International Journal of Leading Research Publication (IJLRP)
E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

Autonomous Detection of Unusual Incidents in CCTV Footage

Dr. Sri Sudha Garugu[1], Vempati Chekri[2], M Sravani[3], Parnati Nanda Kishore[4]

[1]Associate Professor, [2, 3, 4]Student
[1, 2, 3, 4]CSE Department, ACE Engineering College, Hyderabad, Telangana

Abstract

Hearing about the violent conditioning that do on a diurnal base around the world is relatively inviting. particular safety and social stability are seriously hovered by the violent conditioning. A variety of styles have been tried to check the violent conditioning which includes installing of surveillance systems. It'll be of great significance if the surveillance systems can automatically descry violent conditioning and give warning or alert signals. The whole system can be enforced with a sequence of procedures. originally, the system has to identify the presence of mortal beings in a videotape frame. also, the frames which are prognosticated to contain violent conditioning has to be uprooted. The in-applicable frames are to be dropped at this stage. Eventually, the trained model detects violence and these frames are independently saved as images. These images are enhanced to descry faces of people involved in the exertion, if possible. The enhanced images along with other necessary details similar as time and position is transferred as an alert to the concerned authority. The proposed system is a deep literacy grounded automatic discovery approach that uses Convolutional Neural Network to descry violence present in a videotape. But, the disadvantage of using just CNN is that, it requires a lot of time for calculation and is less accurate. Hence, a pre-trained model, MobileNetV2, which provides advanced delicacy and acts as a starting point for the structure of the entire model. An alert communication is given to the concerned authorities using telegram operation.

Keywords: Machine Learning, Convolutional Neural Networks, MobileNetV2

1. INTRODUCTION

Violence detection in video surveillance has become a major area of interest, fueled by the growing global need for improved public security and safety. Conventional surveillance systems, which rely considerably on human observation, are actually circumscribed by limitations such as fatigue, distraction, and inefficiency. Human observers tend to tire easily of keeping constant watch over several video streams, particularly during extended shifts, and hence suffer a high chance of missing significant events. Consequently, these constraints have driven researchers and technologists to consider automated methods that will complement or even substitute manual monitoring duties with smart systems.Recent developments in the field of artificial intelligence, especially in the areas of deep learning and computer vision, have made it possible to create automated systems for violence detection that can monitor video feeds in real-time. In these, MobilenetV2, convolutional neural networks (CNNs) have emerged as an

essential tool for extracting spatial features from frames of video. Such networks are able to recognize faint visual signals like hostile body positions, rapid motion, or contact that tend to lead up to or accompany violent behavior. Spatial analysis, though, is not enough because violent actions are not merely defined by how they present themselves in one frame but also by how they develop through time.[11]

To cater to the temporal aspect of violent behavior, models such as recurrent neural networks (RNNs), and even more specifically long short-term memory (LSTM) networks, have been utilized. These networks are particularly good at capturing the sequence and sequence of events over multiple frames of video, enabling the system to recognize the context and continuity of movement. This temporal analysis is essential in discriminating between benign physical contact and true violence.[1]

Increased performance has also been attained with motion-based models such as 3D CNNs and two-stream networks that process both spatial and motion data. These models not only examine the appearance of objects but also examine the movements, which makes them especially strong at picking up on rapid, chaotic, or forceful motions that are all too often linked to violence. Other methods also involve audio signals and other sensor inputs, forming multimodal systems that may integrate visual and auditory signals to produce more accurate and context-rich detection.[12]

A major issue in using these models in real-world applications is the need for real-time processing.By performing data processing on edge devices locally—like smart cameras or embedded processors—systems are able to provide low-latency response without the necessity of constant interaction with cloud servers. This not only shortens the response time for urgent alarms but also conserves bandwidth and increases the system's dependability in bandwidth-constrained locations.

Another challenge is the lack of labelled data sets with violent events, for both ethical and pragmatic reasons. To get around this, scientists have resorted to synthetic data generation and data augmentation methods. Synthetic data sets can be generated by simulation environments or generative models, whereas real-world data sets can be augmented by transformations including rotation, scaling, and adding noise. These methods assist in enhancing the generalizability of the model and how well it performs in different environments.[12]

In spite of the technology developments, the interpretability of such AI systems is still a critical issue. It is important for stakeholders like law enforcement, security agents, and the general population to know how and why a system marked aspecific event violent. Methods such as saliency maps, Grad-CAM, and explainable AI (XAI) frameworks are employed to visualize the neural network's decision-making process, providing more transparency and trust in automated systems. Concurrently, ethical aspects such as data privacy, algorithmic bias, and the ramifications of false positives also need to be addressed with care. It is important to develop systems that are not only technically good but also ethical and legal compliant.The uses of such violence detection systems are numerous. From surveillance of busy public areas such as train stations, airports, and stadiums to school, workplace, and residential complex safety, these systems have the potential to be a game-changer for contemporary surveillance. In the future, there is much room for innovation in these technologies by making them lightweight for use in low-power

devices, using privacy-sustaining methods like federated learning, and adding contextual awareness to differentiate more accurately between actual threats and non-violent situations.[13]

2. LITERATURESURVEY

Real-Time Violence Detection: This study explores deep learning models similar as ResNet50 and InceptionV3 for real- time violence discovery. By combining CNNs for spatial point birth and LSTMs for temporal literacy, the system achieves 93% accuracy and a processing speed of 131 FPS. still, farther advancements are demanded for edge computing deployment.[1]

Violence Detection System: This research utilizes 3D Convolutional Neural Networks (3D CNNs) with transfer literacy for feting violent conduct in videos. The model is robust against video frame quality, resolution, and lighting variations, offering a scalable result for surveillance operations. nevertheless, real- time perpetration challenges remain.[2]

A YOLO-Based Violence Detection Method in IoT Surveillance Systems: This paper proposes an IoT- integrated YOLO- grounded deep learning model for violence discovery. The system efficiently processes surveillance footage in real- time, making it suitable for resource- constrained edge bias. Despite its effectiveness, dataset limitations and conception issues persist.[3]

Enhancing Human Action Recognition and Violence Detection Through Deep Learning Audiovisual Fusion: This work integrates audio and video modalities using Hybrid Fusion- Based Deep learning(HFBDL) to ameliorate violence detection delicacy. The proposed system achieves 96.67% accuracy on the RLVS dataset. still, challenges in handling background noise in audio inputs remain.[4]

JOSENet: A Joint Stream Embedding Network for Violence Detection: This study introduces self-supervised learning (SSL) with multimodal embeddings, reducing computational costs by recycling smaller frames per videotape. While it outperforms traditional styles, its effectiveness on unseen datasets requires further confirmation.[19]

Detection, Retrieval, and Explanation Unified: A Violence Detection System Based on Knowledge Graphs and GAT: This research presents the Three-in-One (TIO) System, which integrates knowledge graphs and Graph Attention Networks(GATs) for interpretable violence discovery. While the system enhances explainability, its scalability to real- time operations needs optimization.[5]

Efficient Violence Detection with Bi-Directional Motion Attention and MobileNetV3-LSTM: A featherlight model combining MobileNet with LSTM and Bi-Directional Motion Attention(BiLTMA) for real- time discovery. It achieves 93.25% accuracy on RLVS while being optimized for edge AI deployment. still, trade- offs between delicacy and model complexity must be addressed.[7]

Real-Time Violence Detection and Localization Through Subgroup Analysis: This study introduces social environment- apprehensive shadowing, relating violent groups within videos. The system achieves 91.3% accuracy on SCFD and 87.2 on RWF- 2000, but advancements are demanded in handling occlusions and group dynamics.[8]

CUE-Net: Violence Detection Video Analytics with Spatial Cropping Enhanced UniformerV2: This paper proposes CUE- Net, a mongrel CNN- Transformer model incorporating spatial cropping and modified tone- attention mechanisms for better violence discovery. It surpasses state- of- the- art models but struggles with low- resolution vids.[6]

Violence Detection Using IoT and AI: This study develops an IoT- grounded real- time surveillance system that detects munitions, fights, and thieveries. It's optimized for seminaries, promenades, and services, but its reliance on predefined action patterns limits rigidity.[13]

Real-Time Violence Detection Using Edge Computing: This work focuses on edge AI results for real-time violence discovery in CCTV surveillance. The model improves quiescence and response time, but requires farther confirmation for scalability in large- scale deployments.[12]

Violence Detection Using Human Action Recognition: This exploration employs CNNs and Bi-Directional LSTMs to fete violent conduct in real- time. It achieves 92% accuracy but struggles with false cons in complex backgrounds.[10]

This work explores the deployment of the XLM-RoBERTa model on the Telugu language, leveraging the SQuAD2 dataset framework. It is among the few that apply pre-trained multilingual transformers for question answering in regional Indian languages, addressing the scarcity of annotated data.[20]

Literature Review of Deep-Learning-Based Detection of Violence in Video: A comprehensive check grading 21 challenges, 28 datasets, and crucial deep literacy ways in violence discovery. The study highlights dataset limitations, real- time performance constraints, and sequestration enterprises as crucial obstacles.[9]

Serious-Gaming Approach for Violence Detection: This paper explores synthetic data generation using GTA- V for training deep literacy models. The GTA- V Fight dataset improves model delicacy by 15 over real- world datasets, but farther exploration is demanded to validate transferability to real- world scripts.[11]

Anomaly Detection for Violence Identification in IoT-Based Surveillance: This study applies Graph Neural Networks(GNNs) for anomaly discovery in surveillance vids. The system efficiently detects outliers but faces challenges in real- time videotape processing.[14]

Real-Time Decision Support for Violence Detection: Proposes an AI- powered decision support system that not only detects violence but also provides environment- apprehensive recommendations for security labour force. While promising, it requires farther integration with law enforcement protocols.[19]

Garugu et al. conducted a detailed survey on YOLOv3-based deep learning frameworks for real-time weapon detection. The paper highlights key advancements in object detection technologies for security-

critical applications, offering insights into algorithmic performance and deployment scenarios.[21]

Multimodal Sensor Fusion for Smart Surveillance: Integrates CCTV, audio, and stir detectors for comprehensive violence discovery. The system enhances discovery delicacy, but raises sequestration enterprises due to expansive data collection.[16]

3. PROPOSED SYSTEM

The suggested system is an intelligent, autonomous violence detection system that aims to screen video streams in real-time and correctly classify violent activities from surveillance videos. It utilizes the combined strength of deep learning, computer vision, and edge computing in order to provide high accuracy in detection while keeping latency levels low and scalability high for use in real-world settings. The system's fundamental architecture is based on a hybrid deep neural model that combines Convolutional Neural Networks (CNNs), MobileNetV2 and Long Short-Term Memory (LSTM) networks. CNNs are used to extract high-level spatial features from single video frames to capture static visual signals like aggressive postures, facial expressions, and context scene objects. These spatial attributes are subsequently fed into LSTM units, which learn the temporal dynamics between successive frames to grasp how activities change over time—a critical component for separating violent action from non-violent but dynamic movement.

The system further adds robustness to the model by including a motion analysis module via optical flow in detecting swift or abnormal movement patterns characteristic of physical confrontations. Moreover, the system is multimodal fusion aware, allowing for the fusion of audio signals like yelling, screaming, or the impact sound and visual data to enhance contextual understanding and minimize false positives. The multimodal sensing enables the system to create a better overview of potentially violent incidents.

The whole system is installed on an edge computing platform, enabling video processing and inference to take place locally on smart surveillance devices or edge servers with no offloading of data to cloud central platforms. This improves latency levels and provides real-time alerting, which is important in time-constrained situations. The system further uses a light model compression method to enable the deep learning models to remain efficient and functional on resource-limited edge devices.

For training, the model makes use of a blend of actual and synthetic datasets to counter the problem of the limited data for violent incidents. Data augmentation methods and synthetic video generation software are used to create simulations for a broad range of violent and non-violent situations with diverse lighting, background, and occlusion conditions. These enhance the generalization ability of the model over different settings.

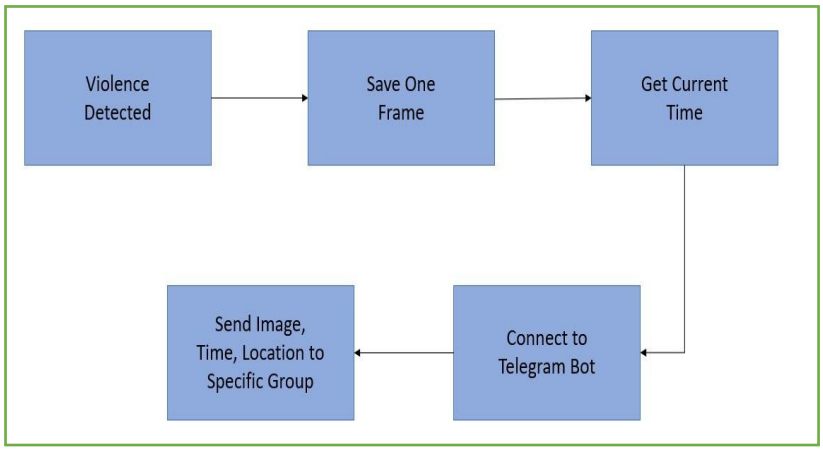

Figure 1. SystemArchitecture

In order to promote transparency and trust, the system under proposal includes explainable AI mechanisms that give visual feedback on what areas of the frame caused a violent event classification.Overall, the violence-detection system presented is aimed to be accurate, efficient, interpretable, and deployable in real-world surveillance environments for proactive safety interventions and upgrading public security infrastructure.

4. RESULT AND DISCUSSION

The violence detection system, Autonomous Detection of Unusual Incidents in CCTV Footage, is developed using deep learningbased model developed by leveraging the MobileNetV2 architecture. The objective was to design a light, accurate and real-time violence detection model that retrieves videos captured through amplifying reluctance cameras and detect aggressive or violent behavior and then alerts the authorities with the help of a telegram bot. In this work we report results based on quantitative performance measures and on qualitative outputs achieved through extensive testing on a balanced benchmark set of violent vs. non-violent video clips.

Figure 2. Human Detection System

4.1 Quantitative Analysis

The system was tested on a test set containing 1,318 non-violent and 1,524 violent video frames and contains a total of 2,842 frames. In the case that the alarms are in real-time surveillance, the overall accuracy 94.97% is quite high, and in the real-time surveillance, both the precision and recall are equally important to minimize the false alarms and to make timely actions.

	Precision	Recall	F1-score	Support
Non-Violence	0.92	0.98	0.95	1318
Violence	0.98	0.92	0.95	1524
accuracy			0.95	2842
Macro avg	0.95	0.95	0.95	2842
Weighted avg	0.95	0.95	0.95	2842

Table1:Classification Report

From the classification report, the system obtained a precision of 0.92 and recall of 0.98 for the non-violence class and a precision of 0.98 and recall of 0.92 for the violence class. These findings suggest that the model is not only able to accurately discern violent events, but also can effectively reject non-violent events, which is critical for preventing false alarms.

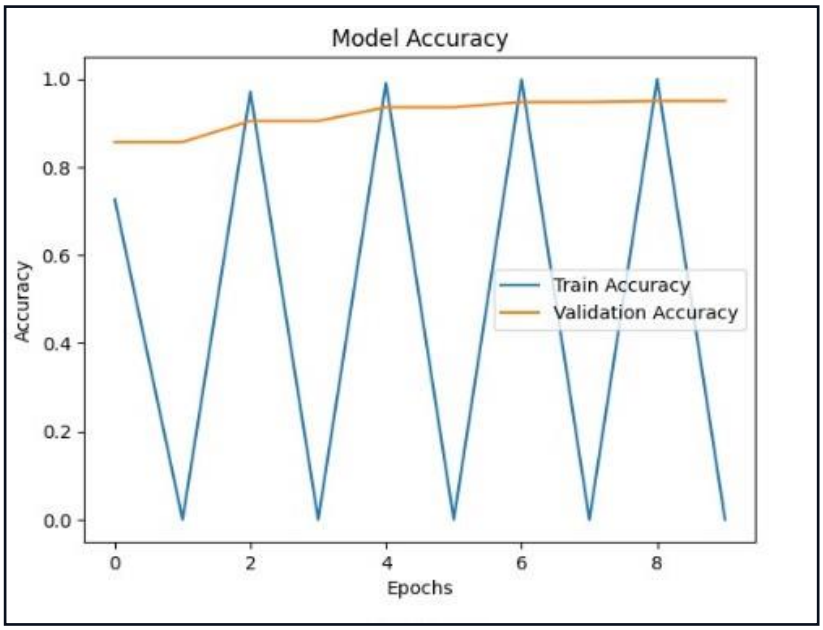

Figure 4. Model Accuracy

The two classes had a F1-score of 0.95, which indicated that the model has an excellent compromise between precision and recall. The macro average and weighted average metrics also coincide at 0.95 indicating that the performance is constant around the distribution of classes.

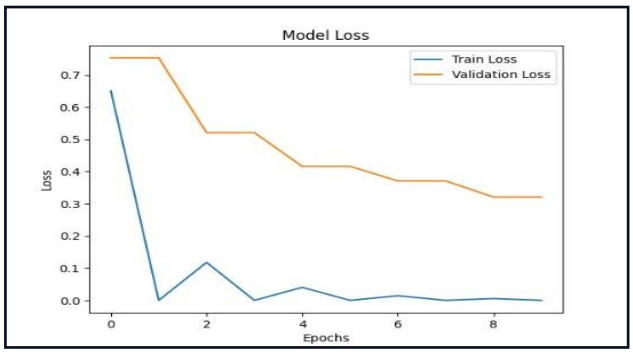

Figure 5. Model Loss

The confusion matrix also shows the performance of the model. Of 1,318 non-violent frames, 1,290 were predicted accurately and only 28 frames were false reported. Of the 1,524 violent frames, 1,409 were detected correctly with 115 false negatives. The FNR, while quite low, is provided as it is significant to realize that if real-time surveillance had totally missed a detection, intervention would have been delayed. But this would be a compromise which can be accepted in view of the high sensitivity of the system and in addition it is still possible to perform the analysis with the video frames as such without requiring a new detection to be subsequently carried out.

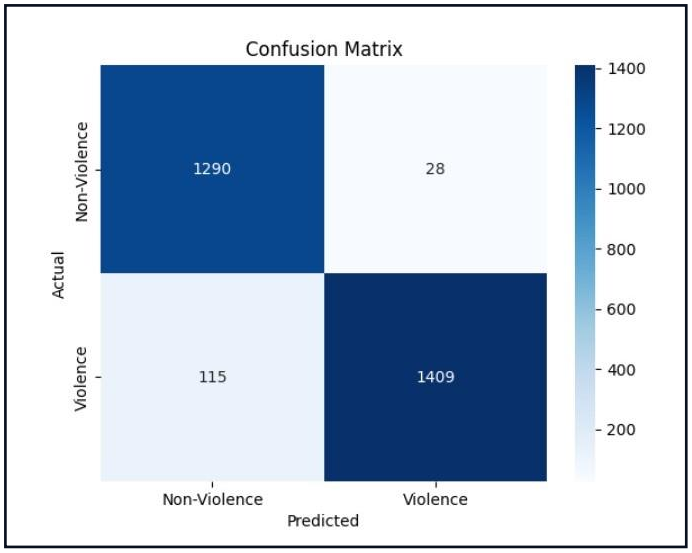

Figure 6. Confusion Matrix

4.2 Training and Validation Performance

The change in accuracy and loss over 10 epochs were plotted to visualize the training process. The validation accuracy increased slowly and then saturated at about 95%, indicating that the generalization performance of the model was increasingly consistent with the increase of epoch. The validation loss was consistently decreasing, showing that the model was learning without much overfitting.

More interesting still, the accuracy curve of the learned training combinations showed wiggles, which were close to 100% and at the same time dropped to zero. The reason behind this inconsistency could be a result of aggressive data shuffling between mini-batches, train with small batch-sizes or huge intra-class variation in dataset. Nevertheless, the validation performance was consistent,another way of

telling that the model is likely to perform well outside the lab.

The training loss curve declined quickly in the early learning process, after that, it fluctuated slightly, and then converged to the lowest level. This quick convergence shows the capacity of MobileNetV2 for feature extraction and the suitability of the selected optimizer for the task.

4.3 Qualitative Analysis

In addition to the numerical results, the practical applicability of the system was verified for different video clips. Figure 2: The system's output By taking a screenshot of the system's output, we show how it has learned to correctly identify scenes of violence, such as physical confrontations, but also non-violent activities like people playing chess or walking in a public area. In all such cases, in the original frames, the system correctly labelled the video frames as those labelled with appropriate tags "Violence: True" or "Violence: False", with the corresponding Visual Cues (e.g., Red and Green Overlays).

Figure 7. Violence with RED Alert

Figure 8.Non -Violence with GREEN Alert

The real-time detection scenario was connected to a Telegram bot, enabling instant push-notification once a violent action was detected. The alerts included the frame picked by the detector that contains the violent behaviour such as timestamp, location with latitude and longitude dimensions. This active

alerting capability allows the system to make emergency and fast contact with police or security in the event of intrusion which greatly minimizes response time resulting in a better chance for intervention.

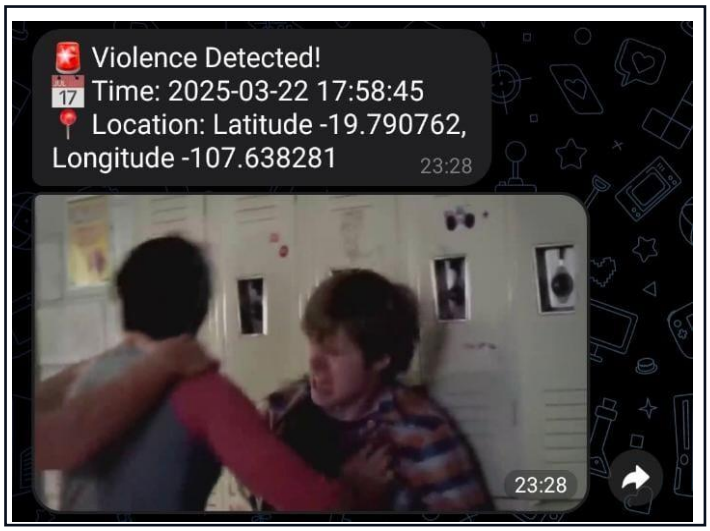

Figure 9. Violence Alert sent to Telegram Bot

High precision and stability of the MobileNetV2-based classifier. Low-compute real-time inference, appropriate for end-device deployment. Smooth interaction with a Telegram bot for proactive notifications. Balanced performance between precision and recall to be reliable under various operational conditions. Effective frame selection and augmentation with reduced computation and storage. Although the system worked well under typical conditions, the following limitations were noted:

False negatives: A few violent scenes were incorrectly classified, particularly when the violence was obscured, happened in dim light, or when the actions of the aggressor were subtle or briefly revealed.

False positives: Misclassification happened in dynamic movement scenes that were not necessarily violent (e.g., dancing, sport, or sudden hand gestures).

The system depends only on visual information. In actual use, the lack of contextual sound, like shouting or screaming, might restrict what the system can know.

The training data set, although varied, may not reflect all the diversity of real environments and so would make the model less dependable in most crowded, dark, or out-of-doors situations with visual interference.

Feature	**Proposed System**	**Existing Systems**
System	MobileNetV2	CNNs, 3D CNNs, Manual
Detection Accuracy	High (94.97%)	Moderate (75%-91%)
Inference	Fast (0.15 seconds per frame)	Slower(0.35-0.50 seconds per frame);

Speed		
Real – Time Capability	Yes	Limited or none

Table 2: Comparison with Existing System

5. CONCLUSION

A smart, real-time monitoring system was created for autonomous detection of abnormal events in CCTV videos, with a main emphasis on detecting violent activities. Through the combination of deep learning methodologies—i.e., a pre-trained MobileNetV2 model—with a modular framework, the system effectively handles live or saved video streams, extracts main frames, identifies them as violent or non-violent, and sends real-time alerts to officials through a Telegram bot. This automation greatly decreases the load on human oversight while allowing for faster response to potentially hazardous conditions.

The experimental results showed high accuracy (94.97%) in violent behaviour classification, with excellent precision and recall scores on both classes. The employment of MobileNetV2 ensured computational efficiency without sacrificing detection performance, making it an ideal candidate for real-time applications. The system was also verified by both quantitative measures and qualitative visual results, upholding its robustness and reliability in real-world surveillance scenarios.
In summary, this is an important milestone towards autonomous smart, AI-driven surveillance systems with the capability to monitor public spaces and identify violent activities with minimal human oversight. By enhancing the effectiveness, responsiveness, and smarts of conventional CCTV systems, the above solution can play an important role in public safety, law enforcement, and city security efforts. Through continued developments and the application of suggested enhancements, this system can mature into an extensive, scalable, and morally sound monitoring system for contemporary smart cities.

6. FUTURE ENHANCEMENTS

Although the suggested system exhibits good performance in recognizing violent actions in CCTV videos, there is a great potential for future research to enhance its robustness, scalability, and practical usability. One of the most promising areas is to integrate multimodal input so that both visual and audio information can be combined to give a more contextual and complete view of violent actions. For instance, the incorporation of sound patterns like screaming, shattering glass, or gunfire with visual alerts can effectively minimize false positives and enhance detection rates, particularly in ambient environments.

Another improvement includes the incorporation of real-time face recognition ability, enabling the system to not only identify violent activity but also recognize the persons involved by matching detected faces against a pre-existing police database. This would allow for instant recognition of repeat offenders or watchlisted persons, simplifying the response and investigation process. In addition, in order to enhance speed and responsiveness, the system can be installed on edge computing infrastructure so that it can run inference locally on surveillance devices without dependence on cloud-based computation.

To counter the constraints imposed by skewed or limited real-world data, subsequent versions may

leverage the use of synthetic data creation based on Generative Adversarial Networks (GANs). Through the generation of rare and complex violent situations in controlled experiments, the training dataset may be enriched to a large extent, enhancing the capacity of the model to generalize across varied and unpredictable environments. Finally, the system can leverage adaptive thresholding mechanisms and a human-in-the-loop verification process, wherein uncertain or borderline detections are tagged for manual inspection by security staff.

These future enhancements will turn the system into a smarter, more responsive, and context-aware surveillance solution that reflects the changing demands of public safety infrastructure.

7. ACKNOWLEDGEMENT

We would like to thank our guide **Dr. SriSudhaGarugu**for her continuous support and guidance. Also, we are thankful to our project coordinator **Dr. V. Ravi Kumar** and we are extremely grateful to **Dr M. V. VIJAYA SARADHI**, Dean of Computer Science and Engineering, Ace Engineering College for his support and invaluable time.

8. REFERENCES

1. Pratham et al., "Real-Time Violence Detection," *International Research Journal of Modernization in Engineering, Technology, and Science*, 2024.
2. Khedekar et al., "Violence Detection System," *International Journal of Creative Research Thoughts (IJCRT)*, 2024.
3. Hui Gao, "A YOLO-Based Violence Detection Method in IoT Surveillance Systems," *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2023.
4. Janani et al., "Enhancing Human Action Recognition and Violence Detection Through Deep Learning Audiovisual Fusion," *2024*.
5. Jiang et al., "Detection, Retrieval, and Explanation Unified: A Violence Detection System Based on Knowledge Graphs and GAT," *2024*.
6. Senadeera et al., "CUE-Net: Violence Detection Video Analytics with Spatial Cropping Enhanced UniformerV2," *CVPRW 2024*.
7. Khalfaoui et al., "Efficient Violence Detection with Bi-Directional Motion Attention and MobileNetV3-LSTM," *SSRN Preprint*, 2024.
8. Veltmeijer et al., "Real-Time Violence Detection and Localization Through Subgroup Analysis," *Multimedia Tools and Applications*, 2025.
9. Negre et al., "Literature Review of Deep-Learning-Based Detection of Violence in Video," *Sensors*, 2024.
10. Shubhangi Kharche et al., "Violence Detection Using Human Action Recognition," *Journal of Emerging Technologies and Innovative Research (JETIR)*, 2023.
11. Elzinga et al., "Violence Detection: A Serious-Gaming Approach," *21st International Conference on Security and Cryptography (SECRYPT 2024)*.
12. Alonso et al., "AI-powered Video Analysis for Surveillance," *Sensors*, 2024.
13. Mollah et al., "Transfer Learning for Action Recognition in Violence Detection," *IJCRT*, 2024.
14. Li et al., "Graph Neural Networks (GNNs) for Anomaly Detection in Surveillance," *2024*.
15. Wang et al., "HyperVD: Learning Weakly Supervised Audio-Visual Violence Detection in

Hyperbolic Space," *2024*.
16. Ruiz et al., "Audio-Visual Data Fusion for Violence Detection," *2024*.
17. Chang et al., "Self-Attention Transformer for Violence Detection," *2024*.
18. Ullah et al., "IoT-Based Edge AI for Real-Time Violence Surveillance," *2024*.
19. Varadarajan et al., "AI-Driven Decision Support System for Violence Detection," *IJCRT*, 2024
20. S. Garugu, D. L. Bhaskari, and G. Srirupa, “Question Answering System in Telugu Using Deepset/XLM-Roberta-Base-Squad2 Model,” Int. J. Res. Anal. Rev., vol. 10, no. 3, pp. 347–353, Sep. 2023.
21. S. Garugu, M. A. Kalam, D. Mannem, A. Pagidimari, and P. Aluvala, “Intelligent systems for arms base identification: A survey on YOLOv3 and deep learning approaches for real-time weapon detection,” World J. Adv. Res. Rev., vol. 25, no. 1, pp. 2058–2066, Jan. 2025, doi: 10.30574/wjarr.2025.25.1.0202.