International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:3048-7722
Volume 4, Issue 4, July-Aug-2026, PP: 01-5

A Deep Learning-Based Framework for Real-Time Violence Detection in Surveillance Videos

Dr.B.Nageshwar Rao[1], K. Sushmitha[2]
[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana –, India.
[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana –, India.

Abstract – **The increasing demand for intelligent surveillance systems has accelerated the adoption of artificial intelligence and deep learning technologies for improving public safety and crime prevention. monitoring, making them susceptible to delayed responses, operator fatigue, and human error, particularly in crowded public environments. Detecting violent incidents in real time remains a significant challenge because aggressive activities often occur unexpectedly and require immediate intervention. automated video analysis systems capable of identifying complex human activities with high accuracy, thereby supporting proactive surveillance and timely emergency response. Lightweight convolutional neural network architectures and temporal sequence learning models have further in real-world surveillance environments. This study presents an intelligent videos by integrating MobileNet and The proposed framework is evaluated using standard classification metrics including Accuracy, Precision, Recall, and F1-Score. Experimental results demonstrate that the integrated MobileNet–BiLSTM model achieves an overall classification accuracy of approximately 96%, with balanced precision and recall for both violent and non-violent activity recognition. The lightweight architecture significantly reduces computational complexity while maintaining high prediction performance, making the framework suitable for deployment on resource-constrained surveillance devices. The developed system enables rapid identification of violent incidents, supports faster emergency response by law enforcement agencies, and provides a scalable solution for intelligent public safety monitoring in transportation hubs, educational institutions, commercial facilities, and other high-risk public environments.**

Keywords - **Deep Learning, Violence Detection, Surveillance Video Analysis, Real-Time Monitoring, Computer Vision, Video Classification**

I. INTRODUCTION

The rapid growth of urban populations, public transportation systems, educational institutions, commercial complexes, and smart cities cameras are now extensively deployed across public and private environments, the majority of existing systems still depend on continuous human observation. Manual monitoring becomes increasingly ineffective when thousands of video streams must be observed simultaneously, often resulting in delayed responses, missed incidents, and operator fatigue. Consequently, there is a growing need for automated surveillance systems that can continuously analyze video streams and immediately recognize violent activities

transformed video surveillance by enabling computers to automatically understand complex human actions from visual data. Deep learning models have demonstrated remarkable success in image classification, object detection, action recognition, and activity analysis by learning discriminative features directly from large-scale datasets.

Recognizing violent activities from surveillance videos presents several challenges because human actions evolve over time and often vary across different cultural, environmental, and social contexts. Factors such as crowd density, lighting conditions, camera viewpoints, occlusions, and background motion further increase the complexity of real-time violence detection. Models trained on limited or region-specific datasets frequently struggle to generalize when deployed in different environments. Therefore, developing a robust framework capable of accurately recognizing violence across diverse real-world scenarios remains an active research challenge.

II. LITERATURE SURVEY

The development of intelligent surveillance technologies has encouraged extensive research into automated violence detection systems capable of recognizing aggressive human behavior from video data. Early surveillance approaches relied primarily on manual observation or conventional computer vision techniques based on handcrafted features such as provided acceptable performance under controlled conditions, they were often sensitive to environmental variations, camera movement, illumination changes, and complex background scenes. As surveillance environments became increasingly dynamic, researchers began adopting deep learning techniques that automatically learn discriminative visual features directly from large-scale datasets.

violence detection because of their ability to extract meaningful spatial information from video frames. Several studies have demonstrated that CNN-based capable of distinguishing violent and non-violent activities. More recently, lightweight architectures such as MobileNet have gained significant attention due to their reduced computational requirements, enabling efficient deployment on mobile devices, embedded platforms, and real-time surveillance systems without substantial loss of prediction accuracy.

While convolutional neural networks effectively capture spatial characteristics, recognizing violent behavior also requires understanding temporal relationships among consecutive video frames. To address this challenge, researchers have incorporated(BiLSTM) networks, into violence detection frameworks. These sequence learning models analyze motion evolution over time and improve recognition of complex human activities by preserving contextual information across video sequences. Bidirectional learning further enhances prediction capability by simultaneously utilizing both past and future temporal dependencies, leading to more accurate activity classification.

Recent investigations have increasingly exploredconsistently report that hybrid CNN–LSTM frameworks outperform individual models in recognizing violent events across challenging datasets containing varying viewpoints, crowd densities, and environmental conditions.

Despite these advances, several research challenges remain unresolved. Many existing violence detection models are trained using relatively small datasets collected from limited geographical or cultural environments, reducing their ability to generalize across diverse real-world scenarios. Furthermore, several approaches prioritize prediction accuracy without considering computational efficiency, making them less suitable for real-time surveillance deployment. The proposed study addresses these limitations by integrating a lightweight MobileNet architecture with BiLSTM temporal learning while utilizing a more diverse training dataset that includes public gatherings and culturally representative scenarios. This combination improves both computational efficiency and recognition performance, providing a practical framework for intelligent violence detection in modern surveillance systems.

III. PROPOSED METHODOLOGY

The proposed framework introduces an AI-enabled violence detection and surveillance system that automatically identifies violent incidents from video streams using deep learning techniques. the proposed solution continuously analyzes video frames and generates instant alerts whenever violent behaviour is recognized. , the framework provides an intelligent mechanism for recognizing violent activities system consists of six major phases: video acquisition, preprocessing, feature extraction, temporal learning, violence classification, and real-time alert generation.

The process begins with video acquisition, where surveillance cameras continuously capture live video streams from public transportation hubs, and other crowded locations. The acquired videos serve as the primary input to the detection framework. To improve model robustness, the training process utilizes multiple datasets, including the AIRTLab dataset, Hockey Fight dataset, and additional video samples representing Indian public gatherings and culturally diverse scenarios. This combination situations.

Following video acquisition, the input data undergo preprocessing to prepare the frames for deep learning analysis. Video streams are decomposed into individual frames, resized to a standard resolution, normalized, and subjected to noise reduction techniques. These preprocessing operations improve visual consistency while eliminating irrelevant background information that could negatively influence model performance. The processed frames are then organized into sequential batches suitable for temporal learning.

The next stage performs spatial feature extraction using the MobileNet architecture. MobileNet efficiently learns discriminative visual representations from individual frames while maintaining a lightweight computational structure. Compared with conventional convolutional neural networks, The extracted feature maps provide detailed information regarding human posture, movement, object interactions, and environmental characteristics.

The generated spatia This bidirectional learning strategy enables the model to better understand motion continuity and behavioural evolution, thereby improving its ability to distinguish violent activities from normal human interactions. The integrated MobileNet–BiLSTM architecture combines spatial and temporal information to produce highly reliable violence predictions.

Finally, the trained model is deployed through a Django-based web application that supports real-time surveillance analysis. Users can upload recorded videos or analyze live camera streams directly through the application interface. Once violence is detected, the system immediately classifies the activity and displays the prediction result, enabling rapid intervention by security personnel or law enforcement agencies. , public safety monitoring, and automated violence detection in real-world environments.

IV. METHODOLOGY

computer vision, and temporal sequence learning to develop an automated violence detection framework capable of analyzing surveillance videos in real time. The methodology consists of six sequential stages: dataset preparation, data preprocessing, feature extraction, temporal sequence modeling, model training, and real-time violence prediction. This structured workflow enables the framework to accurately distinguish violent activities from normal human behaviour while maintaining computational efficiency suitable for practical surveillance applications.

The first stage involves dataset preparation, where video samples representing both experimental dataset includes the AIRTLab dataset, the Hockey Fight dataset, and additional videos recorded from public gatherings and culturally representative Indian scenarios. By incorporating diverse environmental conditions, crowd behaviours, and social interactionst and performance evaluation.

During the data preprocessing stage, each video is converted into individual image frames suitable for deep learning analysis. visual artifacts. Frame standardization ensures consistent input quality while reducing computational complexity during feature extraction. The processed frames are then organized into sequential frame windows that preserve temporal continuity throughout each video sequence.

Spatial information is extracted using the MobileNet convolutional neural network. MobileNet automatically learns meaningful visual representations related to human posture, object interactions, body movement, and scene characteristics while requiring significantly fewer computational resources than conventional deep convolutional architectures. Its lightweight design enables efficient processing of surveillance videos without sacrificing prediction performance, complex behavioural transitions associated with violent activities. This bidirectional temporal modeling significantly improves activity recognition by considering contextual information before and after each frame, thereby reducing misclassification of visually similar actions.

After training, the optimized MobileNet–BiLSTM model is integrated into a Django-based web application that enables real-time violence detection through an intuitive user interface. Live surveillance streams or uploaded video files

V. EXPERIMENTAL RESULTS AND DISCUSSION

MobileNet–BiLSTM framework was experimentally evaluated to determine its effectiveness in identifying violent activities from surveillance videos under diverse environmental conditions. The evaluation considered both recognition accuracy and computational efficiency, with the objective of developing a reliable solution for real-time surveillance applications. Experiments were performed using a dataset consisting of Hockey Fight dataset, and additional videos representing Indian public gatherings and cultural events. The inclusion of multiple data sources enabled the framework to learn a wider variety of human interactions and improve its adaptability to practical surveillance environments.

Prior to model training, all video samples were preprocessed through frame extraction,representative spatial and temporal patterns while maintaining an independent dataset for performance evaluation. MobileNet was employed to extract discriminative visual features from each video frame, whereas the BiLSTM network analyzed temporal dependencies between consecutive frames to improve recognition of complex human activities.

The trained model was its ability to reliably distinguish violent activities from normal human behavior. The classification report further shows that both violent and non-violent classes achieved balanced Precision, Recall, and F1-Score values of approximately 0.96, confirming that the model performs consistently across different activity categories while minimizing false predictions. These results indicate that combining MobileNet with BiLSTM successfully captures both spatial characteristics and temporal motion patterns required for accurate violence recognition.

An important observation from the experimental analysis is the benefit of incorporating culturally representative video samples into the training dataset. Existing violence detection models often experience performance degradation when deployed in The trained model was subsequently integrated into a Django-based web application for real-time surveillance analysis. Users can upload recorded surveillance videos or analyze live camera feeds through the application interface, where the system automatically classifies each video as violent or non-violent. The lightweight MobileNet architecture enables efficient processing with relatively low computational requirements, while BiLSTM effectively models temporal information without introducing excessive processing overhead. This combination allows the framework to deliver rapid predictions suitable for practical surveillance systems deployed in public transportation, educational institutions, commercial facilities, and other security-sensitive environments.

Overall, the experimental findings confirm that the proposed MobileNet–BiLSTM architecture provides l feature extraction, bidirectional temporal sequence learning, and culturally diverse training data significantly improves detection performance while supporting real-time surveillance applications. These results demonstrate the potential of the proposed framework to assist security personnel and law enforcement agencies in rapidly identifying violent incidents and responding more effectively to threats in public spaces.

VI. FUTURE ENHANCEMENT

and practical deployment. Future research may focus on expanding the training dataset by incorporating surveillance videos from different countries, weather conditions, lighting environments, camera viewpoints, and public events. A larger and more diverse dataset would improve the model's generalization capability and reduce false detections in previously unseen environments.

Another promising direction involves integrating advanced deep learning architectures such as Vision Transformodels. These architectures can capture more complex spatial and temporal relationships than conventional recurrent networks, potentially improving recognition accuracy for subtle violent behaviors and crowded surveillance scenes. Comparative analysis with MobileNet–BiLSTM may identify more effective architectures for large-scale intelligent surveillance.

Future implementations may also incorporate real-time alert generation by automatically notifying nearby security personnel or law enforcement agencies whenever violent activity is detected. Integration with emergency response systems, CCTV networks, mobile applications, and cloud-based monitoring platforms would enable faster intervention and significantly reduce response time during critical incidents.

sensors, and edge computing technologies. Deploying lightweight deep learning models directly on edge devices would reduce network latency while enabling continuous surveillance without requiring high-bandwidth cloud communication. Explainable Artificial Intelligence (XAI) techniques may also be incorporated to provide visual explanations for detected violent events, thereby improving system transparency and assisting security operators in validating automated decisions.

Finally, future research may investigate multi-modal violence detection by integrating video analysis with audio signals, environmental sensors, and behavioral analytics. Combining multiple information sources has the potential to improve recognition reliability, reduce false alarms, and create intelligent surveillance systems capable of operating effectively in complex real-world environments. These enhancements would contribute toward developing next-generation AI-powered public safety systems that provide accurate, scalable, and real-time violence detection across diverse surveillance applications.

VII. CONCLUSION

automatically identify violent activities from surveillance videos using the complementary strengths of MobileNet and efficient spatial feature extraction with temporal sequence learning to accurately distinguish violent incidents from normal human activities. By integrating lightweight deep learning models with real-time video analysis, the system provides a practical solution for automated surveillance while reducing dependence on continuous manual monitoring.

The framework Another significant contribution of this work is the implementation of the trained model within a Django-based web application, enabling real-time analysis of uploaded videos and live surveillance streams. The lightweight MobileNet architecture minimizes computational requirements, while the BiLSTM network effectively models temporal relationships between consecutive frames, allowing the framework to deliver accurate predictions without excessive processing overhead. These characteristics make the proposed system suitable for deployment in surveillance environments such as transportation hubs, educational institutions, commercial complexes, and other public spaces where rapid incident detection is essential.

In conclusion, the proposed MobileNet–BiLSTM framework demonstrates that combining lightweight convolutional networks with bidirectional temporal learning provides an effective approach for intelligent violence detection. The . Furthermore,s

REFERENCES

1. M. Divya, G. S. Priya, R. Abitha, K. Sirisha, A. Manikanta, and K. Jayanth, “Automated crime intention detection using deep learning,” International Research Journal of Modernization in Engineering, Technology and Science, vol. 4, no. 6, pp. 5670–5676, 2022.
2. A. O. Hashi, A. A. Abdirahman, M. A. Elmi, and O. E. R. Rodriguez, “Deep learning models for crime intention detection using object detection,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 4, 2023.
3. M. Patel, “Real-time violence detection using CNN-LSTM,” arXiv preprint arXiv:2107.07578, 2021.
4. S. Baloch, S. A. R. Syed Abu Bakar, M. M. Mokji, and S. Waseem, “Violence detection using affective features from upper body expressions,” in Proc. 15th Int. Conf. Digital Image Processing, 2023, pp. 1–6.
5. M. Baba, V. Gui, C. Cernazanu, and D. Pescaru, “A sensor network approach for violence detection in smart cities using deep learning,” Sensors, vol. 19, no. 7, Art. no. 1676, 2019.
   P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, “Deep learning for automatic violence detection: Tests on the AIRTLab dataset,” IEEE Access, vol. 9, pp. 160580–160595, 2021.
6. V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, “Crime prediction using machine learning and deep learning: A systematic review and future directions,” IEEE Access, 2023.
7. A. Datta, M. Shah, and N. D. V. Lobo, “Person-on-person violence detection in video data,” in Proc. Int. Conf. Pattern Recognition, 2002, pp. 433–438.

8. S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in Proc. IEEE Int. Conf. Advanced Video and Signal-Based Surveillance, 2017.
9. J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2625–2634.
10. T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in Proc. IEEE CVPR Workshops, 2012.
11. S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 2, pp. 288–303, 2010.
12. V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Human activity recognition using a dynamic texture-based method," in Proc. British Machine Vision Conference (BMVC), 2008.
13. M. Sharma and R. Baghel, "Video surveillance for violence detection using deep learning," in Advances in Data Science and Management, Springer, 2020, pp. 411–420.
14. S. U. 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. 22, Art. no. 4963, 2019.
15. T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, "Cloud-assisted multiview video summarization using CNN and bidirectional LSTM," IEEE Transactions on Industrial Informatics, vol. 16, no. 1, pp. 77–86, 2020.
16. A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
17. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
18. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
19. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
20. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learning Representations (ICLR), 2015.
21. D. P. Ramesh Babu and P. Krishna Rao, "Artificial intelligence-based intelligent surveillance and violence detection using deep learning," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 5, pp. 3568–3581, 2024.