



Real-Time Violence Detection in Smart Cities Using Lightweight Spatiotemporal Deep Learning Models

Muhammad Ahsan^{1,*}

¹ School of Mathematical Sciences, Jiangsu University, Jiangsu 212013, China

Email: ahsan1826@gmail.com

Received: January 1, 2025 Revised: February 5, 2025 Accepted: May 4, 2025 ★ Corresponding author

ABSTRACT

The rapid development of smart-city infrastructure and the complexity of urban environments increase the need for automated systems that detect violence in surveillance footage. This paper presents a deep-learning architecture for detecting violence and weapon-related activity in CCTV footage using the Smart-City CCTV Violence Detection (SCVD) dataset. MobileNetV2 extracts spatial features from video frames through TimeDistributed layers, and a stacked Long Short-Term Memory (LSTM) network models temporal dependencies. The system processes 15-frame sequences at a resolution of 128×128 pixels. Batch normalization and dropout are used to improve generalization and limit overfitting, while a sigmoid output produces binary classifications. On the SCVD evaluation, the model achieved 99.58% accuracy and a binary cross-entropy loss of 0.0139. The Normal class achieved precision, recall, and F1-score values of 0.99; the Violence class metrics were reported as 1.00 after rounding. These results indicate the potential of compact spatiotemporal models for reliable surveillance and deployment on edge devices such as smart cameras and embedded systems.

Keywords: Violence detection ▪ Smart surveillance ▪ MobileNetV2 ▪ LSTM ▪ CCTV analytics

1. INTRODUCTION

Rapid urbanization places increasing demands on metropolitan areas to protect citizens, particularly in densely populated and high-risk locations. The development of smart urban ecosystems encourages public-safety agencies to use intelligent surveillance for rapid threat identification and response. Closed-Circuit Television (CCTV) networks provide broad coverage, operate as warning systems, and supply data for AI-powered analytics [1]. Deep-learning techniques have expanded their capabilities by enabling automatic identification of violent events and suspicious objects across extensive networks without constant human oversight [2]. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Transformers are used in video understanding, anomaly detec-

tion, action recognition, and spatiotemporal analysis [3, 4, 5]. These data-driven systems can learn complex features and adapt to the behavioral patterns and contextual variation of urban environments, improving on traditional rule-based approaches [6].

High-quality datasets remain a constraint on progress. NTU CCTV-Fights and the Real-Life Violence Situations (RLVS) dataset are two widely used resources for training violence-recognition models [7]. Despite their value, existing datasets can be limited in size, input modalities, and representation of operational conditions. Many videos are recorded on mobile phones, published on YouTube, or staged for dramatic effect rather than captured by fixed-position cameras in active smart cities. This mismatch between training data and deployment conditions restricts the direct application of models in real surveillance systems [8]. Public weapon-identification

datasets also tend to focus on images of common firearms or knives, with less coverage of object types and use scenarios [9]. Improvised weapons such as tools, bottles, bats, and other blunt objects are often absent, limiting a model's ability to generalize to threats that may appear in practical surveillance.

This work introduces the Smart-City CCTV Violence Detection (SCVD) dataset as an open-access benchmark for video-based violence and weapon detection under realistic smart-city surveillance conditions. It uses static-view sequences representative of fixed public-infrastructure cameras [1] and treats potentially harmful handheld objects as a distinct weapon category [10]. The dataset has separate directories for original footage and processed clips, allowing researchers to test supervised and semi-supervised training as well as action classification, temporal event segmentation, and spatiotemporal object detection. Making the data available is intended to support research collaboration and the development of systems that work across heterogeneous urban environments. By connecting the dataset design to CCTV deployment requirements, the study seeks to improve the practical relevance of violence and weapon detection in public-safety applications.

Alongside dataset development, the study evaluates contemporary deep-learning networks for video-based violence recognition. The approach combines convolutional networks for spatial feature extraction with recurrent layers for temporal modeling, balancing accuracy and execution speed for real-time use. MobileNetV2 extracts frame-level spatial features, and LSTM layers process temporal patterns across video sequences. The proposed system is evaluated against VGG16, ResNet50, and EfficientNetB0 as architectural alternatives. TimeDistributed wrappers, normalization, and regularization are used to handle CCTV footage with varied scene complexity and action duration. The comparative analysis is intended to inform architecture selection for scalable surveillance systems.

2. LITERATURE REVIEW

Urbanization and the growing use of CCTV have made intelligent surveillance a core component of smart-city safety. Deep CNNs have improved the rapid identification of violent and weapon-related activity. One study developed a dataset of genuine CCTV footage containing weaponized violence, non-weaponized violence, and non-violent activity. It also proposed combining sequential video frames into salient images that summarize temporal evolution for deep-learning classification. The dataset included videos from public online sources, and evaluations of several DCNN architectures reported accuracy near 99%. The study further examined optimization choices intended to retain computational efficiency and support smart-city deployment [11].

Ramya et al. proposed an EfficientDet-based surveillance method for weapon detection in CCTV footage, addressing the need for automated public-safety monitoring with an object-detection model [12].

To address computational constraints in smart-city surveillance, Khan et al. proposed a two-stream violence detector that combines 3D convolutional processing of background-

suppressed RGB frames with a second stream for optical flow. The model uses 3D depth-wise convolutions to reduce parameter count and was evaluated on RWF2000, RLVS, and Movie Fight. The authors report that it is suitable for edge devices [13].

CrimeNet addresses false positives in video-based violence detection by combining a Vision Transformer with Neural Structured Learning and adversarial training. Evaluation on four challenging violence-related datasets showed improved ROC AUC, while cross-dataset experiments indicated stronger generalization than competing methods [14].

Technical methods are complemented by research into surveillance deployment and its geographic effects. A study of the MiCalle program in Mexico City used a spatial difference-in-differences model to test the local and spillover effects of police-monitored CCTV on delivery robberies. It found no evidence connecting the program with neighborhood delivery-robbery rates, underscoring the need for geographically grounded evaluation rather than assuming that camera placement reduces crime [15].

Machine-learning-based intelligent video surveillance (ML-IVSS) has combined motion detection, pedestrian tracking, and deep learning to detect trespassing, intrusion, criminal behavior, and falls. Feature extraction begins with relevant object regions and then embeds features in a deep neural architecture. On a smart-CCTV dataset, the reported abnormal-behavior accuracy was 98.8%, with 95.6% precision, 96.2% recall, and a 97.1% F1-score [16]. A 2023 survey of smart-city video surveillance reviews system functions, applications, and technical challenges, including the integration of deep learning, edge computing, and cloud computing [17].

Human and institutional factors also affect surveillance outcomes. A study in Gandhinagar, Gujarat, examined CCTV operators' views on crime prevention and concluded that effective citywide security depends on suitable system deployment and continuous operator training. The study was sponsored by the National Forensic Sciences University and the Bureau of Research and Development [18]. For real-time violence detection on resource-limited devices, another study combined a lightweight EfficientNet-B0 frame encoder with LSTM temporal modeling. It reported 86.38% accuracy and 86.39% F1-score and demonstrated deployment on a Raspberry Pi [19].

Video-based violence analysis has also combined localization and action classification. Ha et al. used SORT object tracking, object-region recovery, VGG16 frame features, and recurrent classifiers to identify school violence; their improved SORT-LSTM method achieved 72.13% accuracy [20]. For violence detection, an attention-based BiLSTM was evaluated with newly collected CCTV, YouTube, and public-security footage. On the Hockey Fight dataset, BiLSTM outperformed LSTM in high-speed scenarios, with temporal attention helping the model respond to rapid events [21].

A separate theoretical framework proposed probability-driven video-feed management to improve scalability. Its micro-governance approach uses small edge models to detect events locally and forwards serious incidents to central systems, thereby reducing unnecessary processing while enabling timely responses. The framework draws on the butterfly

effect and sustainable smart-city design to motivate efficient, distributed surveillance [22].

3. DATASET

3.1 Dataset Description

The Smart-City CCTV Violence Detection (SCVD) dataset is intended to address limitations in datasets for surveillance-based violence and weapon detection. Unlike NTU CCTV-Fights and RLVS, SCVD uses static camera views similar to city CCTV installations [1, 7]. It covers peaceful behavior, physical confrontations, and weapon handling. Its weapon category includes potentially harmful handheld objects beyond firearms and knives, such as bottles, bats, and tools [10]. The dataset has separate directories for raw footage and pre-processed clips, enabling experiments in action recognition and anomaly detection. The two directory structures allow researchers to work with the original recordings or with clips prepared for deep-learning workflows, and to compare approaches for supervised classification, semi-supervised learning, temporal event segmentation, and spatiotemporal object detection. This organization is intended to make the benchmark useful beyond a single model architecture or surveillance task.

SCVD contains 399 video files in three classes: 100 weapon videos, 99 violent videos, and 200 non-violent videos. These groups provide examples of dangerous events alongside routine activity and are intended to support training for realistic surveillance distributions. The larger Normal class reflects the relative rarity of weaponized and violent events in typical CCTV footage. Keeping weapon and violence labels separate allows experiments to distinguish object-related threats from physical confrontations, while also enabling joint analysis of object and context patterns. The three-part classification design provides examples for detecting dangerous situations against ordinary surveillance behavior. The dataset's raw-video and processed-clip directories support different training and evaluation workflows, including temporal classification and anomaly detection, and are intended to provide diversity and scalability for smart-city applications.

Figure 1 shows the frame-count distributions across the three classes and the range of clip durations in each group. Violence clips tend to be longer on average than weapon and Normal clips. Normal videos are more numerous, but their durations span a wider range, with many shorter clips, while the weapon and violence groups show distinct ranges of clip lengths. Understanding this variation is important for temporal models such as LSTMs and Transformers: sequence truncation, padding, and sampling strategies depend on clip-length distributions. The frame-count statistics therefore inform preprocessing decisions about temporal windows and fixed-length model inputs, including when to truncate a long sequence, replicate or pad a short one, or sample frames across a clip.

The frame-per-second (FPS) distributions in Figure 2 show that the collected videos are grouped around 30 FPS across the Weapon, Violence, and Normal classes. This consistency simplifies preprocessing by avoiding frame-rate normalization or resampling, which can otherwise alter motion dynamics. It also preserves comparable temporal granularity for

sequence-based models such as LSTMs and Temporal Convolutional Networks. The uniform frame rate is representative of CCTV deployments and helps focus temporal learning on activity rather than differences in video acquisition rates.

3.2 Data Preprocessing

All video samples were standardized to prepare SCVD for deep-learning-based violence and weapon detection [23]. Videos vary in frame count (Figure 1), while their frame rates are similar (Figure 2). Videos were decoded and re-sized to match common CNN input specifications. Original frame rates were preserved because downsampling or interpolation can distort motion dynamics needed for violence detection [24]. Depending on the model, training and evaluation used segments of 16–64 frames. Short clips were extended by frame replication or zero-padding to meet minimum sequence lengths; longer clips were divided into overlapping segments or sampled across their duration to preserve temporal relationships [25]. Mean and standard-deviation normalization was applied to stabilize learning. Training-time augmentation included brightness adjustment, random cropping, and horizontal flipping to improve generalization. Source labels were retained so that clips could support multi-class or multi-label tasks, including weapon detection in violent contexts. This standardized pipeline is intended to provide reliable input features for LSTM and BiLSTM models and remains adaptable to methods requiring temporally structured sequences. Each stage addresses a different source of variation: resizing standardizes spatial dimensions, temporal segmentation standardizes sequence length, and normalization reduces scale differences in the input values. Optional augmentation is applied during training rather than changing the source labels. Figure 3 summarizes the stages from raw video decoding through the creation of the processed clip dataset.

4. RESULTS

This section reports the experimental evaluation of CNN architectures combined with LSTM networks for video-surveillance violence detection. MobileNetV2, ResNet50, VGG16, and AlexNet are compared within a common spatiotemporal framework to represent lightweight, residual, classical, and relatively shallow architectures. The evaluation has three related components: an overview of the network designs, quantitative evaluation metrics, and comparative analysis of training behavior and classification outcomes. Accuracy, loss, class-specific precision, recall, and F1-score are considered together with convergence curves and confusion matrices. These results are used to identify architectures suitable for reliable real-time violence-detection platforms and to explain their relative strengths and weaknesses.

4.1 Deep Learning Models

Violence detection in surveillance footage is challenging because movement must be analyzed across both space and time. The evaluated models combine CNNs for spatial feature extraction with LSTMs for temporal sequence modeling [26].

4.1.1 Architecture Overview

Each model sends video frames through a TimeDistributed CNN encoder, then passes the resulting features through

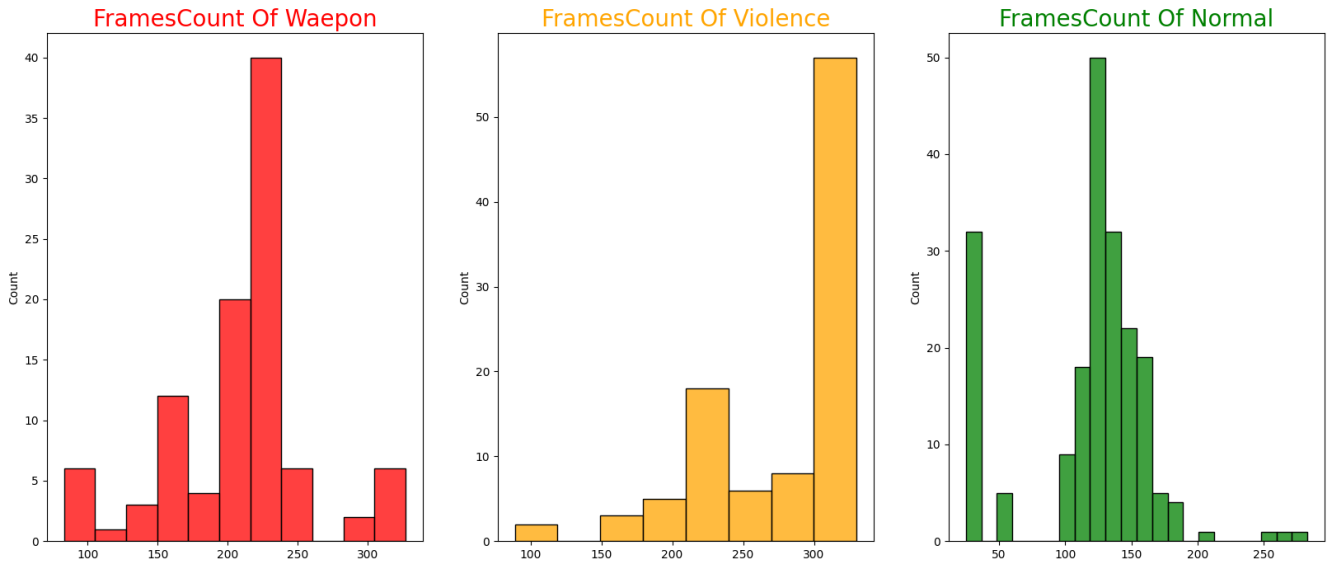


Figure 1. Distribution of frame counts for the Weapon, Violence, and Normal classes in the SCVD dataset.

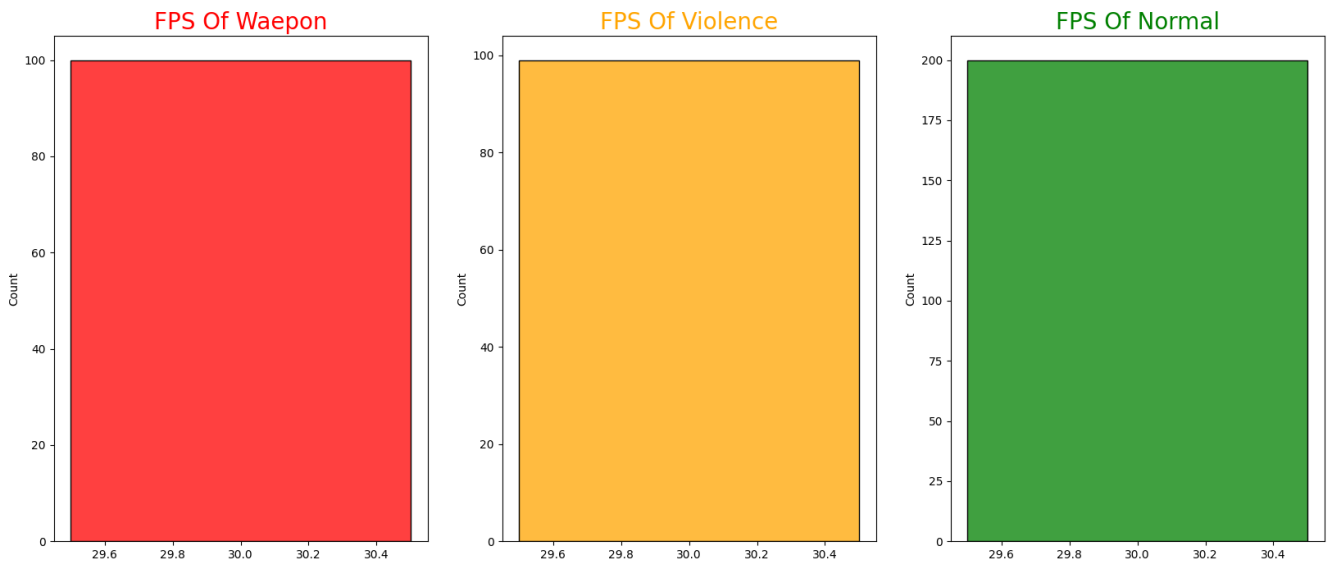


Figure 2. FPS distribution of films in the Normal, Weapon, and Violence classes. The classes maintain a frame rate of about 30 frames per second.

stacked LSTM layers to learn temporal relationships. Dense output units classify a sequence as Violence or Normal, combining appearance information from individual frames with motion information across the clip. CNN backbones are initially frozen using ImageNet-pretrained weights to retain visual representations learned from ImageNet. Selected layers are unfrozen during the final training phase for fine-tuning. Learning rate, batch size, and input resolution are held consistent across models so that the architectural comparison uses a common experimental setup.

4.1.2 Proposed Model: MobileNetV2 + LSTM

The proposed model uses MobileNetV2 for efficient spatial feature extraction. Depthwise separable convolutions and inverted residual blocks reduce the parameter count while retaining representational capacity, making the architecture suitable for edge devices with limited processing power and memory. A TimeDistributed wrapper applies MobileNetV2 separately to each frame in 15-frame sequences at $128 \times 128 \times 3$ resolution. GlobalAveragePooling2D reduces feature-map

size, followed by dropout and batch normalization. Two LSTM layers with 128 and 64 units model temporal patterns in the feature sequence. The classification head uses ReLU activation, batch normalization, and dropout, followed by a sigmoid output representing the probability of violence. The resulting design is intended to balance recognition performance and inference speed for urban security systems.

4.1.3 VGG16 + LSTM

VGG16 contains thirteen convolutional layers followed by three fully connected layers. Its convolutional layers use 3×3 filters, stride 1, ReLU activations, and max-pooling. This deep but comparatively simple design offers strong image representations; however, its 138 million parameters make it computationally demanding and increase training time and overfitting risk on smaller surveillance datasets. In this pipeline, TimeDistributed VGG16 extracts features from individual frames, its fully connected layers are removed, and LSTM layers model temporal patterns from the spatial features. Dropout and batch normalization provide reg-

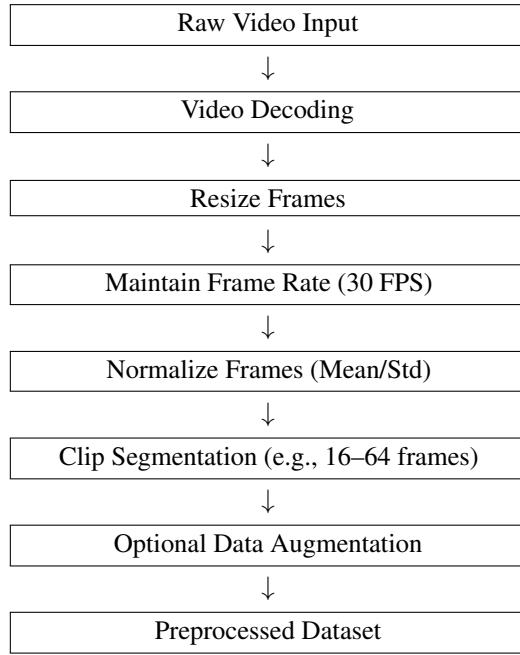


Figure 3. Preprocessing pipeline for SCVD videos. Raw videos are decoded, resized, temporally standardized, optionally augmented, and segmented into fixed-length clips.

ularization. A CCTV surveillance study compared CNN, VGG16-LSTM, ResNet50-LSTM, EfficientNetB0-LSTM, InceptionV3-LSTM, DenseNet121-LSTM, and LRCN models for suspicious-activity recognition [27].

4.1.4 ResNet50 + LSTM

ResNet50 uses shortcut connections for residual learning, helping mitigate vanishing gradients in deep networks. Its 50 layers preserve feature information across blocks. In this pipeline, ResNet50 extracts elaborate hierarchical features from each frame, and stacked LSTM layers analyze temporal information. The architecture provides strong spatial representations and balances depth with feature learning, but requires more computation than MobileNetV2. Regularization is important to control overfitting; batch-normalization and learning-rate settings were adjusted during evaluation. The model achieved high accuracy and generalization in the reported tests, although these outcomes depend on suitable training settings [28].

4.1.5 AlexNet + LSTM

AlexNet was a major CNN architecture in the 2012 ImageNet Large Scale Visual Recognition Challenge. It has five convolutional layers and three fully connected layers, with ReLU activations and dropout. Here, an LSTM is added to model temporal dependencies between video frames. Although AlexNet is a useful historical baseline because it is comparatively simple and helped establish deep learning for image recognition, its relatively shallow design can limit the extraction of complex spatiotemporal features. During evaluation, it showed sensitivity to training-data randomness, unstable validation loss, and substantial misclassification, including false negatives. These limitations motivate deeper or attention-based models for detailed temporal patterns [29].

4.1.6 Temporal Sequence Modeling with LSTM

LSTM layers capture temporal changes and motion patterns in video, which are not represented by static image features alone. The CNN backbones feed frame-level features into two stacked LSTM layers with 128 and 64 units. Their sequential processing models longer-term dependencies and smooth transitions between actions, including sudden gestures, aggressive behavior, and escalation that may not be evident in individual frames. This temporal component is therefore central to the proposed video-based classification approach [30].

4.2 Training and Implementation Details

All models were implemented and trained using TensorFlow with the Keras API. Clips were normalized and converted to 15-frame sequences at 128×128 pixels. Early stopping monitored validation loss, and Adam used a learning rate of 1×10^{-3} . Batch normalization and dropout were applied at different stages to reduce overfitting, and checkpoints were saved at peak validation performance to support generalization. MobileNetV2 + LSTM offered the most favorable combination of reported performance, efficiency, and scalability. ResNet50 also performed strongly in classification but has a larger computational footprint that may reduce its suitability for real-time systems. VGG16 is historically important but expensive to compute, and the reported implementation lacked sufficient regularization. Although EfficientNetB0 is theoretically efficient, it did not generalize as expected in the video experiments. These comparisons motivate a lightweight, well-regularized MobileNetV2 + LSTM configuration for practical smart-city surveillance.

4.3 Evaluation Metrics

Violence detection requires a balance between overall predictive performance and the correct handling of errors that have different operational consequences. Violent events are less common than Normal behavior, so a model that predicts the dominant class can achieve an apparently high accuracy while failing to detect the events of interest. The evaluation therefore uses accuracy, precision, recall, F1-score, and binary cross-entropy loss. Accuracy measures the fraction of correct predictions but should not be read in isolation when classes are imbalanced. Precision measures how often positive predictions correspond to actual violent events and is relevant to the rate of false alarms. Recall (sensitivity) measures how many actual violent events are found and is relevant to missed threats. The F1-score combines precision and recall through their harmonic mean, providing a balanced class-specific measure. Binary cross-entropy loss measures the difference between predicted probabilities and the true labels; lower loss indicates closer predicted probabilities. Confusion matrices summarize true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), showing which errors result from false alarms and which from missed events. This distinction is important in public-safety applications, where both unnecessary alerts and undetected violence carry costs. Table 1 summarizes the measures, formulas, and their intended interpretation.

The metrics are interpreted for both classes rather than only for the aggregate sample. This is particularly important for

Table 1. Evaluation metrics for binary violence classification.

Metric	Definition	Formula
Accuracy	Proportion of correct positive and negative predictions among all predictions.	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	Proportion of predicted positive instances that are positive; higher precision reduces false alarms.	$\frac{TP}{TP + FP}$
Recall (Sensitivity)	Proportion of actual violent events correctly identified; higher recall reduces missed threats.	$\frac{TP}{TP + FN}$
F1-Score	Harmonic mean of precision and recall, balancing the two measures.	$\frac{2 \text{ Precision Recall}}{\text{Precision} + \text{Recall}}$
Loss	Binary cross-entropy measures the divergence between predicted probabilities and ground-truth labels; lower values are better.	–

Table 2. Performance comparison of the evaluated models. Class 0 denotes Normal and class 1 denotes Violence.

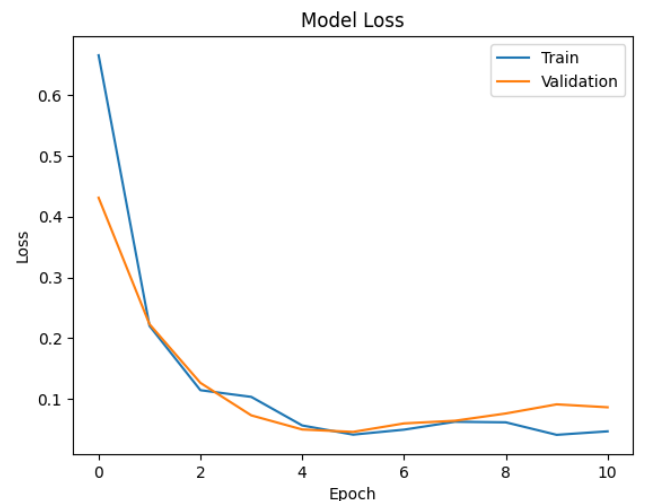
Model	Accuracy	Loss	Class	Precision	Recall	F1-score	Support
VGG16 + LSTM	0.5325	0.0099	0	0.36	0.40	0.38	169
			1	0.65	0.61	0.63	308
ResNet50 + LSTM	0.9790	0.0678	0	0.94	1.00	0.97	169
			1	1.00	0.97	0.98	308
MobileNetV2 + LSTM	0.9958	0.0139	0	0.99	0.99	0.99	169
			1	1.00	1.00	1.00	308
AlexNet + LSTM	0.6436	0.6507	0	0.50	0.30	0.37	169
			1	0.68	0.83	0.75	308

a surveillance setting in which the Normal class is more frequent, but successful detection of the smaller Violence class is a principal safety objective. Precision and recall provide complementary information: increasing sensitivity can increase false alerts, whereas a classifier that avoids false alerts may miss genuine incidents. Reporting class-wise values alongside the confusion matrix makes these trade-offs visible and supports more informed deployment decisions.

4.4 Performance Evaluation and Comparative Analysis

Four CNN architectures with LSTM temporal modeling were trained and evaluated under a common setup on SCVD using labeled Violence and Normal clips. Despite shared hyperparameters and preprocessing, the models differed in learning behavior, convergence, generalization, and class-identification capability. The comparative analysis uses training and validation loss and accuracy, class-wise precision, recall, and F1-score, and confusion-matrix results to characterize these differences. Table 2 reports overall accuracy and loss together with class-specific scores and support. The source table's overall-accuracy entries for VGG16 and AlexNet conflict with their displayed class-wise values and confusion matrices; those two accuracies are recalculated from the confusion-matrix counts below. MobileNetV2 + LSTM achieved 99.58% accuracy and a loss of 0.0139, with high class-wise scores. Its depthwise separable convolutions reduce the computational burden while its LSTM captures relationships across video frames, providing an effective combination of spatial and temporal representation.

The table shows that the tested architectures differ in both aggregate performance and class-wise behavior. MobileNetV2 + LSTM reports the strongest combination of low loss and high scores for the two classes. ResNet50 + LSTM provides high scores with a more substantial computational cost. VGG16 + LSTM and AlexNet + LSTM show weaker class-wise results in the displayed validation matrices, despite the different

**Figure 4.** Training and validation loss for MobileNetV2 across 10 epochs.

overall-accuracy values printed in the source table; the converted table uses matrix-derived overall accuracy for those two models. The learning curves and confusion matrices that follow provide additional evidence about convergence and errors for each architecture.

The MobileNetV2 training and validation losses decrease rapidly during the initial epochs (Figure 4). The curves remain close and roughly parallel, and validation loss stays low during later training, suggesting limited memorization of the training set. Such stability is important in video classification, where temporal noise can otherwise produce unstable predictions. Training and validation accuracy also rise smoothly, reaching nearly 100% by the end (Figure 5).

ResNet50 shows a downward training-loss trend, with small validation-loss fluctuations beginning around epoch 7 (Figure 6). The irregular changes suggest that batch-level varia-

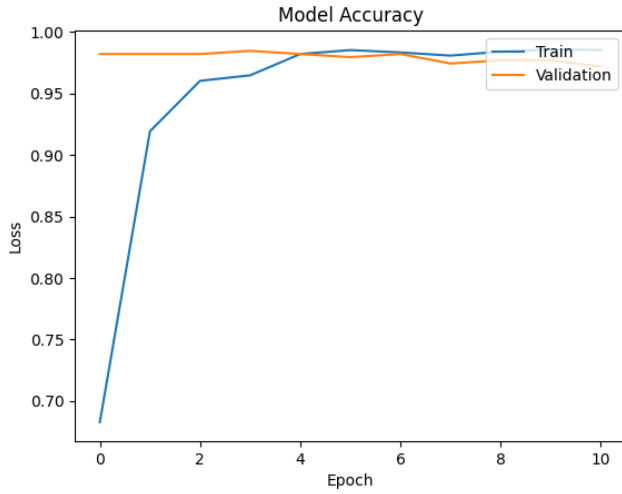


Figure 5. Training and validation accuracy for MobileNetV2 across 10 epochs.

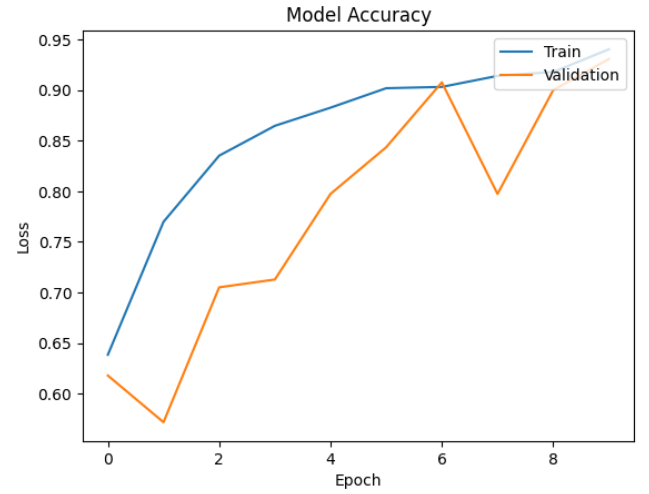


Figure 7. Training and validation accuracy for ResNet50 across 9 epochs.

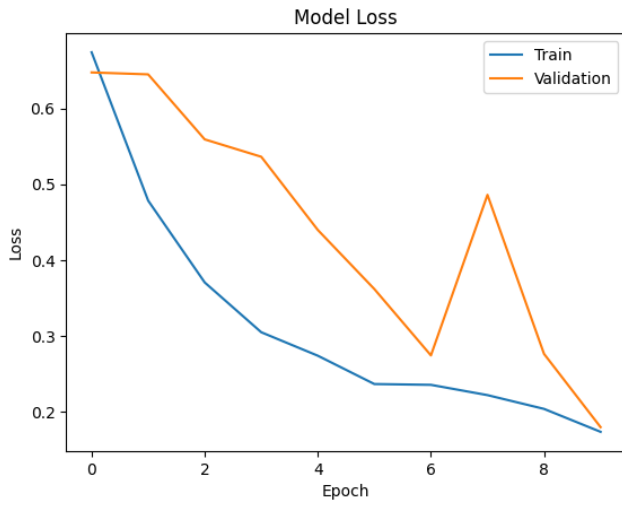


Figure 6. Training and validation loss for ResNet50 across 9 epochs.

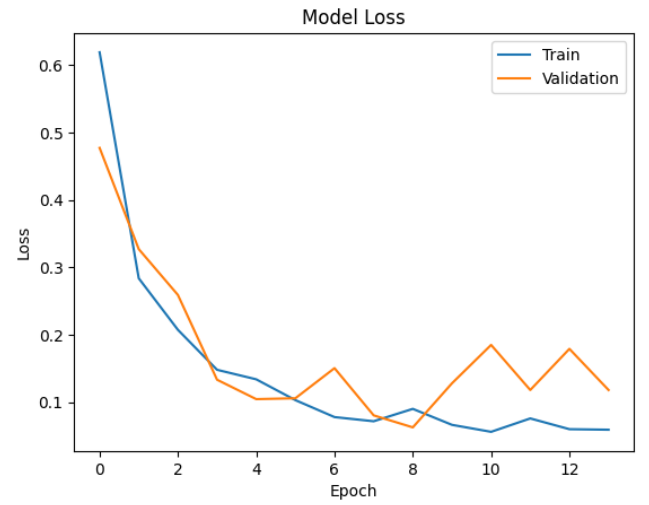


Figure 8. Training and validation loss for VGG16 across 13 epochs.

tion and regularization settings can affect the deeper network. Its validation accuracy remains above 95% throughout the reported epochs (Figure 7), indicating strong feature extraction under the evaluated regularization settings.

VGG16 exhibits signs of overfitting from approximately epoch 6: training loss declines while validation loss fluctuates upward (Figure 8), indicating insufficient generalization. Its validation accuracy remains relatively stable (Figure 9), although the model's class-wise scores indicate missed Violence cases. The conservative behavior and lack of normalization layers described for this configuration may reduce its response to uncertain video segments.

AlexNet has the weakest results. Its loss remains elevated and validation loss fluctuates (Figure 10), indicating unstable optimization. Accuracy stays below 60% for much of training and varies sharply (Figure 11). Its limited architecture and feature-flow design reduce its ability to represent complex temporal patterns and subtle indicators of violence.

Figure 12 shows the validation confusion matrices. MobileNetV2 correctly classified 168 of 169 Normal samples and 307 of 308 Violence samples. ResNet50 correctly classified all 169 Normal samples and 298 of 308 Violence samples. VGG16 correctly classified 67 Normal and 187 Violence sam-

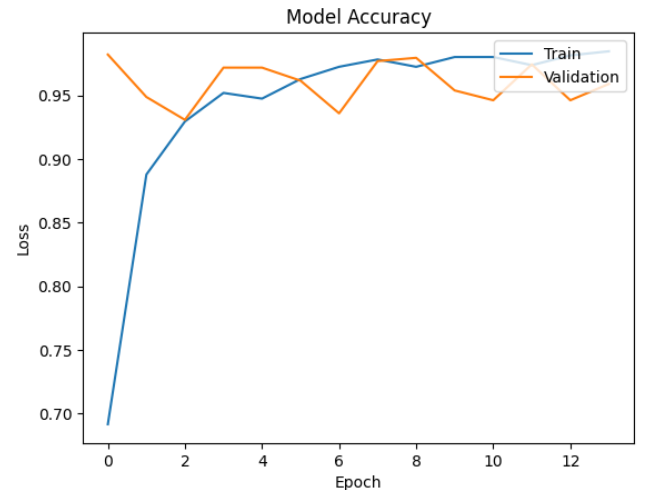


Figure 9. Training and validation accuracy for VGG16 across 13 epochs.

ples, while AlexNet correctly classified 51 Normal and 256 Violence samples.

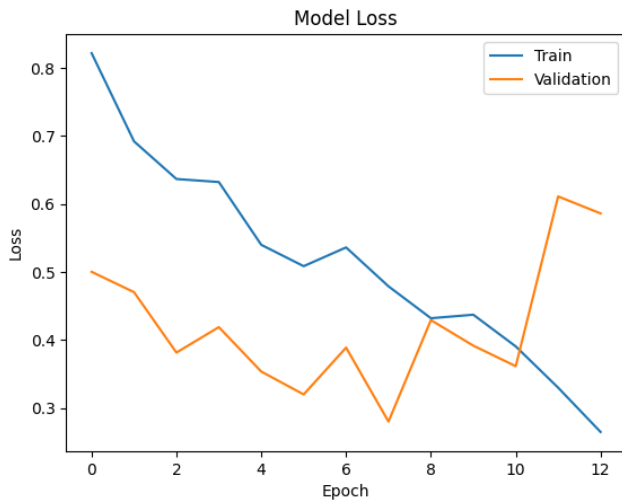


Figure 10. Training and validation loss for AlexNet across 19 epochs.

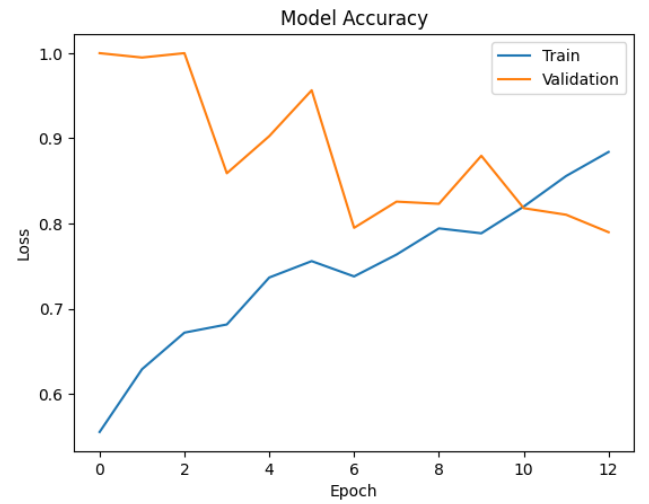
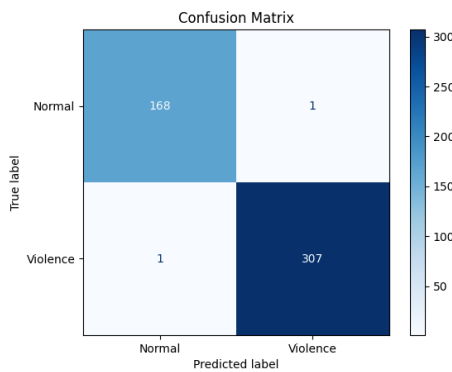
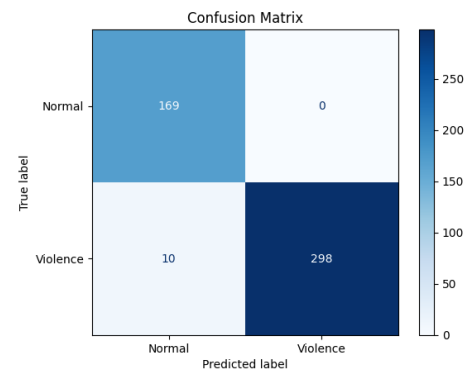


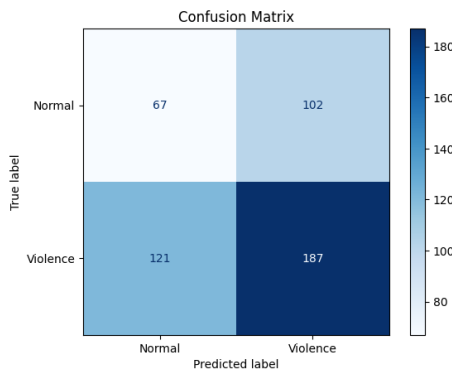
Figure 11. Training and validation accuracy for AlexNet across 19 epochs.



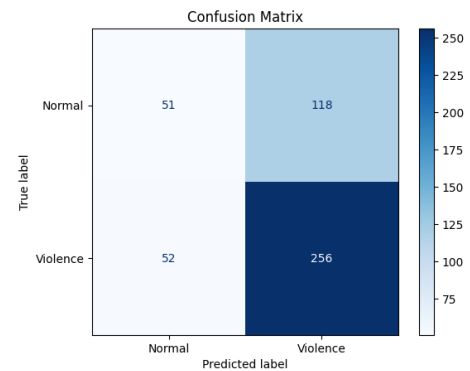
(a) MobileNetV2



(b) ResNet50



(c) VGG16



(d) AlexNet

Figure 12. Confusion matrices for each model on the validation set for the Normal and Violence classes.

5. DISCUSSION

The SCVD evaluation shows that architecture affects convergence, classification, and generalization. All four models were trained under the same experimental setup, yet their learning curves and confusion matrices reveal different balances between spatial representation, temporal modeling, and computational cost. The results therefore provide both a technical comparison and an indication of which configurations may suit operational surveillance, where missed events and false alarms have different consequences.

Under the reported training settings, MobileNetV2 has high validation accuracy and a smooth loss curve. Its training and

validation curves remain close, indicating stable optimization, and its low loss is consistent with strong predictive performance. Its confusion matrix, however, shows one misclassification in each class rather than perfect classification. The matrix and the rounded per-class scores indicate that it detects almost all Violence samples while also correctly recognizing nearly all Normal samples. This balance is important for a compact model intended for edge-based operation. Its depthwise separable convolutions offer a lower-computation spatial encoder, while stacked LSTM layers provide temporal context that cannot be recovered from a single static frame.

ResNet50 is stable across training and correctly classifies

all Normal samples, with ten Violence samples classified as Normal in the displayed validation matrix. Its residual connections preserve feature information through deep layers, supporting robust spatial and temporal modeling. The loss decreases steadily, and its validation accuracy remains high, although its larger architecture increases resource demands compared with MobileNetV2. Thus, it provides a dependable alternative where computational capacity is available and where stable class-wise behavior is prioritized over the smallest deployment footprint.

VGG16 identifies some Normal samples but misses many Violence cases in its confusion matrix. The training loss decreases while validation loss fluctuates, particularly after the early epochs, suggesting overfitting and limited generalization. Its relatively stable validation-accuracy curve does not remove the class-specific weakness shown by the confusion matrix. The model's large parameter count and computational cost further limit its practicality for low-latency surveillance. AlexNet generalizes poorly and shows unstable validation loss and accuracy. Its lower representational depth is not sufficient to capture complex spatial-temporal dependencies in these sequences, and its false-negative errors make it less suitable for safety-sensitive detection.

The comparison reinforces the need to match architecture complexity to both the data and operational task. Overall accuracy by itself does not reveal how a system handles rare violent events; class-wise precision, recall, and F1-score should be considered, particularly for the Violence class. In a public-safety setting, an undetected violent incident may be more consequential than a false alarm that triggers human review. Deployment decisions should therefore make the acceptable error trade-off explicit. ResNet50 offers stable class-wise performance in the reported matrices, while MobileNetV2 remains attractive for edge deployment because of its computational efficiency. Class imbalance remains a concern and may be addressed with class weighting or focal-loss strategies.

Future improvements could use metaheuristic hyperparameter optimization, including Genetic Algorithms, Particle Swarm Optimization, and Bayesian Optimization, to automate parameter selection and reduce lengthy manual tuning. Ensemble methods could combine model predictions to improve robustness. Additional input modalities, including audio, thermal imagery, and sensor signals, could broaden the available evidence and improve situational awareness in difficult surveillance environments. Overall, reliable violence detection requires more than a high-capacity network: model selection, task-specific optimization, class-wise evaluation, and deployment priorities must be considered together.

6. CONCLUSION

This paper presents a deep-learning framework for violence detection in smart-city surveillance using the Smart-City CCTV Violence Detection (SCVD) dataset. Lightweight CNN backbones combined with LSTM layers model spatial and temporal features in fixed-camera CCTV video. The MobileNetV2 + LSTM configuration achieved 99.58% accuracy with a loss of 0.0139 and high reported class-wise metrics, while providing a compact architecture with potential for edge deployment. The comparative evaluation also

shows that model behavior varies by class and architecture: ResNet50 is stable, VGG16 misses a substantial number of violent cases, and AlexNet performs poorly on temporal patterns. These results underline the importance of evaluating spatiotemporal models through more than a single overall score.

Future work should address multi-class categorization and anomaly detection, improve robustness under dim lighting and occlusion, and incorporate additional data such as audio and multiple camera views. Attention-based mechanisms, temporal Transformers, evolutionary hyperparameter optimization, and ensemble learning may further improve generalization and interpretability. The practical use of deep-learning models in smart-city surveillance depends on both efficient computation and an explicit understanding of class-wise errors. The results support such models as a component of proactive public-safety infrastructure, while deployment decisions should account for the risks associated with missed detections and false alarms.

REFERENCES

- [1] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveillance," *Sensors*, vol. 22, no. 6, p. 2216, 2022.
- [2] F. U. M. Ullah, M. S. Obaidat, A. Ullah, K. Muhammad, M. Hijji, and S. W. Baik, "A comprehensive review on vision-based violence detection in surveillance videos," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–44, 2023.
- [3] J. Wensel, H. Ullah, and A. Munir, "ViT-ReT: Vision and recurrent transformer neural networks for human activity recognition in videos," *IEEE Access*, vol. 11, pp. 72227–72249, 2023.
- [4] M. Xu, Y. Xiong, H. Chen, X. Li, W. Xia, Z. Tu, and S. Soatto, "Long short-term transformer for online action detection," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 1086–1099, 2021.
- [5] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee, and S. W. Baik, "TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106173, 2023.
- [6] X. Wu, R. Wang, J. Hou, H. Lin, and J. Luo, "Spatial-temporal relation reasoning for action prediction in videos," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1484–1505, 2021.
- [7] W. Tan and J. Liu, "Detection of fights in videos: A comparison study of anomaly detection and action recognition," in *European Conference on Computer Vision*, pp. 676–688, Springer, 2022.
- [8] H.-T. Duong, V.-T. Le, and V. T. Hoang, "Deep learning-based anomaly detection in video surveillance: A survey," *Sensors*, vol. 23, no. 11, p. 5024, 2023.
- [9] S. Narejo, B. Pandey, D. Esenarro Vargas, C. Rodriguez, and M. R. Anjum, "Weapon detection using YOLO v3

- for smart surveillance system,” *Mathematical Problems in Engineering*, vol. 2021, no. 1, p. 9975700, 2021.
- [10] P. Y. Ingle and Y.-G. Kim, “Real-time abnormal object detection for video surveillance in smart cities,” *Sensors*, vol. 22, no. 10, p. 3862, 2022.
- [11] T. Aremu, Z. Li, R. Alameeri, M. Aloqaily, and M. Guizani, “Towards smart city security: Violence and weaponized violence detection using DCNN,” preprint, arXiv, 2022.
- [12] R. Ramya, C. Lasya, N. M. Sai, and S. Paneerselvam, “An intelligent surveillance system for weapon detection based on EfficientDet algorithm,” in *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 453–459, IEEE, 2023.
- [13] M. Khan, W. Gueaieb, A. El Saddik, G. De Masi, and F. Karray, “An efficient violence detection approach for smart cities surveillance system,” in *2023 IEEE International Smart Cities Conference (ISC2)*, pp. 1–5, IEEE, 2023.
- [14] F. J. Rendón-Segador, J. A. Álvarez-García, J. L. Salazar-González, and T. Tommasi, “CrimeNet: Neural structured learning using vision transformer for violence detection,” *Neural Networks*, vol. 161, pp. 318–329, 2023.
- [15] C. Vilalta, P. Lopez-Ramirez, and G. Fondevila, “Testing the local and spatial spillover effects of police monitored CCTV systems on crime,” *Applied Geography*, vol. 151, p. 102873, 2023.
- [16] M. A. J. Maktoof, I. H. M. Ibrahim, M. A. A. Razzaq, A. Abbas, and A. Majdi, “Machine learning-based intelligent video surveillance in smart city framework,” *Fusion: Practice & Applications*, vol. 11, no. 2, pp. 35–47, 2023.
- [17] Y. Myagmar-Ochir and W. Kim, “A survey of video surveillance systems in smart city,” *Electronics*, vol. 12, no. 17, p. 3567, 2023.
- [18] S. Mathur and K. Sood, “Policing perspective on preemptive and probative value of CCTV architecture in the security of the smart city—gandhinagar, gujarat,” *International Journal of Electronic Security and Digital Forensics*, vol. 15, no. 3, pp. 252–258, 2023.
- [19] M. S. N. B. Abdullah, H. Abdul Karim, and N. Al-Dahoul, “A combination of light pre-trained convolutional neural networks and long short-term memory for real-time violence detection in videos,” *International Journal of Technology*, vol. 14, no. 6, pp. 1228–1236, 2023.
- [20] N. D. Ha, N. Y. Tran, L. N. L. Thuy, I. Shimizu, and P. T. Bao, “Violence region localization in video and the school violent actions classification,” *Frontiers in Computer Science*, vol. 5, p. 1274928, 2023.
- [21] P. Wang, P. Wang, and E. Fan, “Violence detection and face recognition based on deep learning,” *Pattern Recognition Letters*, vol. 142, pp. 20–24, 2021.
- [22] M. Patel, “Real-time violence detection using CNN-LSTM,” preprint arXiv:2107.07578, arXiv, 2021.
- [23] C. El Morr, M. Jammal, H. Ali-Hassan, and W. El-Hallak, “Data preprocessing,” in *Machine Learning for Practical Decision Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics*, pp. 117–163, Springer, 2022.
- [24] S. Alam and N. Yao, “The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis,” *Computational and Mathematical Organization Theory*, vol. 25, no. 3, pp. 319–335, 2019.
- [25] A. A. Khan, “Balanced split: A new train-test data splitting strategy for imbalanced datasets,” preprint arXiv:2212.11116, arXiv, 2022.
- [26] A. Mosavi, S. Ardabili, and A. R. Varkonyi-Koczy, “List of deep learning models,” in *International Conference on Global Research and Education*, pp. 202–214, Springer, 2019.
- [27] D. Saluja, H. Kukreja, A. Saini, D. Tegwal, P. Nagrath, and J. Hemanth, “Analysis and comparison of various deep learning models to implement suspicious activity recognition in CCTV surveillance,” *Intelligent Decision Technologies*, vol. 17, no. 4, pp. 917–942, 2023.
- [28] C. Shripriya, J. Akshaya, R. Sowmya, and M. Poonkodi, “Violence detection system using ResNet,” in *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1069–1072, IEEE, 2021.
- [29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
- [30] S. Zargar, “Introduction to sequence learning models: RNN, LSTM, GRU,” tech. rep., Department of Mechanical and Aerospace Engineering, North Carolina State University, 2021.