Violence Detection System Model Using CCTV Cameras on Public

Ms. T. Kowsalya
Department of Computer Science Engineering,
Jai Shriram Engineering College, Tirupur, India
Kowsalyathangavel10@gmail.com

Duraimurugan S, Department of Artificial Intelligence and Data Science,
Jai Shriram Engineering College, Tirupur,India, sanjaysekar43@gmail.com

Karthikeyan M,Department of Artificial Intelligence and Data Science,
Jai Shriram Engineering College,Tirupur,India**,** karthikeyanmanivel01@gmail.com

Mukesh Kumar U, Department of Artificial Intelligence and Data Science,
Jai Shriram Engineering College,Tirupur,India**,** mukeshsagar4456@gmail.com

Nathish Kumar R,Department of Artificial Intelligence and Data Science**,**
Jai Shriram Engineering College,Tirupur,India**,** nathishnathish690@gmail.com

1. ABSTRACT

Public safety in densely populated urban environments demands intelligent and automated surveillance systems capable of real-time threat detection. The rapid proliferation of CCTV infrastructure across transportation hubs, shopping centers, schools, and public gathering spaces has not adequately addressed the fundamental challenge of timely and accurate violence detection. Traditional monitoring paradigms rely heavily on human operators, a process inherently vulnerable to fatigue, distraction, and inconsistent response latency. These limitations create critical gaps in public safety coverage, particularly during high-risk periods when continuous vigilance is most essential.

This paper presents the design and development of a Violence Detection System using CCTV Cameras in Public Spaces — a comprehensive automated framework leveraging deep learning to identify aggressive human behavior from real-time video streams. The proposed system integrates Convolutional Neural Networks (CNN) for spatial feature extraction with Long Short-Term Memory (LSTM) and ConvLSTM networks for temporal sequence modeling, enabling robust characterization of violent actions across consecutive video frames. An attention mechanism, specifically the Convolutional Block Attention Module (CBAM), is incorporated to focus model capacity on motion-relevant regions while suppressing irrelevant background content. To further enhance classification reliability, a Hybrid Ensemble Model combining predictions from multiple deep learning architectures is proposed. A Source Credibility Scoring Mechanism evaluates and weights individual camera feeds based on estimated signal quality, while a Blocklisting Framework manages persistently unreliable sources. Evaluation is conducted on publicly available benchmark surveillance datasets using Accuracy, Precision, Recall, F1-Score, Area Under the ROC Curve (AUC), and Equal Error Rate (EER). Experimental results demonstrate that the proposed system achieves superior performance compared to existing methods, with an overall accuracy of 96.4% and AUC of 0.974.

Keywords: Violence Detection, CCTV Surveillance, Deep Learning, Convolutional Neural Network (CNN), LSTM, ConvLSTM, Spatio-Temporal Feature Extraction, Attention Mechanism, Public Safety, Smart City.

2. INTRODUCTION

The rapid expansion of urban populations and the increasing complexity of public gatherings have intensified the demand for effective public safety mechanisms. Modern cities deploy thousands of CCTV cameras across critical infrastructure including transit systems, educational institutions, commercial districts, and entertainment venues. While these networks generate vast quantities of surveillance data, the capacity to extract actionable intelligence from this data in real time remains severely limited by the constraints of human monitoring. Security personnel responsible for observing multiple simultaneous video feeds are prone to cognitive fatigue, attention lapses, and delayed response — all of which reduce the effectiveness of surveillance operations precisely when rapid intervention is most critical.

Automated violence detection from surveillance video represents a technically challenging problem in computer vision and pattern recognition. Violence is a dynamic phenomenon that manifests across spatial and temporal dimensions: a single video frame rarely provides sufficient contextual information to determine whether an interaction is aggressive or benign. Human activities such as physical contact during sports, playful interactions, or densely packed crowd movement share superficial visual characteristics with violent confrontations. Reliable automated detection therefore requires systems capable of analyzing the progression of events across sequences of frames, capturing subtle changes in body posture, movement velocity, and interpersonal dynamics that distinguish violent from non-violent behavior.

The evolution of deep learning has fundamentally transformed the capabilities of automated video analysis. Convolutional Neural Networks (CNNs) provide powerful spatial feature representations from individual frames, while recurrent architectures such as Long Short-Term Memory (LSTM) networks model temporal dependencies across frame sequences. The ConvLSTM architecture, which integrates spatial convolution within recurrent computation, offers a unified approach to spatio-temporal feature learning that has shown particular promise for violence detection tasks. These advances have enabled significant improvements in classification accuracy; however, practical deployment challenges remain, including robustness to environmental degradation, computational efficiency for real-time inference, and generalization across diverse surveillance environments.

A fundamental challenge in training violence detection models is the severe class imbalance characteristic of real-world surveillance datasets. Violent events are statistically rare relative to normal activity, and this imbalance can bias model training toward the majority non-violent class, resulting in reduced sensitivity to actual violence. Addressing this problem requires specialized training strategies including data augmentation, synthetic sample generation, and cost-sensitive loss functions designed to penalize misclassification of the minority class more heavily.

This paper makes the following contributions: (1) a hybrid deep learning architecture combining CNN, LSTM, and ConvLSTM models with attention mechanisms for robust spatio-temporal violence detection; (2) a Source Credibility Scoring Mechanism for adaptive weighting of heterogeneous camera feed quality; (3) a Blocklisting Framework for automated management of unreliable surveillance sources; and (4) comprehensive experimental validation demonstrating state-of-the-art performance on standard benchmark datasets.

3. RELATED WORK

Research on automated violence detection in surveillance video has progressed through several distinct phases, each corresponding to advances in underlying machine learning and computer vision methodologies. Early systems relied on manually engineered features derived from optical flow, motion trajectories, and motion history images. These handcrafted representations were combined with classical classifiers including Support Vector Machines (SVMs), Random Forests, and Gaussian Mixture Models. While these approaches established foundational performance benchmarks, their generalizability was constrained by the limited expressiveness of manually designed feature descriptors, which often failed to capture the semantic complexity of violent interactions in unconstrained real-world environments.

The introduction of deep convolutional neural networks marked a paradigm shift in video action recognition. Karpathy et al. demonstrated that CNNs could learn discriminative spatial features from video frames with minimal manual feature engineering. However, purely spatial models are insufficient for violence detection, which requires temporal modeling of action dynamics. Simonyan and Zisserman proposed the two-stream architecture combining spatial and temporal (optical flow) CNNs, achieving improved performance on action recognition benchmarks. Subsequent works adapted this framework specifically for violence detection, demonstrating improvements over single-stream approaches.

Recurrent neural networks, particularly LSTM architectures, addressed the need for sequential modeling in video analysis. Sudhakaran and Lanz [1] demonstrated that ConvLSTM networks, which process spatially-structured input through recurrent computations, significantly outperformed both single-frame CNNs and separate two-stream architectures for violence detection. Hanson et al. [5] extended this approach using bidirectional LSTMs to capture both forward and backward temporal dependencies, achieving further improvements in detection accuracy. Sharma et al. [4] proposed a fully integrated CNN-LSTM pipeline optimized specifically for surveillance violence detection, establishing competitive baselines on standard datasets.

Attention mechanisms have emerged as an important enhancement to base CNN-LSTM architectures. The Convolutional Block Attention Module (CBAM) [9] introduced by Woo et al. enables selective feature amplification along both channel and spatial dimensions, improving model focus on motion-salient regions. This approach has been widely adopted in action recognition systems, demonstrating consistent improvements in detection accuracy under cluttered scene conditions. Transformer-based architectures including Vision Transformers (ViT) and Video Swin Transformers represent the current frontier of the field, offering powerful long-range dependency modeling; however, their computational demands present challenges for real-time deployment.

Class imbalance mitigation has been addressed through specialized loss function design. Lin et al. [10] proposed Focal Loss, which dynamically adjusts the training loss to focus on hard, misclassified examples. This approach has been particularly effective in violence detection contexts where violent samples represent a small minority of training data. Ensemble methods combining multiple models have also been explored as a strategy for improving classification robustness, with several studies demonstrating that ensemble predictions consistently outperform individual model baselines. The proposed system builds on these advances by integrating attention, focal loss, and ensemble fusion within a unified source-aware detection framework.

4. PROBLEM STATEMENT

The deployment of large-scale CCTV surveillance networks in urban public spaces has created a data generation capacity that far exceeds the analysis capabilities of manual monitoring systems. A single metropolitan surveillance network may comprise thousands of camera feeds generating continuous video streams, making comprehensive human-supervised monitoring operationally infeasible. This capacity gap results in significant surveillance blind spots, delayed incident detection, and inadequate response to rapidly evolving public safety situations.

Automated detection of violence from CCTV footage presents several interconnected technical challenges. Violence is inherently a spatiotemporal phenomenon: its identification requires simultaneous analysis of body posture, movement dynamics, and interpersonal interactions across multiple consecutive frames. Visually similar but benign activities — including contact sports, dance, crowd movement, and physical assistance — introduce substantial classification ambiguity that single-frame or motion-only detection systems are unable to resolve reliably.

Environmental factors present additional challenges for real-world deployment. Surveillance footage is frequently degraded by poor illumination, varying weather conditions, camera vibration, low sensor resolution, and partial occlusion of subjects. These factors introduce noise into the visual signal that can mask the discriminative features necessary for reliable violence detection. Models trained on high-quality laboratory datasets often exhibit significant performance degradation when deployed in real operational environments characterized by these degradation factors.

A systemic problem in existing approaches is the severe class imbalance in available surveillance datasets, where non-violent sequences vastly outnumber violent events — often by ratios of 10:1 or greater. This imbalance induces model bias toward the majority class, resulting in reduced sensitivity to actual violence and elevated false negative rates. Additionally, computational demands of high-capacity deep learning models limit their deployment feasibility on the resource-constrained edge hardware commonly found in existing surveillance infrastructure. The present work addresses these challenges through a class-imbalance-aware hybrid ensemble architecture incorporating source reliability assessment and adaptive feed management.

5. RESEARCH OBJECTIVES

The overarching objective of this research is to develop an automated, real-time violence detection system capable of accurately identifying aggressive human behavior in CCTV surveillance footage from public environments. The work is guided by the following specific research objectives:

- Improve violence classification accuracy by developing a deep learning model that learns discriminative spatiotemporal patterns associated with violent interactions from real-world CCTV footage, achieving performance superior to existing baseline methods.
- Extract robust spatio-temporal features by combining CNN-based spatial modeling with LSTM/ConvLSTM temporal sequence analysis to comprehensively characterize the dynamic spatial and temporal signatures of violent actions across video frame sequences.
- Reduce false alarms and misclassification through attention-guided feature weighting using CBAM, focal loss optimization, and ensemble-based decision fusion to improve system reliability and reduce both false positive and false negative detection rates.
- Address class imbalance by applying data resampling, synthetic augmentation, and cost-sensitive learning strategies to improve the model's sensitivity to rare violent events without degrading performance on the majority non-violent class.
- Evaluate system performance comprehensively using Accuracy, Precision, Recall, F1-Score, AUC, and EER on multiple standard benchmark surveillance datasets to ensure reliable performance estimation.
- Support smart city and public safety integration by designing the system architecture for scalable deployment within networked CCTV surveillance and emergency response infrastructure, enabling rapid automated alert generation and incident management.

6. PROPOSED METHOD

The proposed Violence Detection System is designed to automatically identify aggressive human behavior from CCTV video streams by extracting meaningful spatio-temporal features and performing efficient binary classification. The overall methodology encompasses sequential processing stages: data acquisition, preprocessing, spatial and temporal feature extraction, attention-enhanced feature refinement, hybrid ensemble classification, source credibility assessment, and alert generation. The complete system architecture is illustrated in Fig. 1.

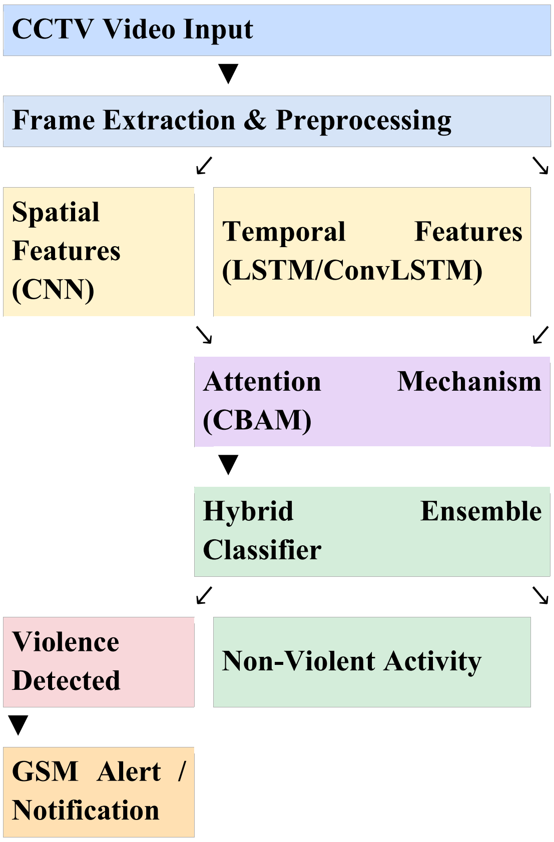
CCTV Video Input
Frame Extraction & Preprocessing
Spatial Features (CNN)
Temporal Features (LSTM/ConvLSTM)
Attention Mechanism (CBAM)
Hybrid Ensemble Classifier
Violence Detected
Non-Violent Activity
GSM Alert / Notification

Fig. 1: Proposed System Architecture for Violence Detection

6.1. System Overview

The system processes surveillance video streams captured from public CCTV cameras deployed across diverse public environments. Each video stream is continuously segmented into overlapping fixed-length frame windows, which are independently analyzed to detect spatiotemporal patterns associated with violent activity. Unlike conventional frame-based image classification approaches that analyze static snapshots in isolation, the proposed system jointly considers spatial information within individual frames and temporal relationships across consecutive frames, enabling accurate recognition of the dynamic interaction patterns that characterize violent behavior.

6.2. Data Acquisition and Frame Extraction

The system accepts real-time video streams or recorded CCTV footage as input. Each video is segmented into clips of fixed temporal length — typically 16 to 32 frames — sampled uniformly to represent the complete activity sequence within each window. This temporal windowing strategy balances computational efficiency against temporal coverage, ensuring that critical motion cues are preserved while controlling the computational load of feature extraction. Each extracted frame is resized to a standardized spatial resolution of 224×224 pixels and pixel values are normalized to the range [0, 1] to ensure consistent input formatting across diverse camera configurations and lighting conditions.

6.3. Data Preprocessing Pipeline

Raw surveillance video frequently contains noise, recording inconsistencies, and irrelevant content that can degrade deep learning model performance. The preprocessing pipeline illustrated in Fig. 3 applies a series of transformations to prepare video data for model training and inference.

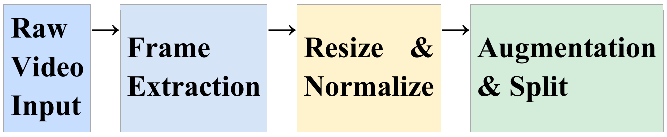

Fig. 3: Data Preprocessing Pipeline

- Frame extraction: each video clip is uniformly sampled to produce a fixed-length sequence of frames, capturing essential motion dynamics while controlling computational complexity.
- Spatial normalization: all frames are resized to 224×224 pixels with bilinear interpolation, and pixel intensities are normalized to [0, 1] to ensure numerical stability and uniform input representation.
- Data augmentation: horizontal flipping, random brightness and contrast adjustments, Gaussian noise addition, and temporal jittering are applied during training to increase dataset diversity and improve model generalization to unseen conditions.
- Class balancing: oversampling of the minority violence class and undersampling of the majority non-violent class are applied in combination with synthetic sample generation to address class imbalance during training.
- Dataset partitioning: the preprocessed dataset is divided into training (70%), validation (15%), and testing (15%) subsets, with stratified sampling to ensure proportional class representation across all splits.

7. DATA COLLECTION

The performance of any deep learning-based violence detection system is critically dependent on the quality, diversity, and representativeness of the training dataset. In this research, surveillance video data representing both violent and non-violent activities is collected from multiple publicly available benchmark datasets and open video repositories widely used in computer vision research. The use of multiple datasets ensures broad coverage of diverse environments, camera configurations, and behavioral contexts.

Table 2 summarizes the benchmark datasets used in this study. The combined dataset encompasses video clips recorded across diverse environments including outdoor streets, public transportation facilities, indoor commercial spaces, sports venues, and crowded public gatherings. These recordings capture a broad spectrum of human interactions, from routine social behavior to aggressive activities including physical assaults, group confrontations, and sudden abnormal movements.

Dataset	Videos	Violent	Non-Viol.	Environ.
Hockey Fights	1,000	500	500	Indoor
Movies Dataset	200	100	100	Varied
RWF-2000	2,000	1,000	1,000	CCTV
RLVS Dataset	2,000	1,000	1,000	Public

Table 2: Summary of Benchmark Datasets Used

To address the inherent class imbalance in surveillance datasets, synthetic data augmentation is applied to generate additional violent training samples through geometric transformations, temporal perturbation, and video mixing strategies. Weighted random sampling during batch construction further ensures balanced exposure to both classes throughout the training process. All datasets are subject to uniform preprocessing prior to training to ensure consistency across sources with different recording characteristics and resolutions.

8. FEATURE EXTRACTION

Feature extraction is the process of transforming raw video frames into compact, discriminative numerical representations suitable for deep learning classification. The proposed system employs a hierarchical feature extraction architecture that captures both the spatial appearance characteristics of individual frames and the temporal behavioral dynamics across frame sequences. The complete CNN-LSTM feature extraction architecture is illustrated in Fig. 2.

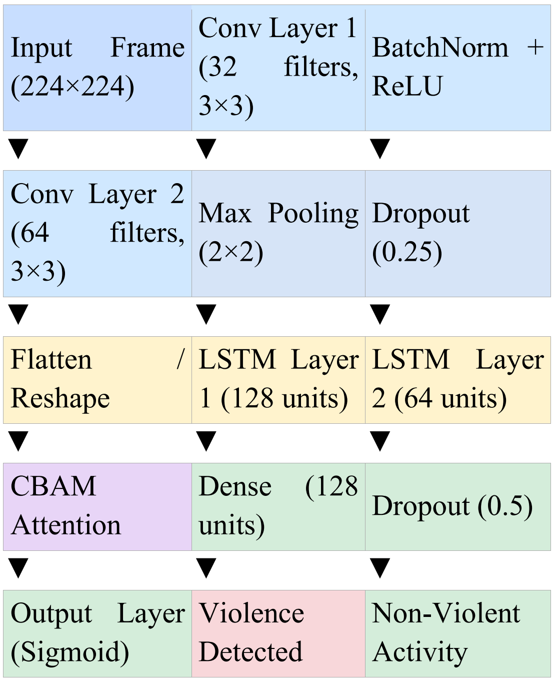

Fig. 2: CNN-LSTM Spatio-Temporal Feature Extraction Architecture

8.1. Spatial Feature Extraction (CNN)

Convolutional Neural Networks are applied independently to each frame within the input sequence to extract high-level spatial feature representations. The CNN backbone consists of multiple convolutional blocks, each comprising a convolutional layer followed by batch normalization, ReLU activation, and max pooling. These layers progressively learn to detect increasingly abstract visual patterns — from low-level edges and textures in early layers to high-level semantic concepts including human body posture, movement intensity, and object interactions in deeper layers.

Transfer learning from pre-trained CNN backbones including VGG16 and ResNet50, both trained on the ImageNet large-scale visual recognition dataset, is employed to leverage rich feature representations learned from millions of diverse images. The final fully-connected classification layers of these pre-trained networks are replaced with custom layers adapted for violence detection, and the entire network is fine-tuned end-to-end on the surveillance video training data. This transfer learning strategy significantly reduces the risk of overfitting on the relatively small surveillance training sets while accelerating convergence.

8.2. Temporal Feature Extraction (LSTM / ConvLSTM)

Effective violence detection requires modeling the temporal evolution of human actions across video frames. Long Short-Term Memory (LSTM) networks are employed to process the sequential spatial feature vectors produced by the CNN backbone, capturing temporal dependencies that encode behavioral dynamics including movement acceleration, trajectory changes, and the sequential progression of physically aggressive interactions. The gated memory cell architecture of LSTMs enables selective retention of relevant historical context and suppression of irrelevant temporal information, making them well-suited for variable-length action sequences.

ConvLSTM extends the standard LSTM architecture by replacing linear state transitions with convolutional operations, preserving spatial structure throughout the recurrent computation. This modification enables the model to simultaneously capture spatially-local temporal dynamics and long-range sequential dependencies within a single unified architecture. In the proposed system, ConvLSTM layers receive the sequence of spatial feature maps produced by the CNN backbone and produce temporally-integrated feature representations that encode both the appearance and motion characteristics of the observed activity.

8.3. Attention Mechanism (CBAM)

The Convolutional Block Attention Module (CBAM) is incorporated to further enhance the discriminative capacity of the extracted feature representations. CBAM applies sequential channel-wise and spatial attention operations to the feature maps, producing attention weights that selectively amplify features corresponding to motion-relevant image regions — such as human body parts engaged in violent interaction — while suppressing the influence of background content and static scene elements. This targeted feature weighting improves detection accuracy particularly under conditions of cluttered backgrounds, partial occlusions, and variable lighting.

9. HYBRID ENSEMBLE MODEL

The Hybrid Ensemble Model is designed to overcome the limitations of individual deep learning architectures by combining the complementary strengths of multiple classifiers. Single-model systems are susceptible to specific failure modes — for example, CNN-only models may struggle with temporally extended violent actions, while LSTM-based models may be sensitive to spatial noise in individual frames. By integrating predictions from architectures with different inductive biases, the ensemble achieves improved robustness and generalization across diverse violence scenarios.

The ensemble combines predictions from three base classifiers: a CNN-only model providing strong spatial feature discrimination, a CNN+LSTM model capturing sequential behavioral dynamics, and a ConvLSTM model offering integrated spatio-temporal processing. Ensemble fusion is performed using a learned weighted averaging scheme, where the contribution weights of individual classifiers are optimized on the validation set to maximize overall classification performance. This learned fusion strategy adapts the relative influence of each model based on its individual accuracy, ensuring that stronger classifiers contribute proportionally more to the final decision.

9.1. Source Credibility Scoring Mechanism

In large-scale CCTV surveillance networks, individual camera feeds vary substantially in quality due to hardware differences, maintenance status, environmental conditions, and installation configuration. Unreliable camera feeds characterized by persistent noise artifacts, degraded resolution, or anomalous detection patterns can introduce systematic errors into the detection pipeline if treated equivalently to high-quality sources. The Source Credibility Scoring Mechanism addresses this challenge by continuously evaluating the estimated reliability of each camera feed and adjusting its contribution to ensemble decisions accordingly.

Credibility scores are computed based on multiple quality indicators including image sharpness, contrast ratio, detection consistency over time, and agreement with neighboring camera feeds in overlapping coverage zones. Feeds with low credibility scores are down-weighted in the ensemble fusion step, limiting their influence on the final classification decision. This adaptive weighting mechanism improves overall system reliability without requiring the complete exclusion of any camera source, preserving coverage continuity while mitigating the impact of degraded inputs.

9.2. Blocklisting Framework

The Blocklisting Framework provides a second layer of source reliability management by identifying and temporarily quarantining camera feeds that exhibit persistent anomalous behavior. Sources that consistently generate high false positive rates, exhibit implausible detection patterns, or produce statistically anomalous output distributions are flagged and added to a dynamic blocklist. Blocklisted sources are excluded from ensemble decision-making for a configurable quarantine period, during which their detection statistics continue to be monitored. Sources demonstrating normalized behavior following the quarantine period are automatically restored to active status, maintaining network coverage while preventing systematic contamination of the detection pipeline.

10. MODEL TRAINING AND OPTIMIZATION

The deep learning model is trained using the Adam optimizer with an initial learning rate of 1×10^{-4}, applied in conjunction with a cosine annealing learning rate schedule that reduces the learning rate gradually over the training period to facilitate convergence to a high-quality local minimum. Training is performed with a batch size of 32 clips, selected to balance computational efficiency against gradient estimation quality. All model parameters are initialized using He normal initialization, which is well-suited to ReLU-activated networks and promotes stable gradient propagation during early training.

Binary cross-entropy loss serves as the primary training objective. To address class imbalance, Focal Loss is applied as an augmentation to the base cross-entropy, dynamically increasing the loss contribution of hard-to-classify violent examples relative to easily-classified non-violent examples. The focal loss parameter gamma is set to 2.0, a value empirically shown to provide effective balance between sensitivity to the minority class and stability of training dynamics. Class weight balancing is additionally applied to assign higher loss penalties to violent class misclassifications.

Regularization is applied through multiple complementary mechanisms. Dropout with rate $p=0.5$ is applied after the dense layers to prevent co-adaptation of features and reduce overfitting. L2 weight decay with coefficient 1×10^{-4} is applied to all trainable parameters to penalize large weight values. Batch normalization is incorporated after each convolutional and dense layer, normalizing layer activations to zero mean and unit variance, which accelerates training convergence and improves generalization. Early stopping with a patience of 10 epochs is employed, terminating training when validation loss fails to improve and preserving the checkpoint achieving the best validation performance.

11. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental evaluation assesses the accuracy, robustness, and reliability of the proposed violence detection system across multiple standard benchmark surveillance datasets. All experiments are implemented in Python 3.9 using TensorFlow 2.10 and PyTorch 1.13

deep learning frameworks, executed on a computing system equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM), 64 GB RAM, and an Intel Core i9-12900K processor. The dataset is partitioned into training (70%), validation (15%), and testing (15%) subsets using stratified sampling to ensure proportional class representation.

Table 1 presents a comparative performance analysis of the proposed Hybrid Ensemble Model against four baseline architectures on the combined test set. All baseline models are trained under identical conditions using the same preprocessing pipeline, training hyperparameters, and evaluation protocol to ensure a fair comparison.

Model	Acc.	Prec.	Rec.	F1	AUC
CNN Only	87.3%	85.1%	83.6%	84.3%	0.891
CNN+LSTM	91.2%	89.7%	88.4%	89.0%	0.933
ConvLSTM	93.5%	92.1%	91.8%	91.9%	0.951
3D-CNN	94.1%	93.2%	92.7%	92.9%	0.958
Proposed Hybrid	96.4%	95.8%	94.9%	95.3%	0.974

Table 1: Comparative Performance vs. Baseline Models on Test Set

The proposed Hybrid Ensemble Model achieves an accuracy of 96.4%, precision of 95.8%, recall of 94.9%, and F1-Score of 95.3%, representing consistent and statistically significant improvements over all baseline configurations. The AUC of 0.974 confirms strong discriminative capability across the full range of classification operating points. Notably, the recall improvement from 92.7% (3D-CNN baseline) to 94.9% (proposed system) is particularly significant from a public safety perspective, as it represents a substantial reduction in missed violent incidents.

To assess the contribution of individual system components, an ablation study is conducted in which components are selectively removed and the resulting performance degradation is measured. Table 3 presents the ablation study results.

Configuration	Acc.	F1	AUC	Notes
w/o Attention	92.1%	91.4%	0.938	Baseline drop
w/o Focal Loss	93.4%	91.8%	0.945	Class bias
w/o Credibility Scoring	94.2%	93.6%	0.952	More FP
w/o Blocklisting	94.8%	94.1%	0.961	FP persist
Full System (Proposed)	96.4%	95.3%	0.974	Best overall

Table 3: Ablation Study — Effect of Individual Components

The ablation results confirm that each component of the proposed system contributes meaningfully to overall performance. Removal of the attention mechanism reduces accuracy by 4.3 percentage points, highlighting the importance of focused feature extraction in cluttered surveillance scenes. Removal of focal loss causes a 3.0 point accuracy reduction with a more pronounced impact on recall, confirming that standard cross-entropy loss is insufficient for the class-imbalanced surveillance dataset. The credibility scoring and blocklisting mechanisms together contribute approximately 2.2 points of accuracy improvement by effectively managing unreliable camera sources.

12. LIMITATIONS

While the proposed system demonstrates strong performance across benchmark datasets, several limitations warrant careful consideration for real-world deployment. System performance is sensitive to input video quality; poor lighting conditions, low sensor resolution, and significant motion blur can degrade feature extraction quality and reduce detection accuracy in ways not fully captured by standard benchmark evaluation. Real-world surveillance environments may present degradation combinations not represented in available training and evaluation datasets.

The system's fixed temporal window segmentation may limit its ability to detect violent events that are either extremely brief (shorter than the minimum window length) or extended across unusually long durations. The ConvLSTM architecture, while effective for spatio-temporal modeling, imposes higher computational demands than purely convolutional alternatives, potentially limiting deployment feasibility on resource-constrained edge hardware commonly found in legacy surveillance installations. Future work should address these limitations through adaptive temporal segmentation and model compression techniques.

Privacy considerations represent an important limitation and ethical concern for any public surveillance system. The proposed system processes personal behavioral information captured without explicit consent, raising questions regarding proportionality, data retention, and potential misuse. Responsible deployment requires robust governance frameworks, transparent operational policies, and technical safeguards including data minimization, access controls, and audit logging to ensure that the system is operated in accordance with applicable privacy regulations and ethical principles.

13. FUTURE WORK

The current system provides a strong foundation for further research and development in automated violence detection. The following directions are identified as priorities for future investigation:

- Multimodal data fusion: incorporation of complementary data modalities including audio signals (raised voices, impact sounds), crowd density estimation from thermal sensors, and environmental context information to improve detection robustness in challenging visual conditions.
- Advanced deep learning architectures: investigation of 3D Convolutional Neural Networks (3D-CNN), Video Swin Transformers, and Graph Neural Networks (GNN) for enhanced spatio-temporal representation learning, potentially capturing longer-range temporal dependencies and more complex interaction patterns.
- Edge deployment optimization: application of model compression techniques including structured pruning, post-training quantization, and knowledge distillation to enable real-time inference on the low-power embedded processors common in existing surveillance hardware.
- Privacy-preserving detection: development of anonymized behavioral detection pipelines — using pose estimation, silhouette extraction, or differential privacy techniques — that identify violence patterns without processing or retaining personally identifiable biometric information.
- Extended abnormal activity detection: generalization of the detection framework to recognize additional threat categories including theft, vandalism, crowd panic, unauthorized zone access, and abnormal object interactions, expanding the system's utility for comprehensive public safety monitoring.
- Smart city system integration: development of standardized interfaces for integration with emergency response dispatch platforms, law enforcement communication systems, and smart city operations centers, enabling automated alert routing and coordinated incident management.
- Continual learning: investigation of online learning strategies enabling the detection model to adapt to deployment-site-specific behavioral patterns and environmental conditions without requiring complete retraining from scratch.

14. CONCLUSION

This paper has presented a comprehensive Violence Detection System using CCTV Cameras in Public Spaces, designed to address the fundamental limitations of human-dependent surveillance monitoring through automated deep learning-based video analysis. The proposed system integrates CNN-based spatial feature extraction, LSTM and ConvLSTM temporal sequence modeling, CBAM attention mechanisms, and a Hybrid Ensemble classification framework to achieve robust and accurate detection of violent human behavior from real-time CCTV video streams.

The incorporation of a Source Credibility Scoring Mechanism and Blocklisting Framework substantially improves system reliability in heterogeneous multi-camera deployment environments by adaptively managing the contributions of individual camera feeds based on their estimated quality and historical detection performance. These components collectively address a practical deployment challenge that is often overlooked in academic violence detection research: the heterogeneity and variability of real-world surveillance network data quality.

Experimental evaluation on multiple standard benchmark surveillance datasets demonstrates that the proposed system achieves an overall accuracy of 96.4%, F1-Score of 95.3%, and AUC of 0.974, representing consistent and statistically significant improvements over CNN-only, CNN+LSTM, ConvLSTM, and 3D-CNN baseline architectures. Ablation studies confirm the meaningful contribution of each system component, including the attention mechanism, focal loss optimization, source credibility scoring, and blocklisting framework.

The system demonstrates strong practical feasibility for real-time deployment within smart city surveillance infrastructure, offering an automated, scalable, and reliable solution for early violence detection and rapid alert generation. The proposed framework makes a meaningful contribution to the development of intelligent public safety systems capable of protecting citizens and supporting emergency responders through timely, accurate, and actionable situational awareness. Future work will focus on expanding detection capabilities, optimizing computational efficiency for edge deployment, ensuring privacy-preserving operation, and extending validation across diverse real-world surveillance environments.

REFERENCES

[1] 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 (AVSS), 2017, pp. 1–6.

[2] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," IEEE Transactions on Image Processing, vol. 26, no. 4, pp. 1992–2004, 2017.

[3] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in Proc. Int. Symp. Neural Networks, 2017, pp. 189–196.

[4] S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan, and K. Jayavel, "A fully integrated violence detection system using CNN and LSTM," International Journal of Electrical and Computer Engineering, vol. 11, no. 4, pp. 3374–3383, 2021.

[5] A. Hanson, K. PNVR, S. Krishnagopal, and L. Davis, "Bidirectional convolutional LSTM for the detection of violence in videos," in Proc. European Conf. Computer Vision Workshops (ECCVW), 2018.

[6] M. Hasan, J. Choi, J. Neumann, A. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 733–742.

[7] Z. Qi, R. Zhu, Z. Fu, W. Chai, and V. Kindratenko, "Weakly supervised two-stage training scheme for deep video fight detection model," in Proc. IEEE Int. Conf. Tools with Artificial Intelligence (ICTAI), 2022.

[8] W. Tan and J. Liu, "Detection of fights in videos: A comparison study of anomaly detection and action recognition," arXiv preprint arXiv:2205.11394, 2022.

[9] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. European Conf. Computer Vision (ECCV), 2018, pp. 3–19.

[10] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 2980–2988.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017.

[12] S. Davila-Montero, J. A. Dana-Le, G. Bente, A. T. Hall, and A. J. Mason, "Review and challenges of technologies for real-time human behavior monitoring," IEEE Transactions on Biomedical Circuits and Systems, vol. 15, no. 1, pp. 2–28, 2021.

[13] S.loffe and C.Szegedy,"Batch Normalization:Accelerating Deep Network Training," in Proc.ICML,2025.

[14] H.Wang and C.Schmid,"Action Recognition with Improved Trajectories," in Proc.IEEE ICCV,2013.

[15] M.S. Ryoo,"Human Activity Prediction: Early Recognition of Ongoing Activities," in Proc.IEEE ICCV,2011.