## 2. Dataset

The UCF Crime dataset has emerged as a cornerstone in the field of violence detection, offering a comprehensive and challenging benchmark for researchers. This dataset consists of images extracted from the UCF Crime dataset, used for real-world anomaly detection. It includes every 10th frame from each video, resized to 64x64 pixels in PNG format, and covers 14 behavior classes. Introduced by [17], the UCF Crime dataset contains over 1,900 video clips from real world surveillance, categorized into normal and violent behaviors. The dataset is notable for its diversity, encompassing a wide range of scenes with varying lighting conditions, camera angles, and resolutions, which mirror the unpredictability and complexity of real-world environments [18]. This diversity makes the UCF Crime dataset an ideal testing ground for violence detection systems, as it pushes models to generalize effectively across different scenarios.

One of the key contributions of the UCF Crime dataset is its role in facilitating the development and evaluation of deep learning models tailored for violence detection. Given its realistic and varied content, the dataset has been widely adopted by researchers to benchmark the performance of advanced models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks [19][20]. These models have been used to capture both spatial and temporal features from video data, helping to improve the accuracy and reliability of violence detection systems. The UCF Crime dataset's relevance in the field is further underscored by its frequent citation in research papers, serving as a standard for comparison and validation of new techniques.

This dataset has played a crucial role in advancing research into more efficient methods of violence detection, such as frame difference techniques. By providing a rich and varied set of real-world scenarios, the UCF Crime dataset allows researchers to test the effectiveness of these techniques in reducing computational complexity while maintaining or even improving detection accuracy. This has been particularly important in the development of real-time violence detection systems, where speed and efficiency are critical. The UCF Crime dataset is a crucial benchmark for violence detection, helping researchers develop and refine accurate and efficient models. Its diverse content significantly contributes to advancing automated surveillance systems for public safety.

## 3. Method

In this study, we compare the performance of two video violence detection approaches: (1) using normal frames as input to a ResNet50-LSTM model, and (2) using frame differences as input to the same model. The goal is to evaluate the effectiveness of frame differences in reducing computational time while maintaining or improving accuracy and AUROC.

### 3.1. Data Preprocessing

Two different approaches to data preprocessing were employed in this study:

- Normal Frames: For the normal frames approach, we extract 20 frames from each video at regular intervals. Specifically, we divide the length of the video by 20 to determine the time-period between two frames. For example, if a video is 10 seconds long, we extract frames at 0.5-second intervals.

- Frame Differences: We extract 20 frame differences from each video for the frame differences approach. We first extract 21 frames at the same intervals as the normal frames approach. Then, we calculate the absolute difference between two consecutive frames to obtain a frame difference, as in (1).

$$\Delta F = F(t) \sim F(t1) \qquad (1)$$

Where F(t) is the current frame and F(t1) is the previous frame. The resulting frame differences are then used as input to the model. The use of frame differences reduces the amount of redundant information and emphasizes the motion dynamics within the video, which are critical for violence detection [21].

We resize each frame to a resolution of 244x244 pixels to match the input requirements of the ResNet model [22]. Each frame was then normalized by scaling pixel values to the [0, 1] range.

We evaluate the performance of our proposed approach using Receiver Operating Characteristic (ROC) curves and Area Under the ROC Curve (AUROC) metrics. “Fig. 2”

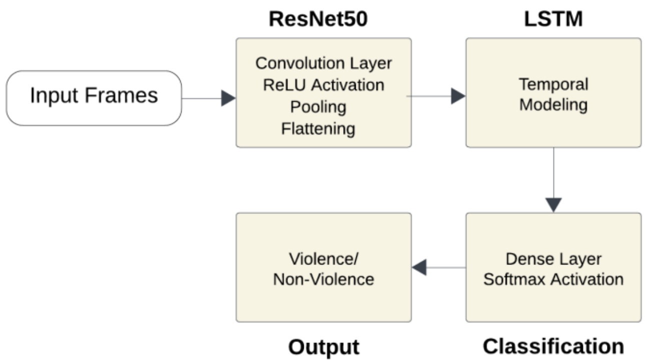


**FIGURE 1.** ResNet50-LSTM Model

### 3.2. Model Architecture

The architecture used for this study is a ResNet50-LSTM model to extract spatial features from frames or frame differences. This is achieved using convolutional blocks, each consisting of a convolutional layer, a batch normalization layer, and a ReLU activation function [23]. These blocks can capture intricate spatial patterns within the input data.

The output from the convolutional blocks is then fed into a max-pooling layer, which reduces the spatial dimensions of the data while retaining the most important features [24]. Following this, a flattened layer is used to convert the multidimensional feature maps into a one-dimensional format.

The resulting features are then fed into an LSTM layer, which models the temporal dependencies between frames. This is critical for capturing the sequence dynamics that are essential for violence detection. Finally, the output from the LSTM layer is passed through a dense layer with a SoftMax activation function, producing the classification results that distinguish between violent and non-violent actions [25].

### 3.3. Training and Evaluation

We train both models using the same hyperparameters and training protocol. We use a batch size of 16, a learning rate of 0.0001, and train for 50 epochs. The Adam optimizer is employed to optimize the model parameters due to its ability to adapt the learning rate for each parameter based on the magnitude of the gradient, which helps to stabilize the training process and improve convergence [26][27]. We evaluate the model’s using accuracy, ROC, The area under the receiver operating characteristic curve (AUROC).

### 3.4. Computational Time Measurement

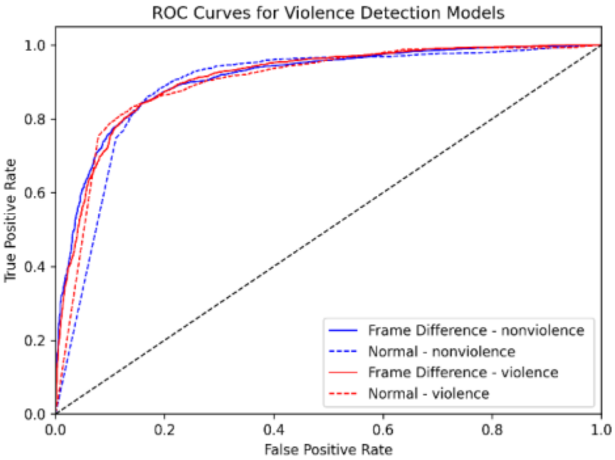
We measure the computational time required for model training and testing using both approaches. We use a NVIDIA GeForce RTX 2060 GPU with 6 GB of DDR6 memory and an Intel Core i5-10400F CPU with 16 GB of RAM.

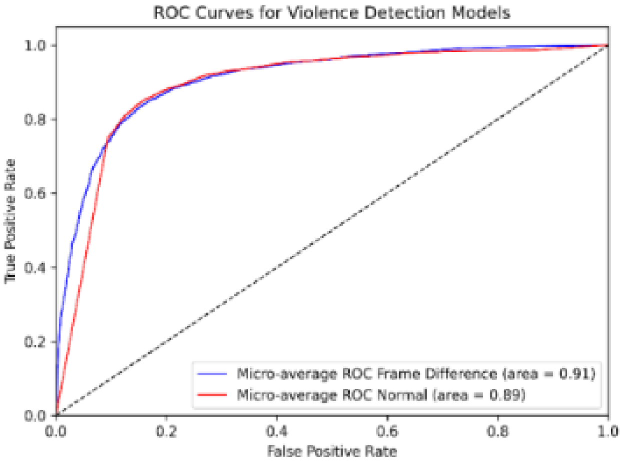
## 4. Result and Discussion

In this section, we present the results of our experiments and discuss the implications of our findings.

### 4.1. Training and Evaluation

This section shows the ROC curves for the violence and non-violence classes using normal frames and frame difference, respectively. The results indicate that the frame difference approach is better than the normal frame approach in terms of AUROC, with values of 0.9091 and 0.8892, respectively, for the violence class, and 0.9091 and 0.8988, respectively, for the non-violence class. These results suggest that the frame difference approach is more effective in distinguishing between violent and non-violent scenes, which is a critical requirement for violence detection systems. The improved performance of the frame difference approach can be attributed to its ability to capture more relevant information about the scene, such as motion and changes in the environment, which are crucial for violence detection. The frame difference approach can extract more discriminative features from the video frames, which enables it to better distinguish between violent and non-violent scenes. Fig 1.1 shows the micro-average ROC curves for normal frames and frame differences. The micro-average AUROC values are 0.89 and 0.91, respectively, further supporting the superiority of the frame difference approach. In addition to the AUROC metric, we also evaluate the accuracy of our approach. The results show that the accuracy remains the same for both normal frames and frame difference, with a value of 85%., but the frame difference approach can do so with a higher degree of confidence, as reflected in the AUROC values. [28][29].





### 4.2. Training and Evaluation

In addition to performance, we also evaluate the computational efficiency of our approach. “Table I.” shows the training and testing times for both normal frames and frame differences. The results indicate that the frame difference approach requires less time for both training and testing, with a reduction of approximately 2.9% and 4.2%, respectively. This reduction in computational time is significant, as it makes the frame difference approach more suitable for real-time applications, where computational efficiency is critical [30].

**TABLE 1.** Time Computation Table

| Approaches | *Training Time/epoch (s)* | *Inference Time On Testset (s)* |
|---|---|---|
| Frame Diff. | 361.4 | 23 |
| Normal Frame | 372.2 | 24 |

The improved computational efficiency of the frame difference approach can be attributed to its ability to reduce the dimensionality of the feature space, which reduces the computational complexity of the approach. The frame difference approach can extract more relevant features from the video frames, which enables it to achieve better performance with fewer computational resources.

### 4.3. Discussion

The results of our experiments demonstrate the effectiveness of the frame difference approach for violence detection in videos. The approach can achieve better performance than the normal frame approach while requiring less computational time. These results have significant implications for the development of violence detection systems, as they suggest that the frame difference approach is a more effective and efficient approach for violence detection. The frame difference approach is particularly well-suited for real-time applications, where computational efficiency is critical. The approach can achieve better performance than the normal frame approach, while requiring less computational time, making it more suitable for real-time applications.

## 5. Conclusion

In this study, we investigated the effectiveness of using frame differences as input to a ResNet50- LSTM model for violence detection in videos. Our findings suggest that the frame difference approach is a more effective and efficient method for violence detection in videos. The improved performance of the frame difference approach can be attributed to its ability to capture more relevant information about the scene, such as motion and changes in the environment, which are crucial for violence detection. The approach is particularly well-suited for real-time applications, where computational efficiency is critical. Overall, our study demonstrates the potential of using frame differences as input to a ResNet50- LSTM model for violence detection in videos. The approach offers a promising solution for developing more accurate and efficient systems for detecting abnormal and violent behavior, including violence detection, which can have significant implications for various applications, such as surveillance, law enforcement, public safety, and security.

## 6. Future Work

In future work, we plan to explore the applicability of our approach to various datasets, experiment with different deep learning models and hyperparameters, and investigate the use of multi-head attention mechanisms and transfer learning techniques to further improve the performance and robustness of violence detection systems.

## Acknowledgements

This paper is partially supported by a grant from the Independent University, Bangladesh (IUB).

## References

[1] World Health Organization, Injuries and violence: the facts 2014, World Health Organization, Geneva, 2014.

[2] Institute for Economics & Peace, Global Peace Index 2019, Institute for Economics & Peace, Sydney, 2019.

[3] X. Zhou, D. Wang, and X. Zheng, “Object detection with deep learning: A review,” IEEE Transactions on Neural Networks and Learning Systems, Vol. 29, No. 7, pp. 3218–3230, Jul. 2019.

[4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, Sep. 2014.

[5] K. Hara, Z. Qiu, and S. Yan, “Learning spatiotemporal features with 3D residual networks for action recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, No. 3, pp. 751–764, Mar. 2020.

[6] W. Liu et al., “SSD: Single shot multibox detector,” Proceeding of ECCV 2016, Amsterdam, pp. 21–37, Oct. 2016.

[7] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, Vol. 9, No. 8, pp. 1735–1780, Nov. 1997.

[8] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” Proceeding of CVPR 2017, Honolulu, pp. 6299–6308, Jul. 2017.

[9] Y. Zhao and Q. Tian, “A survey of deep learning-based object detection,” IEEE Access, Vol. 7, pp. 155–167, Jan. 2019.

[10] Y. Myagmar-Ochir and W. Kim, “A Survey of Video Surveillance Systems in Smart City,” Electronics, Vol. 12, No. 17, p. 3567, Sep. 2023.

[11] J. Iqbal, M. Iqbal, I. Ahmad, and M. O. Alassafi, “Real-Time Surveillance Using Deep Learning,” Security and Communication Networks, Vol. 2021, pp. 1–17, Jan. 2021.

[12] I. Mugunga et al., “A Frame-Based Feature Model for Violence Detection from Surveillance Cameras Using ConvLSTM Network,” Proceeding of ICSPCS 2021, pp. 1–7, Dec. 2021.

[13] X. Wang and Y. Qiao, “Temporal segment networks: Towards good practices for deep action recognition,” Proceeding of ECCV 2016, Amsterdam, pp. 20–36, Oct. 2016.

[14] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, “Temporal segment networks: Towards good practices for deep action recognition,” Proceeding of ECCV 2016, Amsterdam, pp. 20–36, Oct. 2016.

[15] S. Sharma et al., “A fully integrated violence detection system using CNN and LSTM,” International Journal of Electrical and Computer Engineering, Vol. 11, No. 4, pp. 3374–3380, Aug. 2021.

[16] N. Dündar, A. S. Keçeli, A. Kaya, and H. Sever, “A shallow 3D convolutional neural network for violence detection in videos,” Egyptian Informatics Journal, Vol. 26, p. 100455, Mar. 2024.

[17] J. Amin et al., “Detection of anomaly in surveillance videos using quantum convolutional neural networks,” Image and Vision Computing, Vol. 135, p. 104710, Jun. 2023.

[18] M. Gadelkarim, M. Khodier, and W. Gomaa, “Violence Detection and Recognition from Diverse Video Sources,” Proceeding of IEEE IJCNN 2022, pp. 1–6, Jul. 2022.

[19] S. Vosta and K.-C. Yow, “A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras,” Applied Sciences, Vol. 12, No. 3, p. 1021, Feb. 2022.

[20] S. Sharma, B. Sudharsan, S. Naraharisetti, and V. Trehan, “A fully integrated violence detection system using CNN and LSTM,” International Journal of Electrical and Computer Engineering, Vol. 11, No. 4, pp. 3374–3380, Aug. 2021.

[21] S. Vosta and K.-C. Yow, “KianNet: A violence detection model using an attention-based CNN-LSTM structure,” IEEE Access, vol. 99, pp. 1–1, Jan. 2023.

[22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” arXiv preprint arXiv:1512.03385, Dec. 2015.

[23] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” Proceeding of ICML 2015, Lille, pp. 448–456, Jul. 2015.

[24] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, Cambridge, MA, 2016.

[25] H. Wang and C. Schmid, “Action recognition with improved trajectories,” Proceeding of ICCV 2013, Sydney, pp. 3169–3176, Dec. 2013.

[26] S. J. Reddi, S. Kale, and S. Kumar, “On the Convergence of Adam and Beyond,” arXiv preprint arXiv:1904.09237, Apr. 2019.

[27] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv preprint arXiv:1412.6980, Dec. 2014.

[28] F. Melo, “Area under the ROC Curve,” in Encyclopedia of Systems Biology, Springer, New York, 2013, pp. 38–39.

[29] F. S. Nahm, “Receiver operating characteristic curve: overview and practical use for clinicians,” Korean Journal of Anesthesiology, Vol. 75, No. 1, pp. 25–36, Jan. 2022.

[30] M. Patel, “Real-Time Violence Detection Using CNN-LSTM,” arXiv preprint arXiv:2107.07578, Jul. 2021.