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

Abstract

Smart city infrastructure development and urban environment complexity increase the need for automated systems that detect violence immediately in surveillance footage. The current CCTV system depends on human operators, which becomes impractical when quick response times are mandatory for extensive deployment domains. This research develops a deep learning architecture that proposes automated detection methods for violence and weapon activities in practical CCTV surveillance through the Smart-City CCTV Violence Detection (SCVD) dataset. The system uses MobileNetV2 as its basic convolutional framework, which can extract spatial frame patterns through TimeDistributed layers from video sequence inputs. The features move to a stacked Long Short-Term Memory (LSTM) network to extract the temporal-based dependencies within violent actions. The system processes video sequences with 15 frames while maintaining a pixel size of 128 × 128 to achieve operational efficiency and representational capability. Regularization techniques Batch Normalization and Dropout are used in every part of the network to improve generalization capability and limit overfitting. The pipeline finishes through dense layers linked in full connection, followed by a sigmoid activation function to achieve binary outputs. The experiments on the SCVD dataset resulted in highly positive outcomes. Evaluation of the model produced a 99.58% accuracy rate together with a minimal cross-entropy loss amounting to 0.0139. This model monitoring system demonstrated exceptional performance metrics because the standard class achieved 0.99 precision and 0.99 recall alongside 0.99 F1-score, and the violent class received a perfect score of 100 on every metric. The model proves effective for detecting and classifying violent activities with excellent reliability under diverse and complex surveillance settings. The research shows that real-time deployment of deep learning models in intelligent city surveillance can be accomplished using robust, compact solutions. The system design incorporates spatial along with temporal feature methodologies thus making it suitable for deployment on edge devices such as smart cameras and embedded systems. Through its work on uniting academic models with practical deployment, this study helps create safer urban environments by developing AI-driven public safety technologies.

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

1 Introduction

Global population urbanization at high speed creates intense demands for metropolitan areas to protect their citizens especially in areas where both density and risks are high. Smart urban ecosystems development forces public security agencies to use intelligent surveillance technology for fast threat identification and evaluation to generate immediate responses. The foundation of this surveillance network depends on wide-ranging CCTV technology which functions both as warning systems and provides AI-powered analytics with their data sources [1]. The developments of deep learning techniques have increased video surveillance capabilities by developing automatic systems that identify violence occurrences detect suspicious objects while monitoring

DOI: https://doi.org/10.54216/JAIM.090202

Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

activities in extensive metropolitan networks without constant human oversight [2]. Various video understanding tasks benefit from CNN-based models as well as RNNs and LSTMs and Transformers which proved their effectiveness in anomaly detection and action recognition and spatiotemporal analysis [3–5]. These systems surpass traditional rule-based approaches by learning data-derived complex features and they easily adapt to urban environments with their various behavioral patterns and contextual variations [6]. High-quality datasets represent an ongoing constraint for implementing algorithm development progress. The two major datasets used for learning model training concerns NTU CCTV-Fights alongside the Real-Life Violence Situations (RLVS) dataset [7]. Despite their usefulness these datasets normally present substantial restrictions when it comes to dataset size and number of input modalities and real world context. These datasets feature mobile phone and Youtube video recordings and theatrical settings instead of using fixed-position CCTV cameras installed in operational smart cities. Their limited application in real-world surveillance practice becomes restricted because of this integration mismatch between training models and operational systems [8]. The current datasets show weaknesses because they present limited class representations particularly for weapon detection purposes. The majority of weapon identification datasets available to the public are images which mainly feature regular firearms or knives while showing limited variety among object types alongside usage scenarios [9]. The existing datasets fail to include improvised weapons that officers encounter in practical surveillance situations such as tools, bottles, bats and blunt objects. The model cannot develop generalized threat recognition ability because of these limitations. The work offers the Smart-City CCTV Violence Detection (SCVD) dataset which serves as an open-access benchmark to facilitate research into video-based violence detection along with weapon detection under realistic smart city surveillance circumstances. The SCVD dataset perfectly reproduces the perspective of smart city CCTV systems through static-view video sequences which represent typical fixed camera installations found in public infrastructure [1]. SCVD introduces a separate weapon classification beyond typical labels by recognizing any handheld item that could harm as a dangerous category [10]. The dataset includes two primary directories which separate surveillance simulation video clips into original format and processed clip segments that support supervised model training and semi-supervised approaches. The two data formats enable researchers to perform flexible testing of different model architectures and evaluation approaches including action classification with temporal event segmentation and spatiotemporal object detection. This research makes the SCVD dataset accessible to the public to enhance innovation and research collaboration which will help develop AI systems that operate efficiently among the heterogeneous smart city conditions. The work creates important missing knowledge by connecting dataset creation to surveillance system specifications which enhances application reliability across public safety areas. The study presents both the creation of SCVD dataset and evaluation of different contemporary deep learning networks for video-based violence recognition. This method combines modest convolutional neural networks with the recurrent layers for temporal modeling which provides ideal performance between accuracy level and execution speed requirements for real-time systems. The spatial features are extracted by MobileNetV2 before LSTM layers process temporal patterns found across video sequences. The proposed model receives performance evaluation by testing with popular architectures VGG16, ResNet50, and EfficientNetB0 as benchmarks. The models are adjusted with TimeDistributed layers combined with normalization and regularization to normalize real-world CCTV footage with diverse scene complexity and action durations. The comparative study reveals complete knowledge about finding suitable architecture designs for scalable and smart surveillance applications.

2 Literature Review

Urbanization continues to escalate rapidly in smart cities, yet the safety and security of citizens rely heavily on intelligent surveillance systems. Urban environments now heavily rely on Closed-Circuit Television (CCTV) systems as core systems for space monitoring because they prevent illegal behavior. DCNNs within artificial intelligence have evolved to improve surveillance system utility because they enable quick identification of violent and weaponized behaviors. Modern work introduced a dependable detecting method of weapons and violence using DCNNs, bringing significant advancements to innovative surveillance research. The presented dataset consists of genuine CCTV footages, which include examples of weaponized violence as well as non-weaponized violence alongside non-violent activities. The YouTube videos used in this dataset displayed actual scenarios because they all came from public domains rather than artificial settings. The main breakthrough of this research involves combining sequential video frames into single salient images that summarize temporal evolution into static deep learning-ready images. The effectiveness of this approach reached

DOI: https://doi.org/10.54216/JAIM.090202
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

very high accuracy levels at 99% across different DCNN architecture implementations. The research examines different factor optimizations to sustain computational effectiveness, allowing deployment in smart city infrastructures [11]. The authors of another study present a complex system architecture that aims to boost emergency responses and real-time criminal identification purposes [12]. The proposed framework unites a stable diffusion model to synthesize data with YOLO v7 for detecting dangerous objects such as knives and pistols and MediaPipe for action analysis. A Long Short-Term Memory network is essential to this system because it models how detected objects create behavior patterns over time. The complete system functions from an edge device containing a dash camera to perform real-time testing and alerting. The system achieved 89.5% mean average precision for violent object detection together with 88.33% accuracy from the LSTM classification network as its core indicators for real-time threat detection in innovative city environments. Many present-day intelligent video surveillance systems continue using centralized solutions, resulting in substantial computational expenses, high delays, and network bandwidth constraints. A research paper presents LAVID as a decentralized video surveillance system that operates efficiently with minimal weight and cost-effectiveness [13]. The LAVID surveillance system applies Depthwise Separable Convolutional Neural Networks (DSCNN) and Bidirectional Long Short-Term Memory (BiLSTM) algorithms, which run directly on smart cameras supported by Raspberry Pi backup platforms. Under power and computational constraints, LAVID works to automatically locate violent incidents in real-time with no human intervention required. The LAVID system achieved results comparable to state-of-the-art methods through evaluations run on public datasets while operating on smart cameras with limited power resources. Its capability to function with other devices through Internet of Things (IoT) standards and its ability to interoperate with smart city infrastructure systems makes the LAVID system highly effective for expanded smart city integration. Research on scalable solutions now involves modifying extensive pre-trained models to identify violence incidents within surveillance setups. The proposed framework has developed a dual-path approach with spatial and motion paths to overcome challenges in fine-tuning cost, temporal modeling and inference overheads [14]. The system applies parameter-efficient fine-tuning strategies that work effectively with ViT-based pre-trained models, delivering contemporary model updates and real-time model inference capabilities. The motion path solves computational challenges of temporal processing by reducing sequential frame input into a single image, boosting efficiency. According to experimental tests conducted on five datasets, the framework demonstrates superiority, providing an efficient solution for implementing violence recognition within intelligent city surveillance networks. The study of actual deployment patterns and the socio-geographic effects of surveillance systems need equal priority to technological advancements. A study analyzed the CCTV installation-crime pattern connection using GIS-based methods, including the Optimized Hot Spot Analysis (OHSA) along with Ordinary Least Squares (OLS) regression analysis to measure surveillance effects on criminal rates [15]. Operation of CCTV systems that include AI-based face recognition reduced crime activities in monitored areas. The research established a 43% relationship between CCTV indicators, which decreased crime rates in locations using AI-powered systems. The study demonstrated crime redirection to locations beyond the observed region covered by CCTV surveillance equipment. These studies' outcomes demonstrate the successful impact of CCTV installations for smart cities and identify population-wide deployment requirements to build community safety infrastructure. The use of ML-IVSS and intelligent fusion techniques marks an innovative method according to [16]. The system achieves detection capabilities through motion detection along with pedestrian tracking and deep learning to recognize unusual behaviors such as trespassing activity, intrusion attempts, criminal behavior and situations involving falls. Feature learning begins when the surveillance system detects important object regions before extracting basic features for deep neural architecture embedding. The smart CCTV dataset evaluation revealed performance excellence attained through these methods, which resulted in 98.8% abnormal behavior detection accuracy along with precision (95.6%) and recall (96.2%), reaching 96.2% and F1-score (97.1%). Urban safety monitoring benefits tremendously from innovative environments through this combined sensing and learning technique approach. A systematized review analyzed 213 scholarly publications between 2001 and 2023 to synthesize technological advancements, research challenges, and current trends within intelligent city surveillance as demonstrated by LR07 [17]. The extensive research demonstrates how video surveillance is an essential foundation for all smart city domains that use crime detection, anomaly monitoring, and fire detection while preventing tampering activities. Through analysis, the review establishes a standardized approach to video-based smart city surveillance, a state-of-the-art technique classification, an existing surveillance dataset comparison, and unresolved limitation identification. The paper establishes research prospects for upcoming smart cities by examining real-time flexibility, standardized systems, and integrated intelligence. Study respondents in Gujarat's city evaluated CCTV operators' views about security systems for crime prevention detection in a human-focused examination [18]. Research findings aligned with the conclusion that citywide security depends on effective CCTV systems, and both strategic deployment improvements and continuous operator training need further development. The National Forensic Sciences

DOI: https://doi.org/10.54216/JAIM.090202
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

University and the Bureau of Research and Development sponsored the study, thus securing its institutional value. Current research analyzes the performance spectrum among 3D CNNs, VGG networks, and MobileNet models integrated with LSTMs and Multi-Stream Networks [19]. This work advances beyond previous violence identification systems by employing CCTV-based datasets that boost authentic security functionality. The system solution provides practical deployment guidance through performance evaluation beyond analytics for memory utilization, prediction speed testing, and infrastructure examination. The proposed system models reached a maximum accuracy rate of 98.46% in evaluations, which indicates integrated resource-aware architectures possess strong potential for urban surveillance network-based violence detection. Different research examines brute force detection and facial recognition through CNNs and trajectory features according to [20]. The researchers created two brand new CNN models named Multi-foot Input CNN and SPP-based CNN to enhance low-resolution facial recognition performance. A combination of these evaluation methods applied to the Crowd and Hockey datasets reached accuracy marks of 92% and 97.6%, proving their suitability for real-time surveillance monitoring systems. Detecting violent activities receives attention through LSTM and BiLSTM architecture modifications with attention layers in research [21]. The model used data from newly built CCTV footage and YouTube videos alongside public security recordings to achieve successful results in crowd annotations. The Hockey Fight dataset evaluation confirmed that BiLSTM surpasses both LSTM models by excelling in demanding high-speed dramatic scenarios mainly because temporal attention enables better system response time. A theoretical framework within scalable surveillance brings forward probability-driven computation to enhance video feed management efficiency [22]. The reduction of global violence over the past 40 years does not diminish the problem of unobserved aggressive behaviors because multiple real-time video monitoring is difficult to scale. The research establishes micro-governance as a new method that lets small edge models detect situations locally but efficiently sends serious incidents to central network systems. The methodology draws inspiration from the butterfly effect to lower unnecessary processing needs while enabling real-time responses corresponding to sustainable smart city design principles.

3 Dataset

Dataset Description

The Smart-City CCTV Violence Detection (SCVD) dataset represents a new accessible benchmark that solves problems found in previous datasets designed for video surveillance-based violence and weapon detection. SCVD distinguishes itself from previous datasets NTU CCTV-Fights and RLVS by using static, fixed camera perspectives due to similarities with city-based CCTV system camera positions [1, 7]. The dataset requires users to process a variety of human encounters that span from peaceful conduct to physical confrontations, together with a specialized class focused on weapon handling. The SCVD expands weapon classification beyond firearms and knives to include any handheld harmful object regardless of its nature - bottles, bats or tools - as identified in [10]. The database consists of two directory systems: raw video footage resides in one, and pre-processed deep learning-friendly clips exist in another. The organizational structure allows researchers to train and evaluate models through this platform for different surveillance applications such as action recognition and anomaly detection. Multi-class classification SCVD offers a practical, intelligent surveillance resource with scalability and diversity to develop security systems that match smart city operational needs. The SCVD database contains 399 video files distributed among three fundamental groups to build training models with balanced representation. The dataset consists of three specific sections with 100 weapon videos alongside 99 violent videos and 200 non-violent clips. The three-part dataset distribution system enables researchers to build models to identify dangerous situations from regular surveillance. The bigger normal class considers the actual distribution of occurrences in real CCTV systems, which show weaponized and violent occurrences rarely happening about typical routine observations. SCVD separates violence and weapon classes to enable detailed classification methods and joint multi-label threat analysis of contextual and object patterns. Figure 1 presents histograms that display the frame distribution per video across the weapon, violence and standard categories. The visualizations provide evidence of the diverse time durations between different samples. On average, the videos within the violence category extend further in sequence duration compared to videos in the weapon and standard categories. The videos in the regular class comprise more significant numbers than the other classes yet show a wider range of clip durations since most clips fall below those containing violence or weapons. The knowledge about frame dimension at this level is essential for temporal modeling applications

DOI: https://doi.org/10.54216/JAIM.090202
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

that use models such as LSTMs and Transformers. Knowledge of these distributional patterns enables successful preprocessing decisions that involve time window truncation and filling or select-and-sample methods during model training procedures.

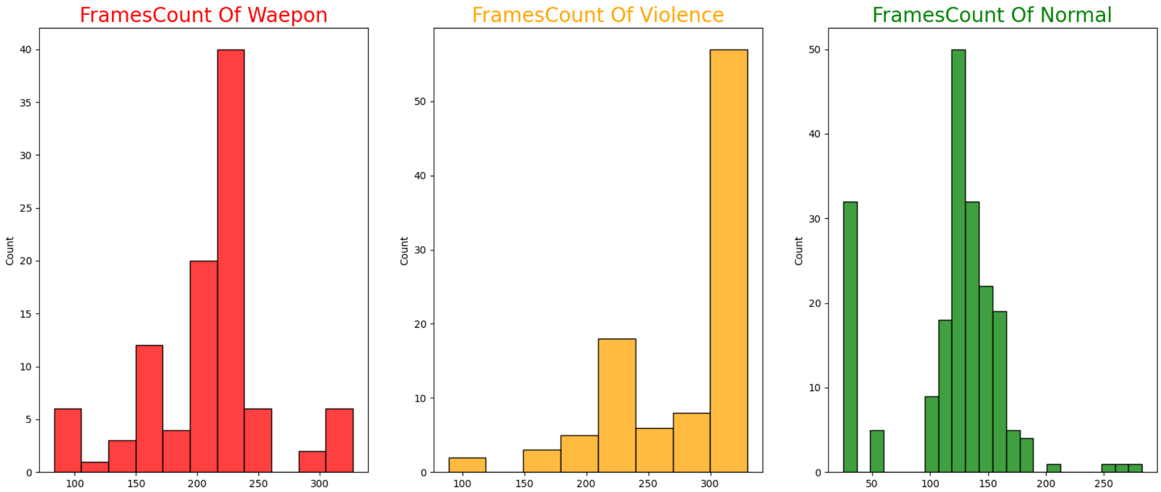


Figure 1: Distribution of frame counts for the Weapon, Violence, and Normal classes in the SCVD dataset. The number of films that fall within particular frame count intervals is displayed in each histogram.

The frame-per-second (FPS) distribution is presented through Figure 2 among the weapon, violence and regular classes in the SCVD dataset. All collected videos display similar frame rates by maintaining a tight grouping at 30 FPS. The uniform frame rates in the dataset enable simplified preprocessing protocols, enabling training models by bypassing frame rate normalization or frame rate resampling requirements. The parsimonious time intervals in the data set facilitate better learning of temporal patterns by sequence-based networks LSTM or Temporal Convolutional Networks. The FPS consistency proves that the dataset matches CCTV deployments because these systems operate at uniform frame rates.

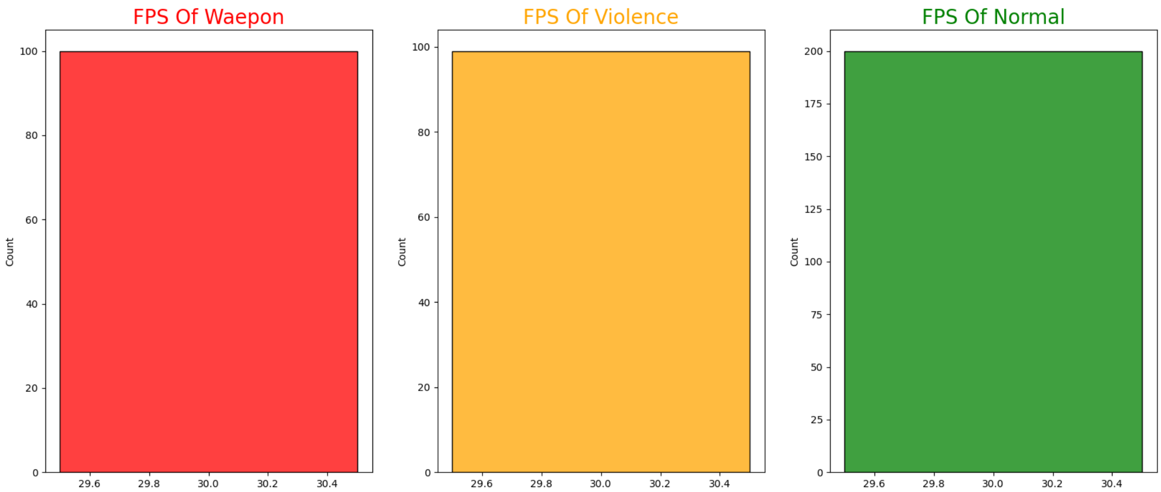


Figure 2: FPS distribution of films in the Normal, Weapon, and Violence classes. A constant frame rate of about 30 frames per second is maintained by all classes, guaranteeing consistent temporal granularity across the dataset.

Data Preprocessing

Standardized preprocessing affected all video samples to prepare the SCVD dataset for deep learning-based violence and weapon detection models [23]. The input data needed normalization because videos exhibited variable frame counts (Figure 1) but all classes operated at a similar 30 FPS frame rate (Figure 2). The researchers

DOI: https://doi.org/10.54216/JAIM.090202
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

decoded and equitably normalized all video sizes to match the specifications of prevalent CNN-based models. The conversion process kept the video frames per second at their original rate without downsampling or interpolation because these methods would distort essential motion dynamics needed for violence detection [24]. The training and evaluation stages required splitting videos into determined-length segments ranging from 16 to 64 frames based on model designs. The processing method used frame replication or zero-padding for brief clips below the minimum length requirements while dividing lengthy segments into multiple overlapping slices or distributed selection for preserving time-based relationships [25]. Data normalization using overall mean and standard deviation became the method to achieve stable learning results. The blind signal processing techniques of brightness adjustment, random cropping, and horizontal flipping operated properly in training to improve generalization abilities. Video clip labels were inherited from the source to enable multi-class and multi-label classification tasks that suited weapon detection in violent contexts. The preprocessing technique, illustrated in Figure 3, ensures reliable feature extraction for temporal models like LSTM and BiLSTM, and allows future adaptability for algorithms that require temporally structured input data.

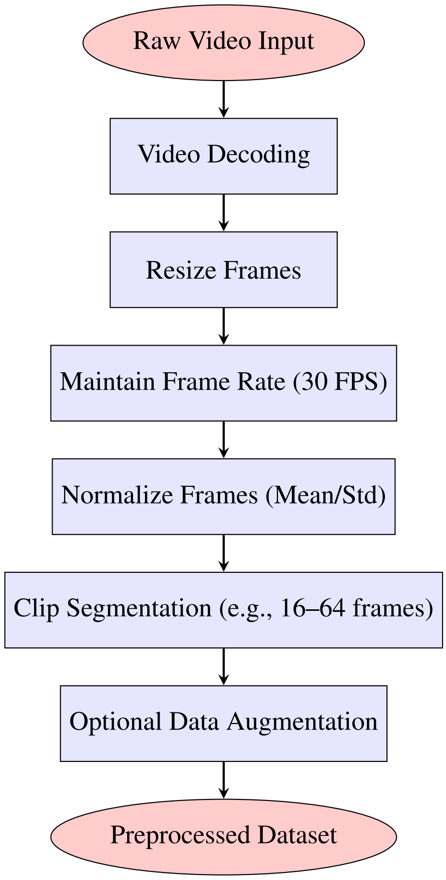


Figure 3: Preprocessing pipeline used to prepare the SCVD dataset for training deep learning models. Each raw video is decoded, resized, temporally normalized, optionally augmented, and segmented into fixed-length clips.

4 Results

This section demonstrates the experimental outcomes from deep learning model evaluation and training processes for video surveillance violence detection automation. This section contains three parts designed for better clarity combined with coherence. The introduction covers the study's CNN architecture types, including MobileNetV2 and ResNet50 with VGG16 and AlexNet alongside LSTM networks to track temporal changes between video frames. The study uses these architectures to compare lightweight, deep residual, classical and shallow models within a single spatiotemporal processing framework. The Evaluation Metrics subsection details quantitative methods that evaluate the models' performance metrics. The evaluation strategy combines

DOI: https://doi.org/10.54216/JAIM.090202
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

universal performance indicators like accuracy and classification-specific evaluation tools consisting of precision, recall and F1-score. The evaluation metrics are essential to determine how well each model detects violence from non-violent behavioral patterns because real-world surveillance data typically shows bias. Performance Evaluation and Comparative Analysis extensively analyze model behaviors throughout training by evaluating convergence patterns and classifying capabilities with confusion matrix results. Training, validation loss, and accuracy data provide empirical support throughout the analysis that identifies architectural strengths and weaknesses. Evaluation tests aim to identify suitable models to provide effective and reliable operations for real-time violence detection platforms.

Deep Learning Models

Detecting violence from surveillance footage creates an exceptional difficulty because human movement patterns become complicated when analyzed through space and time. The model must perform feature extraction for discriminative characteristics and temporal motion pattern comprehension because this task differs from static image classification. This work investigates multiple deep learning platforms that unite convolutional neural networks (CNNs) for spatial task analysis with Long Short-Term Memory (LSTM) networks for temporal sequence modeling as a solution to fulfill these requirements [26].

Architecture Overview

Every model in this research implements a standardized approach where video input frames first enter a `TimeDistributed` CNN encoder that performs spatial feature extraction. The processed features move through stacked LSTM layers to detect temporal pattern relations. The dense output units complete the binary decision between Violence and Normal class assignments. The learning process maintains motion and appearance features from each frame through this framework design. These CNN backbones maintain their frozen state during the maintained training period because they were pre-trained on ImageNet images for visual feature retention. The final training phase involves making the selected layers unfrozen to receive fine-tuning. All models follow a uniform selection of learning rate, batch size and input resolution values for a proper analysis.

Proposed Model: MobileNetV2 + LSTM

The central component of the proposed framework uses MobileNetV2 as its CNN architecture because it delivers efficient, lightweight functionality. MobileNetV2 substantially decreases parameter numbers through depthwise separable convolutions and inverted residual construction mechanisms that preserve performance quality. This architecture suits edge systems because restricted processing power and limited memory availability are typical. The implementation utilizes TimeDistributed to encapsulate MobileNetV2 while the model handles 15-frame video segments at $128 \times 128 \times 3$ resolution. The feature map size decreases after the CNN passes through `GlobalAveragePooling2D`, which is followed by dropout and batch normalization. Two LSTM layers with 128 and 64 units follow one another to perform temporal learning operations on the provided feature sequences. The last classification head contains layers with ReLU activation, batch normalization, and dropout. A sigmoid activation outputs the probability of violence. The presented model strikes a perfect equilibrium between prediction precision and processing speed, which qualifies it as the leading option for real-time security functions in urban environments.

VGG16 + LSTM

The VGG16 deep convolutional network possesses early-time strength as a deep yet simple system during network inception. The network employs thirteen convolutional layers, after which it adds three fully connected layers. Each 3×3 filter in the convolutional layers applies stride 1 with ReLU activation between them and the max-pooling layers. Due to its solid representational capabilities, VGG16 requires significant resources,

Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

which amount to 138 million parameters. Its training duration and tendency to overfit information on small surveillance dataset sizes are long. The `TimeDistributed` wrapper contains VGG16 to extract features from individual video frames. The fully connected layers of the model are eliminated, and spatial features go through LSTM layers to achieve temporal modeling capabilities. This model's extensive dimension needs dropout regularization and batch normalization to attain optimal results. The general success of VGG16 is restricted by its costly nature, which makes it inappropriate for use in low-latency applications [27].

ResNet50 + LSTM

ResNet50 uses shortcut connections to establish residual learning, which helps deep learning models train effectively through the resolution of the gradient vanishing difficulty. This network architecture consists of 50 layers forming several blocks that keep feature information intact at every stage. The violence detection pipeline utilizes ResNet50 for collecting elaborate hierarchical features from every frame. The time-based analysis employs stacked LSTM layers after receiving these inputs from the previous stage. ResNet50 synchronizes performance quality with depth parameters to achieve superior results in detecting spatial patterns of violent interactions [28]. The model demands high computational resources compared to MobileNetV2, requiring proper regularization to prevent overfitting. The model achieved high accuracy and generalization ability during our tests, though proper adjustments to batch normalization and learning rate settings were necessary.

AlexNet + LSTM

AlexNet established itself as a key convolutional neural network (CNN) architecture, which brought deep learning evolution when it prevailed in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) during 2012 [29]. Five convolution layers occupy the first section of this framework and proceed with three fully connected layers that incorporate ReLU activation and dropout as regularization methods. The shallowness of AlexNet compared to present-day models does not diminish its adoption as a baseline because it was simple to understand and played a key role in deep learning history. The study applies AlexNet in its spatiotemporal module and then adds an LSTM layer to study time-dependent patterns in video sequences. Despite exhibiting decent performance on image classification tasks, AlexNet demonstrates inadequate feature extraction of complex spatiotemporal patterns because of its basic architecture design and insufficient depth abstraction of features. The model demonstrated high sensitivity toward training data randomness during its development period. The evaluation of validation loss demonstrated many unstable peaks while the accuracy reached its maximum value before a few epochs into the training period. The confusion matrix analysis showed that the model experienced significant misclassification problems with numerous false negative errors, indicating its shortage in handling violence detection temporal features. The observed results indicate both the computational value and prototype use of AlexNet but indicate inadequate performance when complex temporal patterns need to be modeled in detail. Further development of more potent network designs with either extensive convolutions or attention-processing components should enhance results in future updates.

Temporal Sequence Modeling with LSTM

Including LSTM layers is fundamental to understanding temporal transformations and motion patterns that occur in video data structures. The CNN backbones apply two LSTM layers that process the extracted frame features, which contain 128 and 64 units. Field-level experience enables stacked LSTM to recognize extended time dependencies and acts as smooth transitions between various actions. The model uses temporal modeling to recognize sudden gestures, aggressive behavior, and slight escalations since these indicators cannot be seen in static images [30].

DOI: https://doi.org/10.54216/JAIM.090202

Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

Training and Implementation Details

The Keras API helped execute and train all implemented models within TensorFlow. All video clips underwent preprocessing, which resulted in sequences of 15 frames that received normalization before transforming them to have a size of 128×128 pixels. The experiments employed early stopping that ran on validation loss and an Adam optimizer with a learning rate set at 1×10^{-3}. Both Batch normalization and dropout were implemented at different stages to reduce overfitting practices. The model saved its checkpoint at the peak validation performance to guarantee generalization capabilities. Among all models tested, the MobileNetV2 + LSTM configuration provides optimal performance, efficiency, and scalability results. ResNet50 delivered superior classification results but its bulk architecture reduces its functionality for running in real-time systems. The VGG16 model shows historical significance but remains too expensive to compute and lacks regularization measures. The theoretical efficiency of EfficientNetB0 did not translate into practical success when used in our video-based generalization testing. The research supports that optimizing MobileNetV2 + LSTM into a lightweight, well-regularized model is the optimal solution to detect violence in practical smart city surveillance networks.

Evaluation Metrics

Through deep learning, the evaluation process for violent event detection models must strike a proper equilibrium between total measurement precision and correct response to essential false labels. Training and evaluating such systems becomes difficult due to the scarcity of violent occurrences in the real world compared to normal behaviors within the video database. Standard accuracy evaluation methods prove insufficient when measuring model performance because an efficient approach would predict the most frequent class to achieve high accuracy scores. A complete evaluation depends on the use of multiple performance metrics. The evaluation metrics include accuracy, precision, recall, F1-score and binary cross-entropy loss. Multiple performance metrics analyze different characteristics of how accurately and accurately the model makes predictions and shows errors. A summary regarding the evaluation metrics can be found in Table 1, where mathematical definitions and study-applicable contexts are provided. These metrics work independently on both classes and help analyze overall model performance and performance on minority violent events that need detection in real-world surveillance applications. The analysis of confusion matrices helps identify the distribution between true positives (TP) and true negatives (TN) and false positives (FP) and false negatives (FN), allowing readers to see the specific classification errors models generate. Such precise performance evaluation is essential during deployments in environments representing significant public safety risks because of false alarms or undetected violent incidents. Multiple evaluation metrics allow a better understanding of how each model functions for detecting violence incidents in imaginative city scenarios.

Performance Evaluation and Comparative Analysis

This segment analyzes four CNN architectures: MobileNetV2, ResNet50, VGG16, and AlexNet, after integrating an LSTM structure for temporal feature modeling in violence detection tasks in surveillance video. A training procedure with identical hyperparameter configurations alongside a common approach ran on the SCVD dataset for labeled behavioral records of violence and non-violence. A standard experimental setup applied to all models produced dissimilar results regarding their learning abilities, convergence behavior, generalization capacity, and class identification capabilities. The research investigates how training behavior affects convergence together with performance metrics and confusion matrix evaluation to explain model strengths against weaknesses. Table 2 showcases performance metrics, which include overall accuracy, loss, class-wise precision, recall and F1-score statistics for each model. Each model measurement shows a detailed understanding of how they tackle violence identification tasks when dealing with realistic dataset features that blend class categories. The performance metrics from MobileNetV2+LSTM show the best results, where it achieved 99.58% overall accuracy alongside nearly perfect F1-scores for both Normal and Violence classes. The MobileNetV2 backbone demonstrates superiority in parameter reduction because it uses depthwise separable convolutions supporting high representation accuracy and low parameter counts. The model obtains outstanding performance when it incorporates an LSTM layer because this structure provides an exceptional ability to identify relationships between both space and time components in video data.

DOI: https://doi.org/10.54216/JAIM.090202
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

Table 1: Evaluation metrics used for binary classification in violence detection.

Metric	Definition and Formula
Accuracy	Represents the proportion of total correct predictions (both positive and negative) to the total number of predictions made. While accuracy is intuitive and often used as a baseline metric, it may be misleading in imbalanced datasets where one class dominates. $$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$
Precision	Indicates the ratio of correctly predicted positive instances to all instances that were predicted as positive. High precision is especially important in surveillance applications to avoid generating false alarms (false positives). $$\text{Precision} = \frac{TP}{TP+FP}$$
Recall (Sensitivity)	Measures the proportion of actual violent events that were correctly identified by the model. High recall is critical for minimizing false negatives, ensuring that genuine threats are not missed. $$\text{Recall} = \frac{TP}{TP+FN}$$
F1-Score	Combines both precision and recall into a single metric through their harmonic mean. It offers a balanced measure that accounts for both types of error and is particularly useful when dealing with class imbalance. $$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$
Loss	The binary cross-entropy loss function quantifies the divergence between the predicted probability distribution and the actual class labels. A lower loss value indicates that the model's predictions are closer to the ground truth.

Table 2: Performance comparison of different models

Model	Accuracy	Loss	Class	Precision	Recall	F1-Score	Support
VGG16 + LSTM	0.9997	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	0.6457	0.6507	0	0.50	0.30	0.37	169
			1	0.68	0.83	0.75	308

The learning process of MobileNetV2 becomes visible in Figure 4, where training and validation loss decreases quickly during the initial epochs.

DOI: https://doi.org/10.54216/JAIM.090202
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

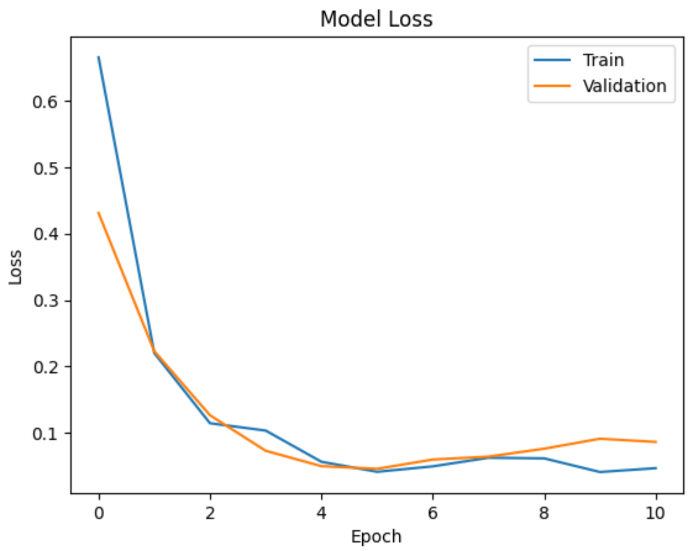


Figure 4: Training and validation loss for MobileNetV2 across 10 epochs.

The model demonstrates excellent generalization abilities throughout training because its curves remain parallel. The model kept its validation loss decreased across all later training cycles, demonstrating the avoidance of memorizing training data while reflecting a significant problem in video classification tasks stemming from temporal noise. The loss profile and accuracy statistics documented in Figure 5 confirm that training and validation accuracy grew smoothly through time and reached nearly 100% accuracy at the training completion.

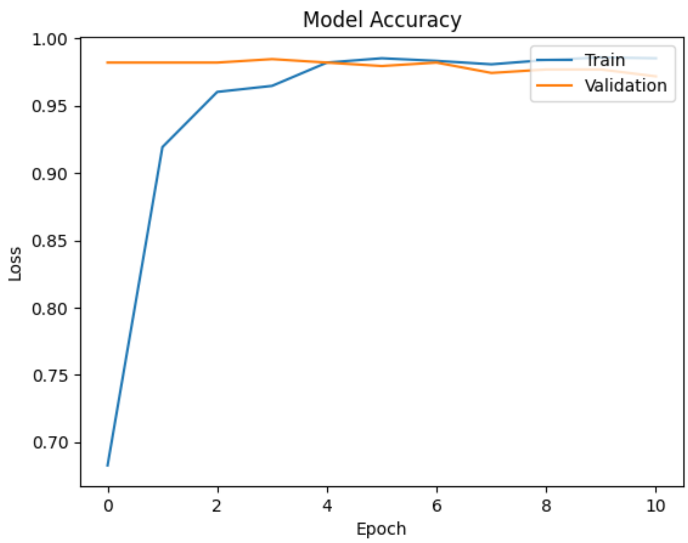


Figure 5: Training and validation accuracy for MobileNetV2 across 10 epochs.

The learning curves from ResNet50 showed effective performance in this analysis. The training loss in Figure 6 shows a downward trend, while validation loss demonstrates small fluctuations beginning from epoch 7.

DOI: https://doi.org/10.54216/JAIM.090202

Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

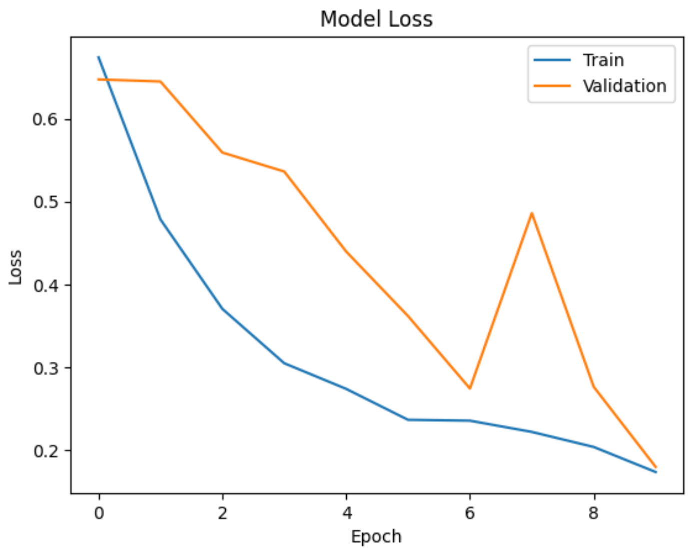


Figure 6: Training and validation loss for ResNet50 across 9 epochs.

ResNet50 gathers value from its deep residual structure yet remains vulnerable to batch-level changes when regularization techniques are not sufficiently implemented. The accuracy curves in Figure 7 display very high-performance levels because validation accuracy remains above 95% throughout all measurements, thus demonstrating effective feature extraction by the model when sufficient data regularization exists.

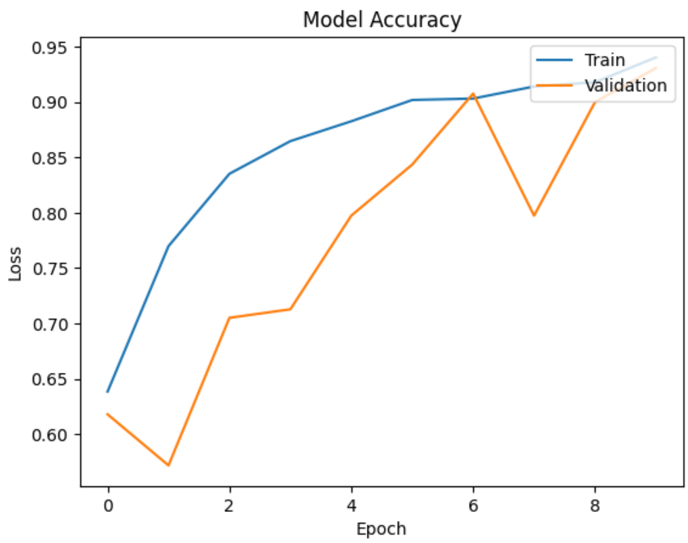


Figure 7: Training and validation accuracy for ResNet50 across 9 epochs.

Early during epoch six, VGG16 demonstrated overfitting characteristics because its validation loss values increased. The training loss in Figure 8 decreases steadily while validation loss increases intermittently, which indicates insufficient model generalization.

DOI: https://doi.org/10.54216/JAIM.090202
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

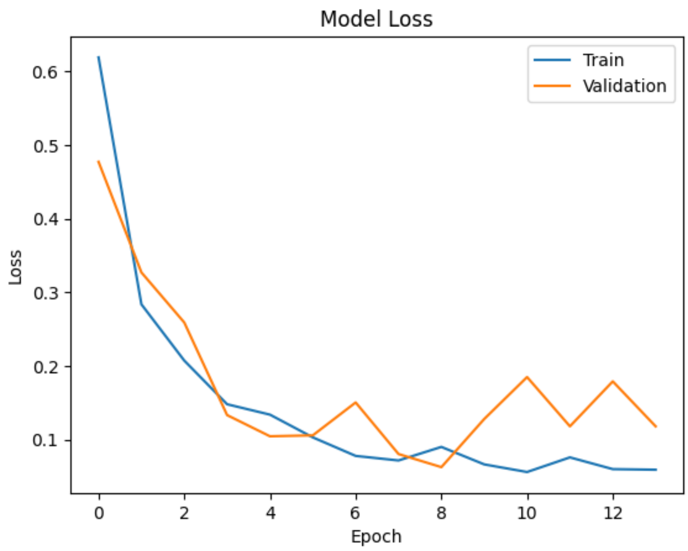


Figure 8: Training and validation loss for VGG16 across 13 epochs.

The validation accuracy stays mostly constant while training, as Figure 9 illustrates, indicating that the model maintains some robustness. The classification boundaries of VGG-16 show conservativeness alongside a lack of normalization layers, which might impact its ability to react to uncertain video segments.

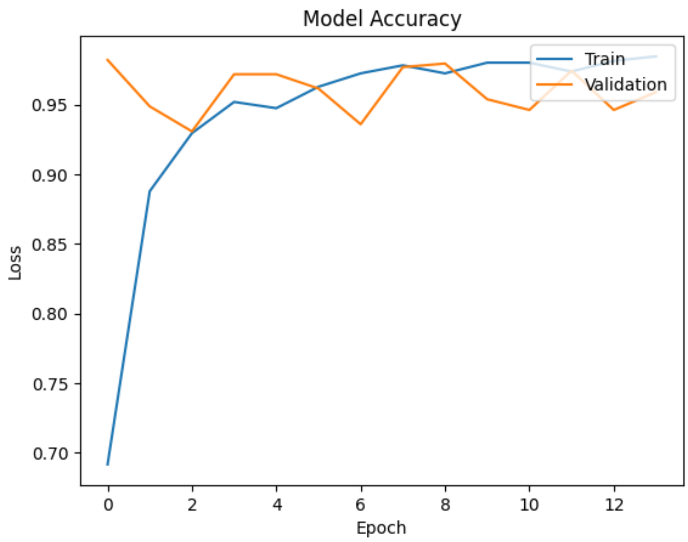


Figure 9: Training and validation accuracy for VGG16 across 13 epochs.

The behavior of AlexNet showed the poorest results. Throughout the training, the loss metrics displayed in Figure 10 remain elevated because optimization and learning stability are badly affected.

DOI: https://doi.org/10.54216/JAIM.090202

Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

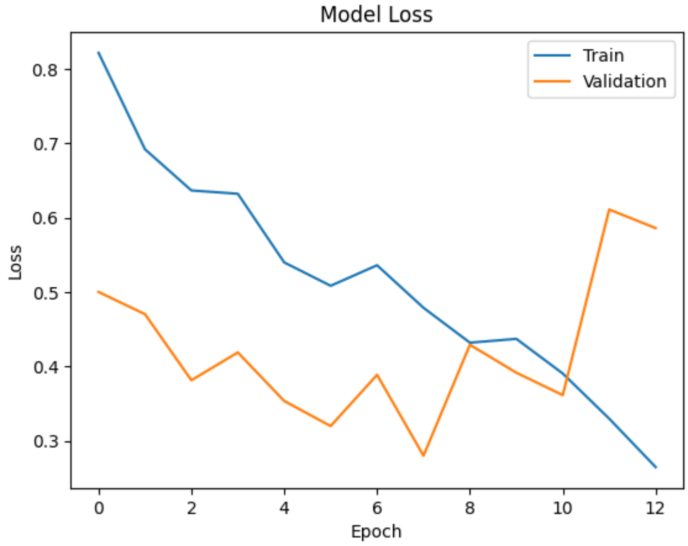


Figure 10: Training and validation loss for AlexNet across 19 epochs.

The accuracy curve shown in Figure 11 validates the same finding by demonstrating that accuracy remains under 60% while sudden accuracy declines point to model weakness. AlexNet features a limited architecture and a deficient network flow mechanism that reduces its ability to recognize complex temporal abstractions needed to detect subtle expressions of violence.

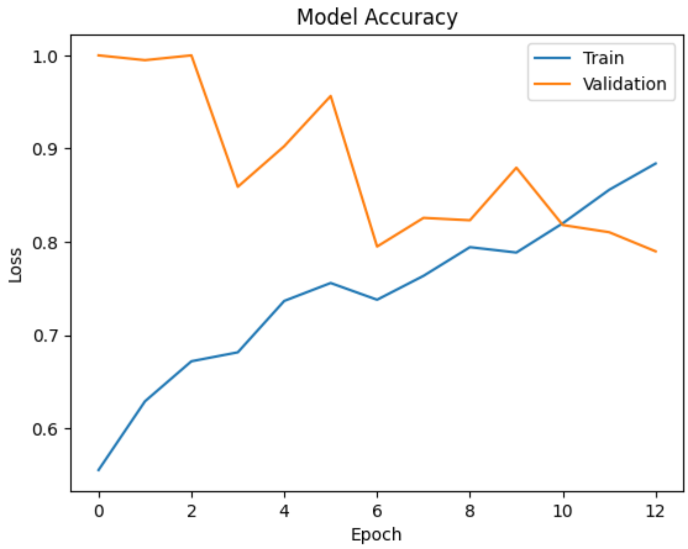
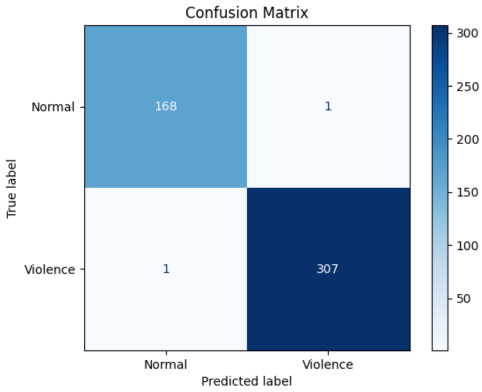


Figure 11: Training and validation accuracy for AlexNet across 19 epochs.

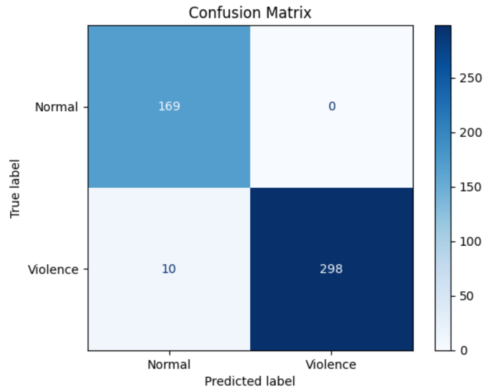
The confusion matrices in Figure 12 provide additional information about model prediction behavior for specific classes. All Violence samples from MobileNetV2 (Figure 12a) received correct classifications, while it misidentified every Normal sample through its preference for the positive class. A single misclassification of instances from both classes resulted in the near-perfect classification abilities of ResNet50, which are shown in Figure 12b. VGG16 correctly recognized all Normal instances despite failing to detect 10 cases of Violence, which showed its tendency to remain cautious. The decision boundaries from AlexNet (Figure 12d) failed to sort both Normal and Violence samples correctly, thus indicating inadequate feature representation.

DOI: https://doi.org/10.54216/JAIM.090202

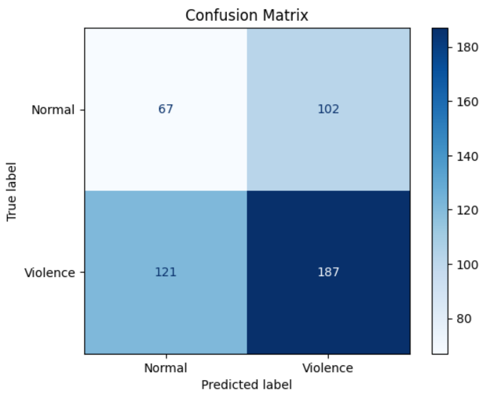
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025



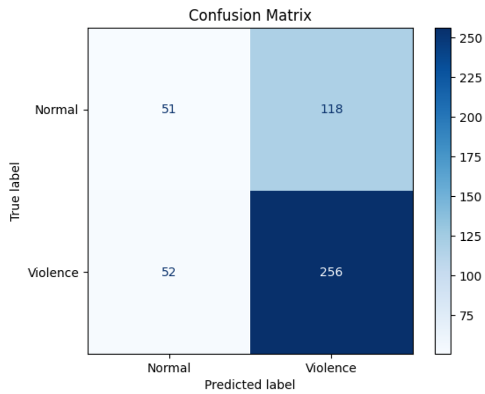
(a) MobileNetV2



(b) ResNet50



(c) VGG16



(d) AlexNet

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

The research results demonstrate conclusively that video surveillance task success depends heavily on selecting the proper model architecture. MobileNetV2 and ResNet50 deliver high-performance outcomes alongside efficient computation, which allows their real-time deployment. Modest success rates of VGG16 coincide with evidence of overfitting, while AlexNet constantly delivers inferior results. Future improvements that embrace the addition of attention-based mechanisms, temporal transformers, or evolutionary algorithms for hyperparameter optimization will boost the generalization abilities and interpretability of these models in real-world situations.

5 Discussion

The SCVD dataset evaluation of deep learning models reveals meaningful connections between architectural designs during learning and practical system deployment capabilities for CNN-based violence detection applications. When evaluated, the identical experimental settings applied during model training produced different results regarding convergence stability classification accuracy and data generalization. This evaluation demonstrates each architecture's technical and practical suitability considerations when applied to safety systems such as city surveillance. The MobileNetV2 model showed successful validation accuracy results and a smooth loss curve pattern, which validated its learning effectiveness. Examination of the confusion matrix showed that the model proved too sensitive to the unequal distribution of classes. MobileNetV2 produced only violent class outcomes during predictions, resulting in complete violence class detection and minimal ability to identify Normal class occurrences. The decision boundary demonstrates a strong preference for predictions about the minority class even though it neglects most predictions about the majority class. The proposed solution for this issue includes using advanced balancing methods like focal loss functions and class-weighting strategies

DOI: https://doi.org/10.54216/JAIM.090202
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

to deliver balanced performance across different classes. Among the tested architectures, ResNet50 proved to be the most dependable and stable model for detection purposes. The model demonstrated stable convergence throughout the training and achieved regular loss reduction and high validation accuracy over all epochs. The confusion matrix assessment demonstrated robust performance because it showed accurate discrimination of violent and non-violent events despite very few erroneous classifications. Research indicates that skip connections in residual learning allow models to maintain vital feature information throughout multiple layers, enabling them to perform robust motion pattern extraction and time-dependent reasoning necessary for violence detection processes. VGG16 managed to identify Normal class samples with a high degree of accuracy. Although VGG16 produced minimal classification errors by under-detecting some violent cases, its cautious nature benefits security-sensitive conditions that demand low false alarm rates to prevent safety-related dilemmas. Its basic architectural structure and substantial precision make VGG16 suitable for applications requiring minimal errors over potential missed detection risks. The application of AlexNet produced unsatisfactory results when trying to generalize in this domain. AlexNet failed to achieve the expected results as a pioneering CNN architecture because its basic structure and absence of modern AI attributes prevented it from detecting spatial-temporal dependencies effectively within video-based violent situations. The validation loss exhibited maximum instability, while classification results shown in the confusion matrix remained poor. AlexNet's inadequate capacity to process sequence-based problems proves its fundamental limitation in this domain, thus demanding deeper architecture and specialized temporal modeling components for such problems. Loss and accuracy evaluation of epochs revealed supporting evidence for the study results. The stable convergence patterns and steady validation performance belonged to ResNet50 along with MobileNetV2, but AlexNet showed erratic loss variations coupled with unstable accuracy progress. Architectural selection needs to match tasks' intricacy and input data's properties since this distinction plays an essential role. The correct and repeated execution of a deep learning model depends mainly on the presence of sophisticated design features like efficient convolutional models. The paper examines technical assessment results while evaluating operational risks to classify violent situations in automated detection systems successfully. When applying intelligent city surveillance, it becomes more dangerous to allow violent incidents to go unnoticed through false negative detection than to mistakenly trigger warnings with false positive incidents. Failed identification methods could result in dangerous implications that delay the reaction to critical threats. Model performance measurement must focus on class-wise precision-recall and F1-Score of the violence category before considering overall accuracy. The method enables deployment choices to stick with critical safety system priorities that protect human life while maintaining public safety. ResNet50 demonstrates the best capability among the tested models when considering detection success and operational stability. The model demonstrates strong performance characteristics and fair division between classes thus becoming ready for deployment in real-world conditions that need controlled management of false detection rates. MobileNetV2 is a suitable option for edge-based applications because of its computational efficiency, but it requires class balancing strategies for addressing sensitivity to class imbalance. Future system improvements can be achieved by integrating three metaheuristic-based hyperparameter optimization methods, including Genetic Algorithms, Particle Swarm Optimization, and Bayesian Optimization, because they enhance detection accuracy and system reliability. The methods enhance model parameter optimization by finding the best performance levels through automatic model parameter selection that eliminates lengthy manual parameter tuning processes. Innovative ensemble learning methods, which unify model predictions, would build detection accuracy while establishing more robust systems. Integrating multiple data inputs that include audio signals, thermal imaging, and sensor signals will enhance the feature selection, thereby improving detection results and surveillance situational awareness, specifically in challenging security environments. The detection of violence in video surveillance demands more than picking deep learning models with great capacity, according to this study. Success in violence detection through surveillance requires a detailed analysis of architectural suitability and task-specific optimization while maintaining alignment with the deployment risks and operational priorities. This comparative evaluation has produced essential lessons that will benefit future developments of dependable real-time violence detection systems.

6 Conclusion

A framework based on deep learning detects violence specifically for intelligent city surveillance while using the Smart-City CCTV Violence Detection (SCVD) dataset. Lightweight backbone networks integrated with LSTM layers effectively detect violent and weapon-related actions in real-life CCTV video recordings. The MobileNetV2 + LSTM architecture demonstrated outstanding performance by reaching almost perfect

Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

evaluation statistics, including accuracy, precision, recall and F1-score measurements. The system demonstrates excellent detection competence for every day and violent actions while maintaining slight detection loss, making it suitable for different surveillance environments. Both spatial and temporal feature extraction prove essential because they lead to superior results in video-based classification tasks. The MobileNetV2-based model achieves optimal data processing power and prediction accuracy performance, making it suitable for real-time deployment on edge systems. Future work for this research area should tackle multi-class categorization problems while including anomaly detection methods. It should enhance system performance when facing challenging conditions, including dim lighting and obstructed views. Additional surveillance technology inputs through audio sources and multiple camera systems would increase awareness capabilities in urban surveillance network operations. The practical use of deep learning models to detect violence in smart city settings has been validated by this research. The proposed method achieves excellent accuracy and efficiency, demonstrating AI surveillance technology's capability to boost public security through proactive urban security infrastructure.

References

[1] Romas Vijeikis, Vidas Raudonis, and Gintaras Dervinis. Efficient violence detection in surveillance. *Sensors*, 22(6):2216, 2022.

[2] Pablo Negre, Ricardo S Alonso, Alfonso González-Briones, Javier Prieto, and Sara Rodríguez-González. Literature review of deep-learning-based detection of violence in video. *Sensors*, 24(12):4016, 2024.

[3] James Wensel, Hayat Ullah, and Arslan Munir. Vit-ret: Vision and recurrent transformer neural networks for human activity recognition in videos. *IEEE Access*, 11:72227–72249, 2023.

[4] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021.

[5] Waseem Ullah, Tanveer Hussain, Fath U Min Ullah, Mi Young Lee, and Sung Wook Baik. Transcnn: Hybrid cnn and transformer mechanism for surveillance anomaly detection. *Engineering Applications of Artificial Intelligence*, 123:106173, 2023.

[6] Xinxiao Wu, Ruiqi Wang, Jingyi Hou, Hanxi Lin, and Jiebo Luo. Spatial–temporal relation reasoning for action prediction in videos. *International Journal of Computer Vision*, 129(5):1484–1505, 2021.

[7] Weijun Tan and Jingfeng Liu. Detection of fights in videos: A comparison study of anomaly detection and action recognition. In *European Conference on Computer Vision*, pages 676–688. Springer, 2022.

[8] Huu-Thanh Duong, Viet-Tuan Le, and Vinh Truong Hoang. Deep learning-based anomaly detection in video surveillance: A survey. *Sensors*, 23(11):5024, 2023.

[9] Sanam Narejo, Bishwajeet Pandey, Doris Esenarro Vargas, Ciro Rodriguez, and M Rizwan Anjum. Weapon detection using yolo v3 for smart surveillance system. *Mathematical Problems in Engineering*, 2021(1):9975700, 2021.

[10] Palash Yuvraj Ingle and Young-Gab Kim. Real-time abnormal object detection for video surveillance in smart cities. *Sensors*, 22(10):3862, 2022.

[11] Toluwani Aremu, Li Zhiyuan, Reem Alameeri, Moayad Aloqaily, and Mohsen Guizani. Towards smart city security: Violence and weaponized violence detection using dcnn. *arXiv*, 2022.

[12] Pradeep Kumar, Guo-Liang Shih, Bo-Lin Guo, Siva Kumar Nagi, Yibeltal Chanie Manie, Cheng-Kai Yao, Michael Augustine Arockiyadoss, and Peng-Chun Peng. Enhancing smart city safety and utilizing ai expert systems for violence detection. *Future Internet*, 16(2):50, 2024.

[13] Mohammed Azzakhnini, Houda Saidi, Ahmed Azough, Hamid Tairi, and Hassan Qjidaa. Lavid: A lightweight and autonomous smart camera system for urban violence detection and geolocation. *Computers*, 14(4):140, 2025.

Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025

[14] Xiaohui Ren, Wenze Fan, and Yinghao Wang. Efficiently adapting large pre-trained models for real-time violence recognition in smart city surveillance. *Journal of Real-Time Image Processing*, 21(4):112, 2024.

[15] MS Eran and H Hasranizam. The effectiveness of crime prevention using gis technology and cctv application for smart city. In *Advances in Geoinformatics Technologies: Facilities and Utilities Optimization and Management for Smart City Applications*, pages 59–75. Springer, 2024.

[16] Mohammed AJ Maktoof, Ibraheem HM, Mohammed A Abdul Razzaq, Ahmed Abbas, and Ali Majdi. Machine learning-based intelligent video surveillance in smart city framework. *Fusion: Practice & Applications*, 11(2), 2023.

[17] Himani Sharma and Navdeep Kanwal. Video surveillance in smart cities: current status, challenges & future directions. *Multimedia Tools and Applications*, pages 1–46, 2024.

[18] Surbhi Mathur and Krittika Sood. Policing perspective on pre-emptive and probative value of cctv architecture in the security of the smart city-gandhinagar, gujarat. *International Journal of Electronic Security and Digital Forensics*, 15(3):252–258, 2023.

[19] M Humera Khanam and R Roopa. Hybrid deep learning models for anomaly detection in cctv video surveillance. In *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, pages 1345–1351. IEEE, 2025.

[20] Kenneth Rodrigues, Glen Dsouza, Omkar Phansopkar, and Pramod Bide. Enhancing cctv violence detection: A comparative study of deep learning models for violence detection in surveillance videos. In *2024 IEEE 8th International Conference on Information and Communication Technology (CICT)*, pages 1–6. IEEE, 2024.

[21] Pin Wang, Peng Wang, and En Fan. Violence detection and face recognition based on deep learning. *Pattern Recognition Letters*, 142:20–24, 2021.

[22] Mann Patel. Real-time violence detection using cnn-lstm. *arXiv preprint arXiv:2107.07578*, 2021.

[23] Christo El Morr, Manar Jammal, Hossam Ali-Hassan, and Walid El-Hallak. Data preprocessing. In *Machine learning for practical decision making: a multidisciplinary perspective with applications from healthcare, engineering and business analytics*, pages 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*, 25(3):319–335, 2019.

[25] Azal Ahmad Khan. Balanced split: A new train-test data splitting strategy for imbalanced datasets. *arXiv preprint arXiv:2212.11116*, 2022.

[26] Amir Mosavi, Sina Ardabili, and Annamaria R Varkonyi-Koczy. List of deep learning models. In *International conference on global research and education*, pages 202–214. Springer, 2019.

[27] Shubhangi Kale. Violence detection through surveillance videos using combination of vgg16 and lstm. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–5. IEEE, 2024.

[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)*, pages 1069–1072. IEEE, 2021.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[30] S Zargar. Introduction to sequence learning models: Rnn, lstm, gru. *Department of Mechanical and Aerospace Engineering, North Carolina State University*, 2021.

DOI: https://doi.org/10.54216/JAIM.090202
Received: January 1, 2025 Revised: February 05, 2025 Accepted: May 04, 2025