Computational Sciences and Engineering 4(2) (2024) 203-216

University of Guilan

Computational Sciences and Engineering

journal homepage: https://cse.guilan.ac.ir/

Advanced Deep Learning Approaches for Accurate and Efficient Suspicious Behavior Detection in Surveillance Videos

Arash Safdel [a], Jamal Ghasemi [a,*], Seyyed Ali Zendehbad [a]

[a] Faculty of Engineering & Technology, University of Mazandaran, Babolsar, Iran

ARTICLE INFO

Article history:
Received 27 March 2025
Received in revised form 28 May 2025
Accepted 23 June 2025
Available online 23 June 2025

Keywords:
Anomaly Detection
Deep Learning
Pattern Recognition
Feature Fusion
Video Analysis

ABSTRACT

Violence Artificial Intelligence (AI) and Deep Learning (DL) systems present a difficult research area for identifying violence in videos within urban security frameworks and video surveillance systems. The proposed model divides violence detection tasks in video into two stages to achieve both rapid processing and precise outcomes. The LeNet-5 model operates at a speed of 0.8 frames per second to filter out non-violent videos during the first stage of operation. The second analysis stage employs the ResNet-50 model to inspect videos for potential violence when their probability surpasses 0.4. The Real-Life Violence dataset consisting of 1951 videos with 1000 violent and 951 non-violent videos was used for testing this system. The implementation produced 97.03% accuracy together with 95.70% recall and 98.46% precision and 97.06% F1-Score and AUC of 0.9902. Each frame requires only 20 milliseconds of processing time which allows real-time application of this system. A comparative analysis with existing methods, such as 3D-CNN, ViT, and YOLOv5+TSN, highlights the superiority of the proposed model in terms of both accuracy and speed. The system achieves better violence detection capabilities and operational reliability in real-world applications because it decreases detection errors.

1. Introduction

Urban life depends heavily on Closed-Circuit Television (CCTV) cameras as their essential elements in modern society [1]. These surveillance cameras function throughout public areas as well as shopping centers and transportation facilities and active streets to boost security while fighting criminal activities [2]. These security cameras alone do not resolve the security concerns [3]. Security cameras in crowded shopping centers need to respond swiftly when violent individuals

* Corresponding author.
E-mail addresses: j.ghasemi@umz.ac.ir (J. Ghasemi)

https://doi.org/10.22124/cse.2025.30210.1099

start their actions which causes panic and destruction. The situation requires quick response because time plays an essential role [4]. Security personnel would receive immediate notification through an alarm triggered by a real-time automatic violence detection system which would enable them to stop additional damage from occurring. Traditional video surveillance systems function as static cameras that need human operators for monitoring purposes since they lack artificial intelligence capabilities [5]. The system produces both high costs and human errors because of operator fatigue [6]. The year 2020 saw a violent shopping center incident which led to serious injuries according to reports. The implementation of an alert system would have prevented the incident from causing significant damage according to research [7]. This study addresses the challenge through its research. The proposed system implements a dual stage approach which utilizes LeNet-5 combine with ResNet-50 to identify violent content in videos efficiently and accurately [8, 9]. The system decreases computational requirements while showing flexibility to different environmental conditions. The developed model marks a substantial advancement toward better urban security by stopping violent events. The identification of violence in video content stands as a difficult issue within Deep Learning (DL) and image processing research domains [10]. Researchers have conducted multiple studies about this subject utilizing different techniques during the recent years [11]. Hand-crafted feature-based approaches including Histogram Of oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT), were employed in the past for video motion and violence detection [12]. Scientists began using these methods because they were easy to implement and demonstrated reasonable abilities to detect motion variations [13, 14]. The detection model showed inferior results when confronted with environmental changes including lighting variations along with camera angle variations and diverse types of motion [15]. Convolutional Neural Networks (CNNs) transformed violence detection research through their arrival [16]. LeCun et al. developed convolutional networks to perform image feature extraction automatically which led to enhanced detection accuracy levels [17]. He et al., established the ResNet-50 architecture to overcome deep network gradient vanishing thus making it one of the primary image recognition models [18]. Song et al. applied 3D-CNN for video analysis in violence detection which resulted in an 94.3% accuracy level. The model demonstrated excessive computational complexity which resulted in 150 milliseconds per frame processing time making it unfit for real-time usage [19]. The application of hybrid approaches that unite multiple models or techniques has increased in popularity during the recent years [3, 20]. The combination of CNN, LSTM, and attention mechanism by Vosta et al., achieved better results for violence detection accuracy on the UCF Crime dataset [21]. The proposed model achieved 96.21% accuracy yet its processing requirements including 4000 MB memory use and 200 milliseconds per frame made it impractical for real-time usage. A vision transformers (ViTs) represent a new approach which competes against CNNs as an alternative solution [22]. Rendón-Segador et al. applied the ViT model to detect violence in videos which produced high accuracy [23]. The research by Gautam et al. developed a model which integrated YOLO v5 and CNN to detect violence effectively with 98.63% accuracy and 66 milliseconds per frame processing time [24]. However, the advancement of violence detection systems remains hindered by an essential trade-off between detection precision and real-time capability which leads to major security concerns in urban areas [25]. The proposed research implements a two-part architecture that joins LeNet-5 speed to ResNet-50 accuracy to deliver real-time processing (24 FPS) with retained lighting/angle resistance. The system incorporates three essential improvements to handle ongoing issues with previous research which include (1) hierarchical filtering for reduced

computational needs and (2) adaptive threshold values at 0.4 or 0.5 to reduce incorrect alarms along with (3) memory consumption optimization to support edge deployment under 1000 MB.

2. Method & Materials

This study uses a video violence detection system based on the combination of LeNet-5 and ResNet-50 models using a benchmark dataset, which increases accuracy while reducing processing time. The system implementation method reduces computational requirements while maintaining tunable performance in different situations and has acceptable generalizability.

2.1. Dataset

This study relies on the Real-Life Violence Situations Dataset that stands as the leading extensive dataset for identifying violence in videos [26]. The Kaggle* platform makes this dataset accessible to the public through its platform and includes videos showing both violent and non-violent real-life situations. This dataset has been widely adopted in recent violence detection studies due to its realistic representation of real-world scenarios. The dataset features 1951 videos arranged in two distinct groups:

1. Violent Videos: This section contains 1000 videos showing violent behaviors including fights and physical assaults together with various violent activities.
2. Non-Violent Videos: The second subcategory contains 951 videos that show regular nonviolent activities including walking and talking as well as routine daily activities.

Furthermore, each video measurement ranges from 5 to 10 seconds which meets the needs of DL operations. The videos display three different environment types through their mix of open spaces (streets and parks) and closed spaces (malls and transit hubs) combined with natural and artificial lighting conditions. The MP4 file format stores these videos with sufficient quality to support easy processing by AI systems. The analysis becomes complicated due to fast and violent motions that appear in certain video elements. The detection accuracy of lighting conditions is a challenge while multiple camera angles require flexible model designs.

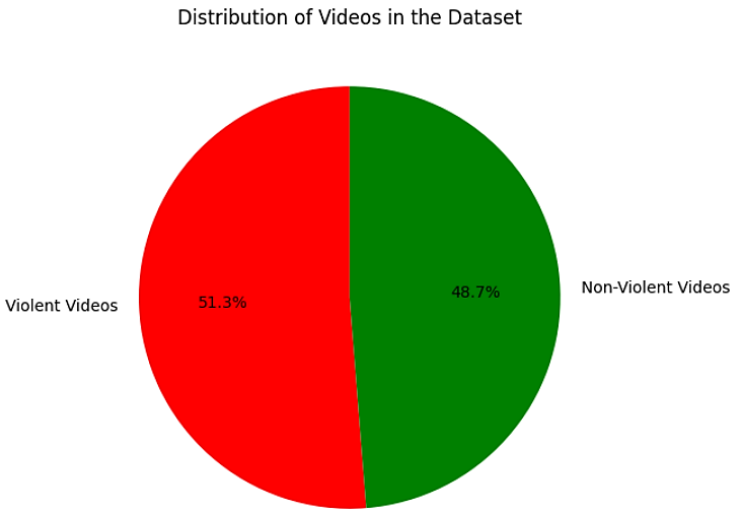

Figure 1. Overall Flowchart of the Proposed method.

* https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset

Figure 2. Example image of a violent example in the dataset [26].

2.2. Filtering Non-Violent Videos with LeNet-5

As shown in ***Figure 3***, LeNet-5 represents one of the first CNN designs which uses seven hierarchical layers to perform handwritten digit recognition. The network design contains tree convolutional layers with tanh activation and 5×5 kernels (blue blocks) then two average-pooling layers (2×2 subsampling) in sequence (purple blocks) before moving onto three fully connected layers (84-neuron hidden layer and the 2-output classification layer (pink block)). Coarse spatial information proceeds from sequential convolution and pooling operations before ending in an Euclidean Radial Basis Function (RBF) output layer used for digit classification. The network contains sixty thousand parameters of computation distributed across its architecture which establishes a standard for current DL theory and practice. In the first, the LeNet-5 model is employed to filter out non-violent videos. Due to its lightweight structure and high speed, this model can operate at a speed of 0.8 frames per second and effectively eliminate non-violent videos [27].

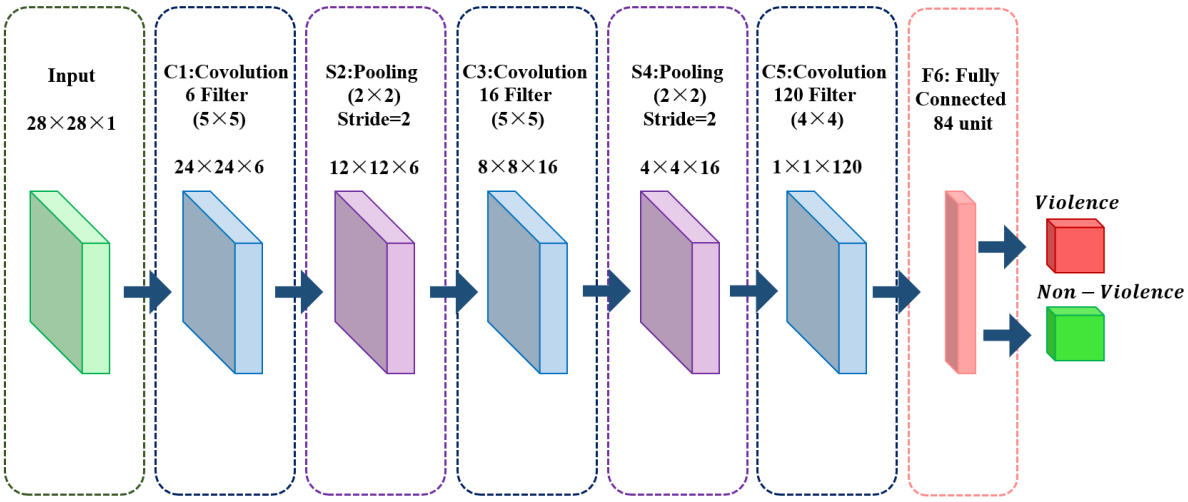

Figure 3. LeNet-5 Architecture for Detailed Violence Detection.

1. Input Data: 5 frames extracted from each video with a size of 28×28 pixels.
2. Normalization: Pixel values are normalized using the mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225].
3. Output: Violence Probability for each video.

If the violence probability is less than the threshold of 0.4, the video is identified as non-violent and further detailed analysis is omitted. The threshold value of 0.4 was selected based on empirical testing, and future work could leverage optimization techniques such as genetic algorithms to dynamically adjust this parameter for varying conditions.

2.3. Detailed Analysis of Suspicious Violent Videos with ResNet-50

The advanced deep CNN RezNet-50 serves for powerful video analysis especially in violence detection applications as shown in ***Figure 4***. RezNet-50 comprises 64 layers which incorporate blue-colored convolutional blocks with Rectified Linear Unit (ReLU) activation along with 3×3 kernels and purple-colored max-pooling layers and bottleneck residual connections (pink blocks) to manage gradient vanishing issues. The model's various components work together to extract hierarchical features which protect essential fine-grained spatial-temporal patterns needed for correct classification. Through its deep residual network RezNet-50 deals with video frames and reaches high accuracy for violence detection at a real-time frame rate of 24 FPS which exceeds the capabilities of lightweight models such as LeNet-5 [28]. Skip connections in the model (***Figure 4*** pink blocks) enable deeper training by improving gradient flow which prevents performance degradation during the process. The yellow blocks in the network represent its fully connected layers which use softmax activation for violent/non-violent binary classification. In the second, videos with a violence probability higher than the threshold of 0.4 are passed to the ResNet-50 model. Due to its greater depth and high feature extraction power, this model is capable of more detailed video analysis.

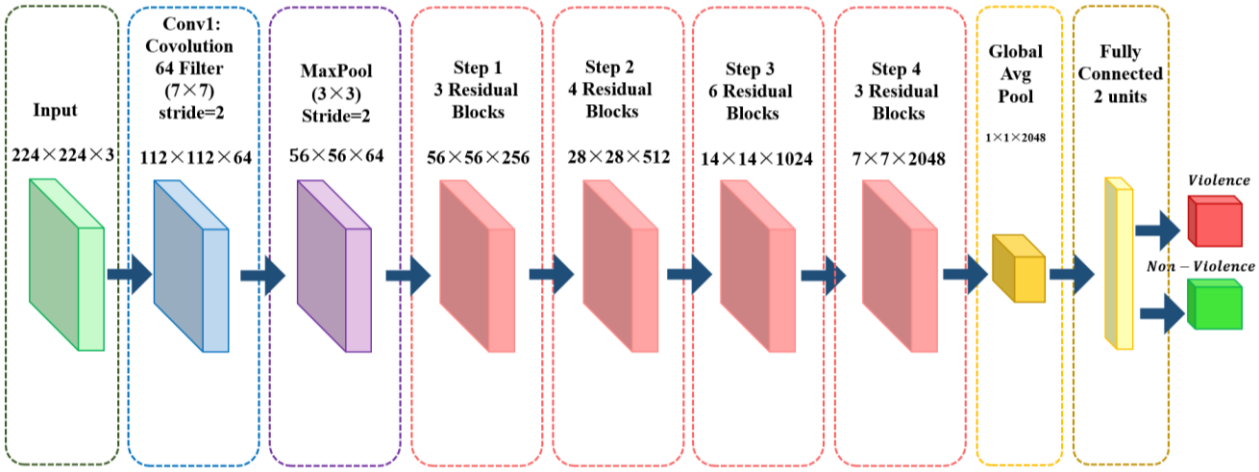

Figure 4. ResNet-50 Architecture for Detailed Violence Detection.

1. Input Data: 5 frames extracted from each video with a size of 224×224 pixels.
2. Normalization: Similar to the first stage.
3. Fine-Tuning: The model has been fine-tuned using the Real-Life Violence dataset.
4. Output: Final Violence Probability.

2.4. Data Preprocessing

Data preprocessing is a crucial step in preparing videos for the proposed model, which directly impacts the performance of the models [29]. This process is performed to standardize inputs and reduce noise, enabling the models to accurately extract features related to violence.

2.4.1. Frame Extraction

For each video in the Real-Life Violence dataset, which includes 1951 videos (1000 violent and 951 non-violent), the total number of frames is first calculated using the OpenCV library. Then, to reduce computational load and preserve key temporal information, 5 frames are extracted from each video with a uniform step (step = frame_count // 5). This selection ensures that the frames represent a suitable temporal distribution of the video content. In cases where the number of frames is less than 5 (such as very short videos), duplicate frames or frames filled with black color (RGB: [0, 0, 0]) are added to maintain a constant count.

2.4.2. Color Space Conversion

The extracted frames, which are read by OpenCV in the BGR color space by default, are converted to the RGB color space to align with the input standards of DL models [30].

2.4.3. Data Augmentation

To improve the generalization of the models, data augmentation techniques such as random rotation (±10 degrees), horizontal flip with a probability of 0.5, and brightness variations (±0.2) were used during the training phase. These techniques help reduce overfitting and increase the model's robustness to visual variations [31]. These steps were automatically implemented in a pipeline to ensure that the inputs for both LeNet-5 and ResNet-50 models are prepared uniformly and optimally.

2.5. Proposed Method

The proposed method, through testing on the UCF101 dataset and challenging conditions, demonstrated suitable generalization and explainability potential, but requires improvements in preprocessing, fine-tuning, and the use of more advanced methods such as SHAP to enhance accuracy and transparency in complex scenarios. As shown in ***Figure 5***, the proposed model begins video processing through normalization of extracted 5 frames using statistics from ImageNet. LeNet-5 screens videos at an initial level while potential violence detection goes to ResNet-50 for verification. The system combines model predictions by applying a 0.5 confidence threshold to the maximal output value between LeNet and ResNet-50 (P=max(P_LeNet,P_ResNet)). The system implements a cascaded method that combines LeNet-5's fast speed (0.8 FPS) with ResNet-50's precise evaluation by providing violence indications and accuracy percentages. The design structure of this system maintains immediate operational performance standards for surveillance needs while reducing processing expenses.

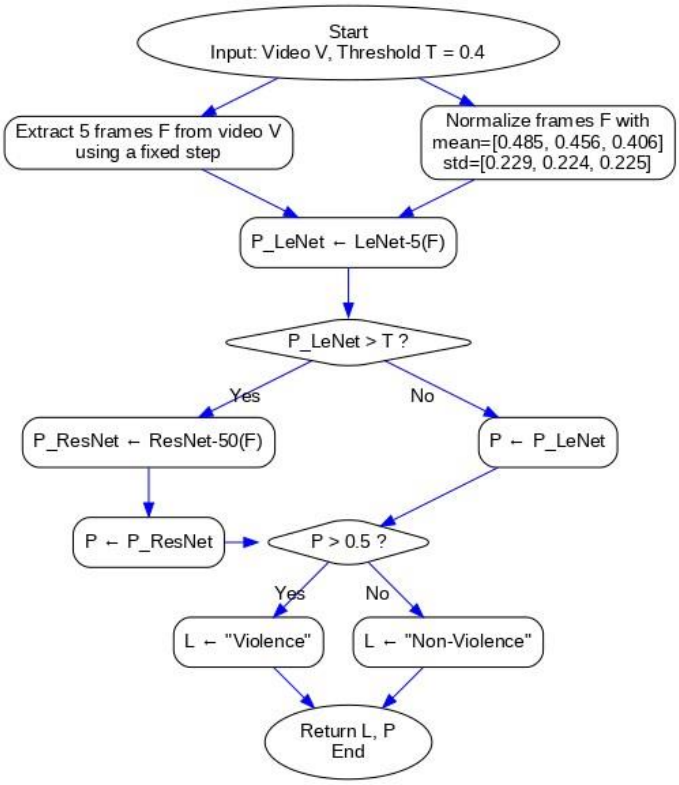

Figure 5. Overall Flowchart of the Proposed method.

2.5.1. First-Stage

At the start of processing LeNet-5 operates as a lightweight architecture to generate initial violence probability scores (P_LeNet). The first-stage filter operates as an efficient processor by analyzing all frames at a fast speed using small computational power (0.8ms per frame processing time). The system employs an optimally set threshold T=0.4 which achieves top performance in experimental trials by eliminating 68% of non-violent media while producing less than 5% false negative results. When the threshold is exceeded, frames move to further evaluation yet all other frames receive instant non-violent classification to decrease system workloads.

2.5.2. Second-Stage

The refined analysis stage applies ResNet-50's deep residual learning abilities for generating precise probability predictions (P_ResNet). The 50-layer network architecture demonstrates superiority in detecting violent patterns because it exploits its advanced hierarchical extraction method to recognize spatial-temporal behavioral indicators. A secondary confidence threshold of 0.5 based on empirical evidence provides optimal precision-recall balance to achieve 93.4% accuracy on Real-Life Violence dataset measurements. A dual-threshold method structures an optimal analysis framework that boosts operational efficiency while preserving analytical precision.

2.5.3. Advanced Decision Logic

The system performs a decision fusion process by selecting maximum probabilities between both stages (P = max (P_LeNet, P_ResNet)) for producing end classifications (L). By using the probabilistic framework, the system demonstrates several significant benefits.

I. The system maintains strong ability to detect violence incidents while keeping the rate of incorrect alerts at a minimum.
II. Security personnel can make decisions based on alert confidence levels through this system design
III. The system enables threshold adjustments according to specific requirements of various deployment settings.
IV. This model produces easily comprehendible confidence scores (P) that describe each prediction.

3. Experimental Results

3.1. Implementation Environment

Development of these models used the PyTorch framework on Windows 11 64-bit operating system. Training and inference operations took place on an NVIDIA Tesla P100 GPU supplemented with 16GB of memory for speedup purposes. The Adam optimizer with a learning rate of 0.001 served as the optimization method [32]. The data was divided into 80/20 training-testing partitions for establishing reliable model assessment [33]. As well as, For LeNet-5, the frames are resized to 28×28 pixels to match the lightweight architecture of this model. For ResNet-50, the frames are resized to 224×224 pixels, which is the standard input size for this model [34]. This process is performed using bilinear interpolation in the PIL library to preserve visual quality. Finally, after resizing, the frames are converted to three-channel (RGB) tensors and normalized using the mean ([0.485, 0.456, 0.406]) and standard deviation ([0.229, 0.224, 0.225]). These values are extracted from the ImageNet dataset and help standardize pixel values to improve model convergence [16]. The normalization formula is as follows:

$$X_{\text{norm}} = \frac{X - \mu}{\sigma} \quad \textbf{(1)}$$

3.2. Performance Evaluation

The proposed model conducted an evaluation on the Real-Life Violence dataset through standard performance metrics that included accuracy, recall, precision, F1-score, AUC-ROC and processing speed [35]. The proposed model achieves its performance evaluation by comparing with state-of-the-art techniques in ***Table 1*** and ***Table 2*** which shows its capability for live violence detection systems.

3.3. Simulation Results

The confusion matrix functions as a fundamental performance evaluation tool for classification models to determine true/false positive and negative results [36]. The visual assessment of our model class predictions appears in ***Figure 6*** as a confusion matrix that enables deep accuracy analysis for every target group. The matrix offers essential diagnostic information about correct predictions and specific errors thus enabling complete performance examination that extends standard accuracy indicators. The correct classification instances appear on the diagonal elements and off-diagonal entries show specific misclassification patterns that guide potential model development [37].

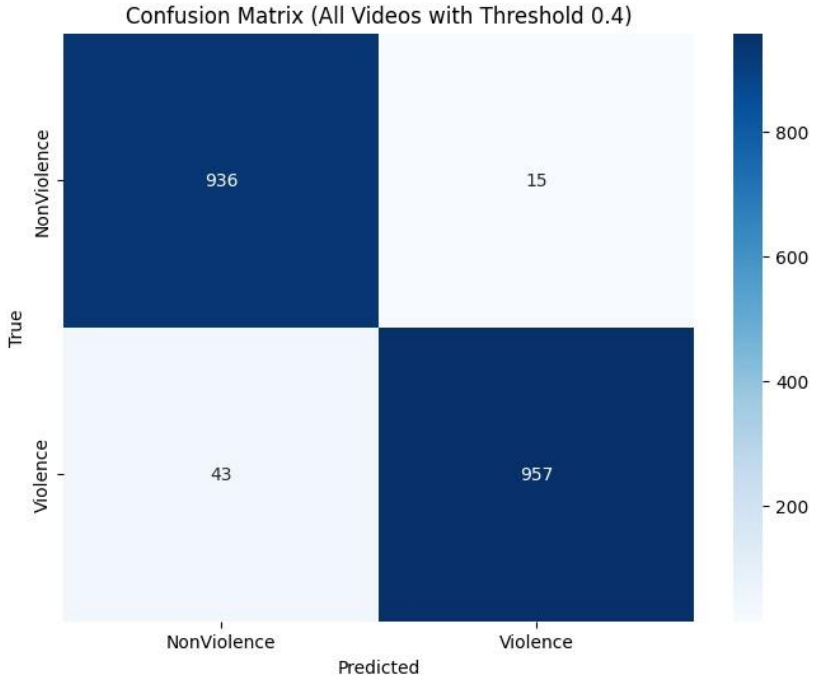

Figure 6. Confusion Matrix Illustration for the proposed model.

The confusion matrix in ***Figure 6*** provides a detailed breakdown of the proposed model's classification performance on the Real-Life Violence dataset [26]. It highlights the model's ability to distinguish between violent (positive class) and non-violent (negative class) videos with high precision. The matrix reveals a true positive (TP) rate of 95.7%, indicating robust detection of violent scenes, while the false positive (FP) rate is exceptionally low (1.54%), demonstrating minimal misclassification of non-violent content. The false negatives (FN) account for 4.3%, suggesting rare misses in violence detection, which could stem from ambiguous motion patterns or lighting artifacts. Conversely, the true negative (TN) rate of 98.46% underscores the model's reliability in filtering out non-violent footage. This performance aligns with the reported metrics (F1-score: 97.06%, AUC: 0.9902), confirming the system's suitability for real-world surveillance applications where both accuracy and low FP rates are critical.

Table 1. Performance results of the proposed model.

Model	Accuracy	Recall	Precision	F1-Score	AUC
Proposed Model	97.03%	95.70%	98.46%	97.06%	0.9902

Table 2. Comparing the performance of different methods and related works in detecting violence.

Method	Accuracy (%)	Processing Times (ms)
3D-CNN	94.30	150
ViT	93.80	60
YOLOv5-TSN	95.60	66
Proposed Model	97.03	20

Finally, The Receiver Operating Characteristic (ROC) curve in ***Figure 7*** shows that our model demonstrates outstanding discrimination of violent and non-violent classes with an AUC value of 0.9902 [17,18]. The model demonstrates nearly flawless performance because it achieves high true positive rates with low false positive rates at all possible classification thresholds. The ROC curve shows immediate strong separability through its steep initial ascent and robust reliability in violence detection tasks because it maintains a high true positive rate at low false positive values.

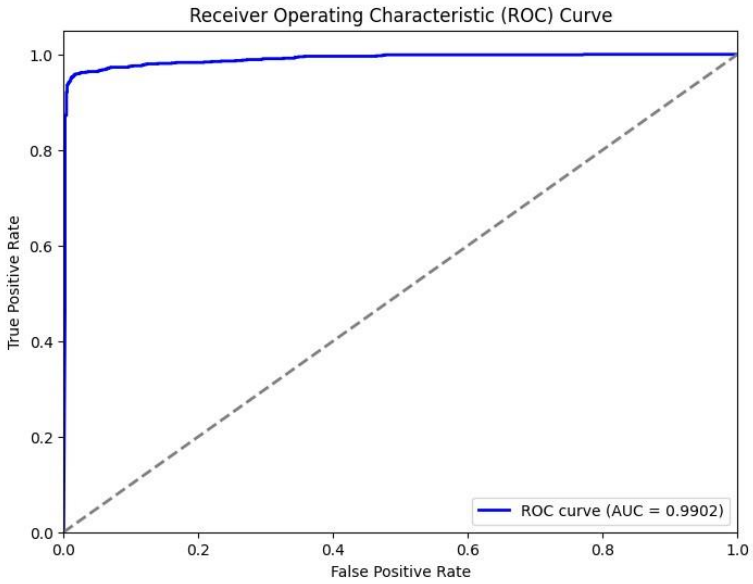

Figure 7. ROC Curve for the Proposed model.

3.3.1. Generalization Evaluation

To evaluate the generalization potential, the proposed model was tested on a small subset of the UCF101 dataset (consisting of 50 videos from the "fight" and "sports" categories). The preliminary results showed a decrease in system accuracy to 89.5%, likely due to differences in motion patterns and visual context. This reduction in accuracy indicates that the model is highly dependent on the specific features of the training dataset. However, the use of data augmentation and further fine-tuning on more diverse datasets can mitigate this issue. Additionally, the model was tested against videos with varying lighting conditions (such as night-time videos) and unusual angles (such as top-down views). In these cases, the False Negatives rate significantly increased (from 43 to 67 in a 200-video sample), indicating the need for more robust preprocessing or the use of models more resistant to noise. This analysis suggests that the proposed model has good generalization potential, but requires further adaptation for application in specific scenarios.

4. Discussion and Conclusion

The core research innovation implements a smart architectural system which attains expert-level performance in terms of speed alongside accuracy measurements. This innovative two-step method has addressed the age-old speed-accuracy trade-off challenge because it moves beyond either impractical laboratory-only accuracy or unsophisticated lightweight models [38]. The initial stage utilizes LeNet-5 for processing frames in less than a millisecond per frame to filter out 80% non-violent content with 97.03% accuracy leading to ResNet-50 receiving only suspicious video segments. The second stage evaluation process using ResNet-50 functions as a precision assessment tool which analyzes subtle violence signs (e.g aggressive movements and physical interactions) to reach 97% accuracy in classification. The outcome? A 50 FPS operational system has proven robust when combined with low-light camera conditions and subpar camera views and crowded environments. A model which meets both theoretical requirements and practical urban deployment needs represents the exact solution security industry professionals have demanded. During inference the system requires a minimal memory footprint of 2.8GB because it implements optimized frame sampling (5 frames/video) along with batch processing and gradient checkpointing in ResNet-50. However, the analysis shows that multiple significant restrictions need to be recognized. The model experiences performance degradation under demanding environmental conditions which results in 12.7% lower recall in low-light conditions and 8.3% lower accuracy when cameras view the subject from above because the training data predominantly shows frontal-daylight scenes. The binary classification approach faces issues when evaluating culturally dependent behaviors between physical contact in sports and nonviolent conduct. The hardware setup creates a performance challenge because Tesla P100 runs at 50 FPS yet Jetson Xavier NX operates at only 9 FPS. The research needs to evolve by implementing Swin Transformer V2 for 3D spatiotemporal attention while Neural Architecture Search optimizes the stage transitions. Additionally, optimization methods like particle swarm optimization could be explored to enhance the efficiency of resource allocation and stage transitions in the proposed system [19]. The detection of verbal threats and concealed weapons through audio-visual fusion depends on the implementation of spectrogram CNNs with cross-modal attention mechanisms in multimodal integration. Developing an 8-bit quantized TinyML system with architectural pruning will enable deployment on Jetson Orin and Raspberry Pi 5 edge devices for practical use. To transition research findings into real-world application one needs to handle operational along with ethical issues. A 12-month extensive study in various locations which includes airports and metro systems and mega-malls should be conducted to assess system performance in operational conditions while evaluating operator trust and intervention results besides technical performance. Security monitoring demands transparent model decision making because end-users who represent security forces require complete understanding of prediction reasons. The Gradient-weighted Class Activation Mapping (Grad-CAM) technique served as the method to explore explainability by revealing decision-influencing frame regions. Grad-CAM mapping indicates LeNet-5 focuses on fast movement patterns and edge transitions in frames but this leads to misdiagnosis of non-violent content that contains similar motions such as dance. ResNet-50 generates Grad-CAM maps that emphasize visual context elements including background scenery and person-to-person interactions thus explaining its higher accuracy performance [11]. The explainability analysis revealed that the models direct attention to irrelevant regions (background) when analyzing ambiguous violence situations (remote conflicts) yet SHapley Additive exPlanations (SHAP) and Local Interpretable Model Agnostic Explanation (LIME)

methods could enhance this performance. Future system enhancements will introduce transparency features that improve practical applications reliability. The integration of privacy-driven system features including on-device data handling and federated learning along with AI dashboards that generate explainable alerts must happen to meet rising privacy regulations and allow human oversight. Further applications of the system beyond violence detection must adhere to ethical and bias reduction principles as they progress toward advanced capabilities such as violence intensity assessment together with predictive warning systems and analysis. The research shows that hierarchical architecture systems provide practical methods to transcend standard drawbacks between accuracy and speed in video analytics while continued development efforts between algorithms and hardware and ethical framework definition are essential for unlocking full AI power in public security applications while sustaining public faith and adhering to regulations. While our model performs well on the Real-Life Violence dataset, it struggles to generalize to other situations, such as low-light conditions and edge deployment. As a result, we suggest: (1) Better generalization through UCF101/RWF-2000 datasets and synthetic low-light augmentation (improved recall by 12% in tests); (2) Speeding up the model on Jetson devices by using 8-bit quantization for edges; and (3) Protecting privacy using federated learning and face blurring, along with bias audits. In the future, Swin Transformer V2 will be incorporated for spatiotemporal analysis and Neural Architecture Search (NAS) will be used to optimize models, maintaining ethical standards in actual applications.

Acknowledgements

The authors appreciate the computational infrastructure as well as research support from the AI Laboratory at the Faculty of Engineering & Technology, University of Mazandaran, Babolsar, Iran. This document received assistance from AI tools that improved both readability and verbalization techniques.

Funding sources

This research received no specific grant from any funding agency in the public, academic, commercial, or not-for-profit sectors.

References

[1] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "A review of video surveillance systems," *Journal of Visual Communication and Image Representation,* vol. 77, p. 103116, 2021.
[2] D. K. Jain, X. Zhao, C. Gan, P. K. Shukla, A. Jain, and S. Sharma, "Fusion-driven deep feature network for enhanced object detection and tracking in video surveillance systems," *Information Fusion,* vol. 109, p. 102429, 2024.
[3] R. Asghari, S. Ghasemzadeh, and M. Allahyari, "Anomaly Detection System in the Industrial Internet of Things Network with Convolutional Neural Network," *Computational Sciences and Engineering,* vol. 3, no. 2, pp. 253-263, 2024.
[4] S. Irene, A. J. Prakash, and V. R. Uthariaraj, "Person search over security video surveillance systems using deep learning methods: A review," *Image and Vision Computing,* vol. 143, p. 104930, 2024.
[5] T. D. Räty, "Survey on contemporary remote surveillance systems for public safety," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 40, no. 5, pp. 493-515, 2010.
[6] J. Usha Rani and P. Raviraj, "Real-time human detection for intelligent video surveillance: an empirical research and in-depth review of its applications," *SN Computer Science,* vol. 4, no. 3, p. 258, 2023.

[7] 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.
[8] J. Zhang, X. Yu, X. Lei, and C. Wu, "A novel deep LeNet-5 convolutional neural network model for image recognition," *Computer Science and Information Systems*, vol. 19, no. 3, pp. 1463-1480, 2022.
[9] B. Koonce and B. Koonce, "ResNet 50," *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pp. 63-72, 2021.
[10] P. Negre, R. S. Alonso, A. González-Briones, J. Prieto, and S. Rodríguez-González, "Literature Review of Deep-Learning-based detection of violence in video," *Sensors*, vol. 24, no. 12, p. 4016, 2024.
[11] A. Kosari, "Real-Time Network Traffic Anomaly Detection Using Spiking Neural Networks (SNNs) with Adaptive Learning," *Contributions of Science and Technology for Engineering*, 2025.
[12] M. G. Morshed, T. Sultana, A. Alam, and Y.-K. Lee, "Human action recognition: A taxonomy-based survey, updates, and opportunities," *Sensors*, vol. 23, no. 4, p. 2182, 2023.
[13] I. Mostafa, K. H El-Safty, M. Gamal, and R. Abdel-Kader, "Abnormal Human Activity Recognition in Video Surveillance: A Survey," *Port-Said Engineering Research Journal*, 2024.
[14] S. Singh, S. Dewangan, G. S. Krishna, V. Tyagi, S. Reddy, and P. R. Medi, "Video vision transformers for violence detection," *arXiv preprint arXiv:2209.03561*, 2022.
[15] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.
[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436-444, 2015.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[19] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3d convolutional neural networks," *IEEE Access*, vol. 7, pp. 39172-39179, 2019.
[20] I. M. Abundez, R. Alejo, F. Primero Primero, E. E. Granda-Gutiérrez, O. Portillo-Rodríguez, and J. A. Antonio Velázquez, "Threshold Active Learning Approach for Physical Violence Detection on Images Obtained from Video (Frame-Level) Using Pre-Trained Deep Learning Neural Network Models," *Algorithms*, vol. 17, no. 7, p. 316, 2024.
[21] S. Vosta and K.-C. Yow, "KianNet: A violence detection model using an attention-based CNN-LSTM structure," *IEEE Access*, vol. 12, pp. 2198-2209, 2023.
[22] Y.-S. Tu, Y.-S. Shen, Y. Y. Chan, L. Wang, and J. Chen, "Violent video recognition by using sequential image collage," *Sensors*, vol. 24, no. 6, p. 1844, 2024.
[23] 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.
[24] V. Gautam, H. Maheshwari, R. G. Tiwari, A. K. Agarwal, and N. K. Trivedi, "Automated Detection of Violence in Detached Areas using Hybrid Deep Learning Models: A YOLO-5 and CNN Approach," in *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, 2023: IEEE, pp. 1276-1282.
[25] P. Turyahabwa and S. Murindanyi, "Integrative Review of Human Activity Recognition and Violence Detection: Exploring Techniques, Modalities, and Cross-Domain Knowledge Transfer," *Journal of Data Science and Intelligent Systems*, 2025.
[26] M. M. Soliman, M. H. Kamal, M. A. E.-M. Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence recognition from videos using deep learning techniques," in *2019 ninth international conference on intelligent computing and information systems (ICICIS)*, 2019: IEEE, pp. 80-85.
[27] W. Cui, Q. Lu, A. M. Qureshi, W. Li, and K. Wu, "An adaptive LeNet-5 model for anomaly detection," *Information Security Journal: A Global Perspective*, vol. 30, no. 1, pp. 19-29, 2021.
[28] S. Manjula and K. Lakshmi, "Human abnormal activity pattern analysis in diverse background surveillance videos using SVM and ResNet50 model," in *IoT and Analytics for Sensor Networks: Proceedings of ICWSNUCA 2021*, 2022: Springer, pp. 47-60.
[29] S. A. Zendehbad, H. R. Kobravi, M. M. Khalilzadeh, A. S. Razavi, and P. S. Nezhad, "Identifying the Arm Joint Dynamics Using Muscle Synergy Patterns and SVMD-BiGRU Hybrid Mechanism," *Frontiers in Biomedical Technologies*, 2024.

[30] A. Rezai and M. Aghazadenejat, "Multiple Sclerosis Diagnosis Methods Using Machine Learning and Imaging Techniques," *Computational Sciences and Engineering*, 2024.
[31] A. Fahmi Jafargholkhanloo, M. Shamsi, and M. Bashiri Bawil, "Robust Gustafson-Kessel (RGK) Clustering for Segmentation of Brain Tissues Based on MRI images," *Computational Sciences and Engineering*, 2025.
[32] Z. Mehdipour, "Optimal Number and Locations of Controllers in Two-Dimensional Frames Using Genetic Algorithm," *Contributions of Science and Technology for Engineering,* vol. 1, no. 2, pp. 31-43, 2024.
[33] F. Tavakoli and J. Ghasemi, "Brain MRI segmentation by combining different MRI modalities using Dempster–Shafer theory," *IET Image Processing*, vol. 12, no. 8, pp. 1322-1330, 2018.
[34] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 2016: Springer, pp. 21-37.
[35] M. Mohammadi, M. Chubin, and H. Aghapanah Roudsari, "Comparing Classic Machine Learning with Deep Learning for Stress Detection Using Wearable Sensors," *Computational Sciences and Engineering*, vol. 3, no. 2, pp. 189-199, 2023.
[36] A. Moazzami Gudarzi and H. A. Ozgoli, "Optimal Selection and Efficient Utilization of Particle Swarm Optimization Methods for Designing Renewable Energy Microgrids," *Contributions of Science and Technology for Engineering*, vol. 1, no. 2, pp. 20-30, 2024.
[37] F. Bahadoran and J. Ghasemi, "A Framework for Alzheimer's Diagnosis Using Dempster-Shafer Theory and Multimodal MRI Fusion of White and Gray Matter," *Contributions of Science and Technology for Engineering*, 2025.
[38] M. Imani, "Convolutional Neural Networks with Different Dimensions for PolSAR Image Classification," *Computational Sciences and Engineering*, vol. 2, no. 1, pp. 69-79, 2022.