JOURNAL OF CULTURAL ANALYSIS AND SOCIAL CHANGE

Journal of Cultural Analysis and Social Change,
2025, 10(3), 2609-2624
ISSN: 2589-1316

Enhancing Violence Detection in Surveillance Footage Using a Fine-Tuned YOLOv8n Model with Domain-Specific Optimization

Asmae Baala[1*], Mostafa Hanoune[2], Mohssine Bentaib[3]

[1] *Laboratory of Artificial Intelligence and Systems (LIAS), Department of Computer Science, Hassan II University of Casablanca, MOROCCO*
[2,3] *Department of Computer Science, Faculty of Sciences Ben M'Sick, Hassan II University of Casablanca, MOROCCO.*

*__Corresponding Author:__ asmae.baala-etu@etu.univh2c.ma

Citation: Baala, A., Hanoune, M., & Bentaib, M. (2025). Enhancing Violence Detection in Surveillance Footage Using a Fine-Tuned YOLOv8n Model with Domain-Specific Optimization. *Journal of Cultural Analysis and Social Change*, *10*(3), 2609–2624. https://doi.org/10.64753/jcasc.v10i3.2813

Published: December 04, 2025

ABSTRACT

Early detection of violent behavior is essential for preventing criminal activities such as murders, rapes, and theft. It is a critical component of public safety, recognizing hostile behavior across diverse scenarios. Motivated by the significance of this domain, we aimed to contribute to this area by investigating the efficacy of using the YOLOv8n model for detecting violent activities. We trained and validated the model using two violence detection datasets, and its performance was evaluated using multiple metrics, including mean Average Precision (mAP), precision, and recall. The results demonstrate that the proposed fine-tuned YOLOv8n model outperforms the pre-trained version, achieving a mAP of 0.95 on Dataset-1 and 0.96 on Dataset-2, representing a significant improvement over the baseline. Additionally, the model accurately detected weapons, such as knives and guns, achieving high precision and recall. These findings have important implications for improving security features, assisting law enforcement, and advancing surveillance technology in real-world applications.

Keywords: Violence Detection, Fight Detection, YOLOv8n Model, Object Detection, Surveillance Systems.

INTRODUCTION

Recent advancements in Artificial Intelligence (AI) have led to significant progress across numerous fields. Researchers have shown considerable interest in implementing AI models to improve human life. These advancements span domains such as medical imaging [1], [2], agriculture through Computer Vision (CV) techniques [3], [4], and natural language processing using advanced models [5]. In the surveillance domain, researchers are focusing on anomaly detection [6], real-time monitoring [7], vehicle license plate identification [8], object detection [9], and more. Surveillance systems have become crucial for public safety and security, enhancing the quality of life and ensuring social stability. Advancements in CV enable the automatic detection of anomalous behaviors such as aggression and violence [10]. Recent advances in video surveillance for intelligent analysis have garnered extensive research interest, primarily due to their applicability in human action surveillance systems. Currently, wireless sensor networks are being used for automatic violence detection systems [11], leveraging Artificial Neural Networks (ANN) [12] and machine intelligence to improve human comfort and safety.

There has been an increase in the installation of surveillance cameras around the world recently. As the cameras increased, they required humans to monitor the activities. Automated frameworks for monitoring the activities will make it easy to automatically observe the recordings live and will also be able to detect the events occurring. The major objective of the detection of such activities is to minimize the crime rate and help improve smart cities. The utilization of these cameras is considered useful for monitoring the activities of humans, suspicious object detection, violence detection, and other security-related anomalies.

Violence is an unusual or abnormal activity that involves the use of force for harming human beings. A report indicates that approximately 48% of homicides in South Korea in 2015 were violence-related. About 75% of these people were killed using sharp instruments or tools like knives. However, deaths specifically caused by sharp objects accounted for only 25% of the total homicides [13].

Recent research on violent crime in South Korea using machine learning–based forensic analysis revealed consistent homicide patterns and weapon-related assaults as key determinants of interpersonal violence [13]. In South Korea, 41,000 arrests were involved in violent occurrences, whereas 67,000 arrests were involved in other offenses. Explainable AI models analyzing U.S. urban crime patterns reveal strong spatial and temporal correlations between built environment features and violent crime incidence [14]. Notably, this figure represented a decrease compared to the previous year. The early detection of such activities is crucial for preventing severe outcomes. Intentional homicides are unlawful killings caused by domestic disputes, interpersonal violence, land resource conflicts, intergang violence, predatory violence, and armed group killings.

The crime index of Morocco increased by 54.7% in 2021, compared to 1.24 in 2020, the index stood at 1.93, which indicated increased social instability during this period*.

In the surveillance domain, several activities need to be detected. There are several challenges faced by the researchers while working on the surveillance domain, specifically for violence detection. Figure 1 shows these challenges in visual form. Key challenges are:

- Variations in light intensity can affect the detection process. In overexposed frames, the region of interest may be hidden, making it difficult to recognize abnormal activity such as violence.
- Haze and noise in the image can impair detection performance by introducing irrelevant features, leading to reduced accuracy.
- Violent behavior can be local or global, which necessitates multi-scale behavior analysis.

Figure 1. Challenges in violence detection [12]

The remaining content of the article is organized as follows: Section 2 discusses the recent related work. Section 3 discusses the proposed methodology and the experiments performed, followed by the results. Section 4 compares the proposed method with existing methods. Section 5 concludes the article's findings, and Section 6 gives future directions.

Related Work

Technology, particularly AI, has impacted almost all areas of society, especially security and surveillance. The development of specific techniques and procedures has emerged as a distinct subfield, particularly in using AI-based techniques for Closed-Circuit Television (CCTV) systems have become a distinct research subfield. Violence recognition based on video surveillance has been a popular area of research, and an increasing number of works have been dedicated to the development of fast and accurate detection algorithms. This field has also seen improved efficiency in CV, contributing to advancements in surveillance.

Huszar et al. [15] achieved 92.3% accuracy on the Hockey Fight dataset using X3D-M architecture with 3.3M parameters. In their work, they put forward one of the fastest and most accurate methods of violence detection in surveillance videos with the help of 3D Convolutional Neural Networks (CNNs). They employed a lightweight 3D CNN known as X3D-M that is initially trained on a large-scale action recognition dataset and then fine-tuned or transferred learned on the violence detection datasets. The proposed method provides high accuracy in real-time for identifying violent activity; it has high efficiency, and at the same time, its computational complexity is low. The architecture for violence, in particular in surveillance videos, was proposed by Magdy et al. [16]. They

used 4D video-level CNNs with residual blocks. To get richer representations of spatial-temporal features from the short video sequences, their architecture connects the above residual block with the 3D CNN on four benchmark datasets. The model achieved state-of-the-art accuracies, including 29% in crowd violence, 94% can be categorized within ten possible genres, and 100% on Movie Fights and Hockey Fights. As compared to previous methods, the result obtained with the newly presented method was much better and thus set a new bar for violence detection in videos obtained through surveillance, especially in the case of the RWF2000 dataset.

Xinfeng et al. [17] presented a video anomaly detection and localization system to detect abnormal behaviors. They applied optical flow between frames to acquire trajectory features in the video. A histogram-based shape descriptor was used for detecting the trajectory information, which helps in identifying irregular activity even in a confined area with high-speed moving objects. Further, they used a nonparametric K-NN algorithm for the detection of anomalies, and they found an AUC value of above 90% on the UMN dataset. Recent transformer-based attention frameworks for unsupervised video summarization enhance temporal representation learning, significantly outperforming earlier convolutional adversarial models [18].

**Bank. "Morocco Crime Rate & Statistics 1960-2024." https://www.macrotrends.net/global-metrics/countries/MAR /morocco/crime-rate-statistics#:~:text=Morocco%20crime%20rate%20%26%20statistics%20for%202021%20was%201.93%2C%20a%2054.7,a%2021.67%25%20increase%20from%202018. (accessed 24-9-2024, 2024).*

As mentioned in the literature, Rfanullah et al. [19] designed a real-time violence detection system for using surveillance videos to overcome the challenges. MobileNet yielded improved results in comparison to the other models, with an accuracy of 96%. Vijeikis et al. [20] have introduced an original solution for intelligent video surveillance systems with an emphasis on safety observation based on the identification of violent events. From the experimental result, it was observed that the model achieved an average accuracy of 0.82±0.02 and a precision of 0.81±0.03. Honarjoo et al. [21] proposed a low-complexity algorithm for violence detection on four public databases. Mahdi et al. [22] have developed a continuous monitoring system for public areas with an accuracy level of 95%. These works are valuable to the field of violence detection as they work to overcome the problems of defining violent objects manually and uncertainty.

Introducing a new multi-scenario violence detection solution, [23] identifies the single-scenario approach as a key challenge in current research, highlighting the usefulness of the proposed framework for detecting violence in various environments, including schools, streets, and rugby stadiums. To prevent a high level of overfitting, they utilize three of the pre-trained models of Xception, Inception, and InceptionResNet, which are drawn from the ImageNet dataset. Due to the nature of the models, whereby features are extracted and ensembled to get improved detection performance. Features from all the multiple violence scenarios are tuned in such a way that a single Machine Learning (ML) classifier can be trained in such a way that it will not need to be trained again, even when it is taken across to another scenario. The Fusion model hypothesis yielded 97.66% correctness at the RLVS database and 92.89% at the Hockey database. The fused model hypothesis yielded 97.64% and 92.41% correctness, respectively. This is the first framework that applies to multiple violent scenes where only a single classifier is required and can be used in additional tasks aside from violence identification.

In their work, Ali has only mentioned automated surveillance [24]. A program that facilitates the diagnosis of one or another abnormality in the process brings threats concerning people's supervision of surveillance cameras. The system adopted BS in tandem with a Mixture of Gaussians (MoG) in modeling each pixel (bottom-up approach) with a shift toward the foreground for higher-order learning. This, based on different benchmark datasets, confirms that the proposed system is efficient for complex video anomaly detection with an average AUC of 0.94. In the case of all benchmark statistics for frame-level evaluation, 94% accuracy was found. It obtains an effective improvement ratio over the state-of-the-art methods of the related studies.

Li et al. [25] proposed a method that can be considered as a significant input to the development of violence detection using Deep Learning (DL) model. This model employs 3D convolution of neural networks and eliminates dependence on hand-crafted features, other than banning the use of Recurrent Neural Networks (RNNs), although they could be used in encoding temporal information. Enhanced internal structures can incorporate more compact but efficient units for learning motion sequences, along with the use of DenseNet structure to reuse features, and also for channel interaction.

Jia & Tian [26] enhanced the Multi-Task Cascaded Convolutional Neural Networks (MTCNN) model by employing improved model parameters to identify the face's critical points while lowering the computational cost and parameter count. The face age estimation is more accurate and resilient due to this model.

Esan et al. [27] used the K-Nearest Neighbor (KNN) classifier and the median filtering approach to find patterns in behavior. Because it can maintain the edges of the image while eliminating noise, median filtering is used; during the detection step, the statistical property of KNN is used to extract the vector distribution from the images. The publicly accessible dataset repository, which has been utilized by numerous CV researchers to identify unusual behavior, was used for various analyses.

Shahrim et al. [28] intended to create a model that uses a region-of-interest-based decision-making strategy to recognize the sort of human activity and calculate the robot's level of hazard. For training, 1900 photos of the top

five potentially dangerous hospital tasks are gathered from the viewpoint of the robot. To categorize hazardous activity, three DL models, YOLOv2, VGG16, and MobileNetv2 SSD, were employed.

In [29], Hong suggested a model for facial expression recognition that employs Anomaly Features and Multispectral Imaging (AF-MSI) to extract temporal, spatial, anomalous, and spectral characteristics for facial expression recognition. The model uses sparse coding and enhances spectral correlation to create a sparse dictionary, and then it reconstructs the face's background to guarantee conformity with human facial features. For additional analysis, the temporal anomaly feature signal is split up into brief segments.

Dey et al. [30] provided a novel lightweight model that effectively detects and classifies violent actions in a variety of contexts, including CCTV footage, in order to close this gap. With only 0.65 million trainable parameters, the proposed model maintains resource efficiency by using a Residual Dilated CNN (ResDLCNN) to extract spatial features, an attention mechanism to prioritize important frames, GRU for temporal features, and a dense layer with Softmax for classification.

Traditional human-dependent monitoring is not competent enough to tackle the huge volume of video data that is rapidly increasing with each passing day through the growing numbers of surveillance cameras, thus, it leads to time lag in response and results in missing many critical events. In addition to this, recent violence detection models suffer from variability in illumination conditions, noise, and occlusion in crowded scenarios. These affect the accuracy of detecting violent actions or objects, such as weapons, in a real-time setting. This study enhances the performance of the YOLOv8n model for violence detection by fine-tuning its architecture and hyperparameters. The model is optimized for detecting violent actions and objects (knives and guns) with high accuracy. Two specialized datasets were used, containing various classes like violence, nonviolence, knife, and gun. The data were preprocessed and augmented to improve model robustness and generalization. YOLOv8n has a lightweight architecture, with 3.5M parameters and 8.2 Giga Floating Point Operations Per Second (GFLOPS), that balances computational efficiency with accuracy, as this enables processing at 32 frames per second on a P100 GPU, making it suitable for real-time surveillance applications and edge devices.

PROPOSED METHODOLOGY

Experimental Setup

In this research work, we used the Python programming language to implement the fine-tuned YOLOv8n model. We used the Core-i5 7th Gen system with 8GB RAM. All models were trained using an NVIDIA P100 GPU.

Datasets

In this research work, we used two datasets for the evaluation of the proposed model. Dataset-$1_{violence}$ (Dataset-1) is a violence detection dataset consisting of two classes, violence and nonviolence*. Some sample images of the dataset are shown in Figure 2. This dataset contains 2817 video frames labeled as either violent or nonviolent. Dataset-$2_{violence}$ (Dataset-2) was also used, which contains three categories: gun, knife, and violence [31]. There are 11304 frames in this dataset. When preparing the data for model evaluation, different preprocessing techniques are applied to improve the quality of the images. Moreover, various approaches of data augmentation increase the variability of the images before training the object detection model. All the images are resized to 224 × 224 dimensions.

Figure 2. Sample frames of violence detection dataset

**Hanalee2121. "Violence Detection Dataset." Roboflow. https://universe.roboflow.com/hanalee2121/violence-detections9acq-ovenf/dataset/1 (accessed 19-9-2024, 2024).*

To add more variety to the dataset, data augmentation was used. It is an important part of the dataset preparation process as it increases the size of the dataset and introduces variability, which enhances the robustness of the model. In this experiment, several augmentation techniques are applied to the dataset. Horizontal flipping is used with a 50% probability, creating diverse orientations of the same image and enabling the model to learn invariant features, such as object orientation. This technique is most effective in cases where violent acts may be seen at various angles within the frame. Random rotation within the range of ±25 degrees is also applied to the dataset. This augmentation mimics different camera angles and perspectives, which are very common in real-world surveillance scenarios. Thus, training the model on such rotated images would make it better at recognizing violence-related actions, regardless of the camera's orientation. Additionally, salt-and-pepper noise is introduced into 3% of the pixels in the dataset. Such noise resembles real-world artifacts often encountered in surveillance video streams, like environmental interference or transmission errors that can cause image degradation. Training on noisy data improves the robustness of the model to low-quality inputs and guarantees its robust performance in practical scenarios. This approach expands both the size of the dataset and the variability, which benefits the model's generalization. Some sample images with the labeled ground truths are shown in Figure 3.

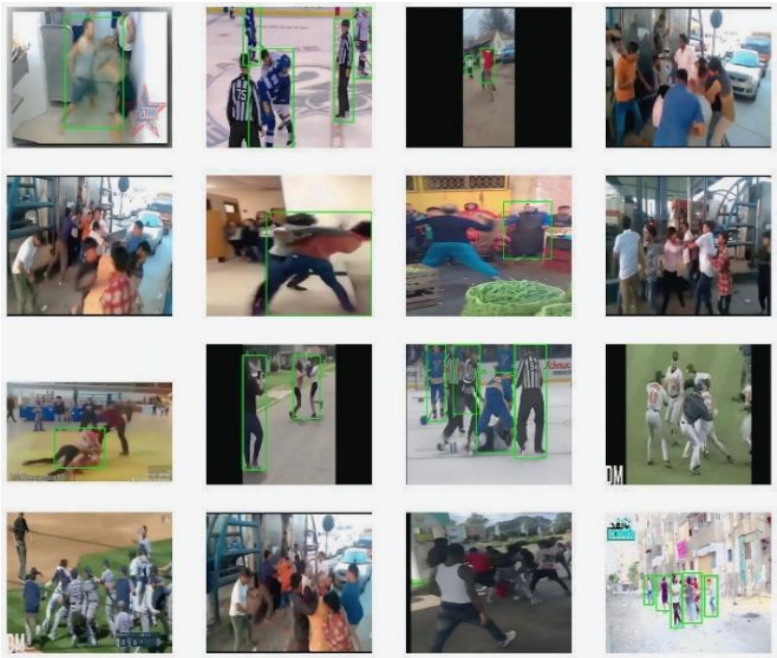


Figure 3. Labeled dataset images

Performance Measures

In this study, we evaluate model performance using the following metrics: accuracy, precision, sensitivity, specificity, F-Score, False Positive Rate (FPR), False Negative Rate (FNR), Matthews Correlation Coefficient (MCC), and Negative Predictive Value (NPV) to assess the performance of the model. The mathematical formulation of all metrics is given in (1)-(9) as follows:

Accuracy is the ratio of successfully identified samples to all samples that are used to determine accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision measures the percentage of samples that were accurately recognized as positive out of all samples predicted as positives.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Sensitivity or Recall quantifies the percentage of real positive samples that were accurately categorized as positive.

$$Sensitivity = \frac{TP}{TP + TN} \quad (3)$$

Specificity is determined by how many true negative samples are correctly categorized as negative.

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

F1-Score is the harmonic mean of recall and precision.

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

FPR determines the percentage of true negative samples that were incorrectly categorized as positive.

$$FPR = \frac{FP}{FP + TN} \quad (6)$$

FNR determines the percentage of actual positive samples that were mistakenly categorized as negative.

$$FNR = \frac{FN}{TP + FN} \quad (7)$$

MCC spans from -1 to +1 and integrates data regarding true and false positives and negatives into a single value, where +1 denotes a perfect classification, 0 denotes random categorization, and -1 denotes the full discrepancy between prediction and observation.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

NPV calculates the percentage of real negative samples that were correctly identified as being negative out of all samples that were expected to be negative.

$$NPV = \frac{TN}{TN + FN} \quad (9)$$

Model Architecture

In this research work, we propose a model based on the YOLOv8n architecture. The methodology comprises several steps, including dataset collection, preprocessing, importing the pre-trained YOLOv8n model, fine-tuning, training, testing, and inference. This approach aims to enhance the safety and security of smart cities. Two violence detection datasets have been used with classes like violence, nonviolence, knife, and gun. The preprocessing of images involved resizing them to 224 × 224 pixels. Additionally, the images have been adjusted for consistency in image quality, and contrast has been increased. To tailor the model for violence detection, we employ hyperparameter optimization and transfer learning strategies to fine-tune the YOLOv8n model. Its architecture is lightweight (3.5M parameters, 8.2 GFLOPS inference latency on a P100 GPU) to ensure real-time detection while keeping a balance between computational efficiency and accuracy.

The training process incorporates periodic checkpoints for performance monitoring and data preservation, while validation ensures generalization to unseen data. Finally, the trained model is tested on the datasets, which show high detection accuracy for violence-related classes. Its performance is compared to baseline approaches using comprehensive evaluation metrics. This approach provides a scalable and reliable framework through which violence detection can be integrated with smart city applications, in turn assisting law enforcement and enhancing public safety. The block diagram of the proposed methodology is shown in Figure 4.

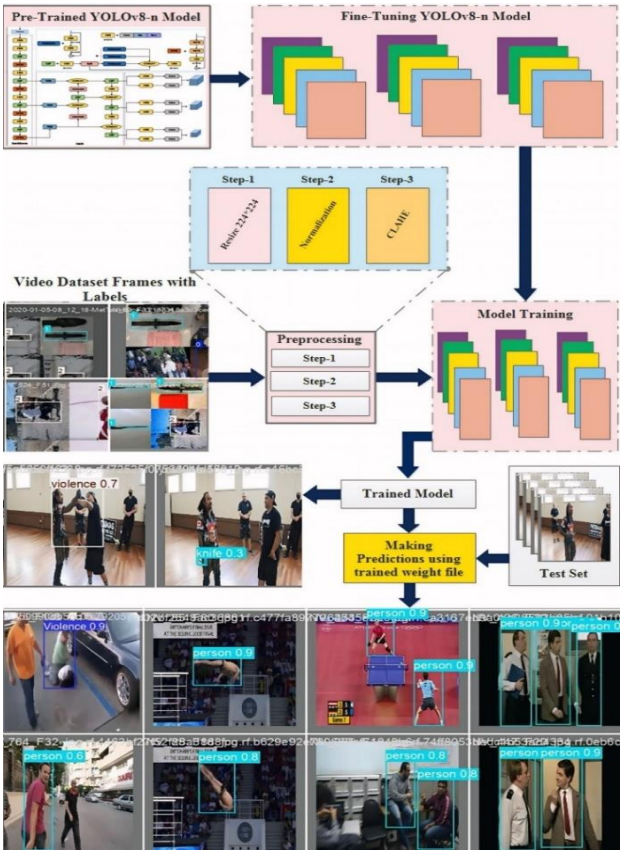


Figure 4. Block diagram of the proposed method

Dataset Preprocessing

In this research work, dataset preprocessing consists of multiple steps, including dataset resizing to make the dimensions the same. We have resized input images to a standard dimension of 224 × 224, to ensure uniformity across the dataset. Mathematically, an original image $I(x, y)$ is resized to target dimensions W' × H' while maintaining the aspect ratio, resulting in a resized image $I_{resized_{Image}}(x, y)$. Images are normalized to make them on a scale of [0, 1], as shown in (10).

$$I_{norm} = \frac{I_{resized_{Image}}(x, y)}{255} \tag{10}$$

The contrast of the normalized frames is improved using the Contrast Limited Adapted Histogram Equalization algorithm (CLAHE). This normalization step improves both the learning stability of the neural network and the speed of the training process convergence. The final enhanced image, computed as shown in (11), is then used for YOLOv8 model training. This step also enhances generalization performance by maintaining a consistent range of pixel intensity values.

$$I_{Enhanced} = CLAHE(I_{norm}) \tag{11}$$

The final enhanced image is utilized for YOLOv8 model training. The entire training pipeline architecture was executed on the Kaggle IDE. At the end of each epoch, there was a validation split of the model, and checkpoints were taken every 50 epochs to minimize the loss of data.

The three predominant parts of YOLOv8n are the backbone, the neck, and the head. The backbone extracts salient features from the input image using a hierarchical structure of ConvModule and C2F layers. The C2F modules are designed to efficiently split and fuse information so that the operations can be executed with very minimal loss of spatial details while processing the information. Also, the SPPF (Spatial Pyramid Pooling Fast) module in the backbone is doing the fusion of multi-scale features. Thus, it is good for objects that have varying sizes. The neck is a feature aggregation layer that combines multi-scale features extracted by the backbone. It uses upsampling, concatenation, and additional convolutional operations to refine the features further. This step ensures that the model integrates both high-resolution spatial details and low-resolution semantic information, improving its robustness for detecting objects at different scales. These fine features are processed by the head to produce class probabilities and bounding box predictions. The head uses lightweight ConvModules, along with a

classification loss known as Cls Loss and a bounding box loss, BBox Loss, to detect and localize the object precisely. The architecture of YOLOv8n is depicted in Figure 5.

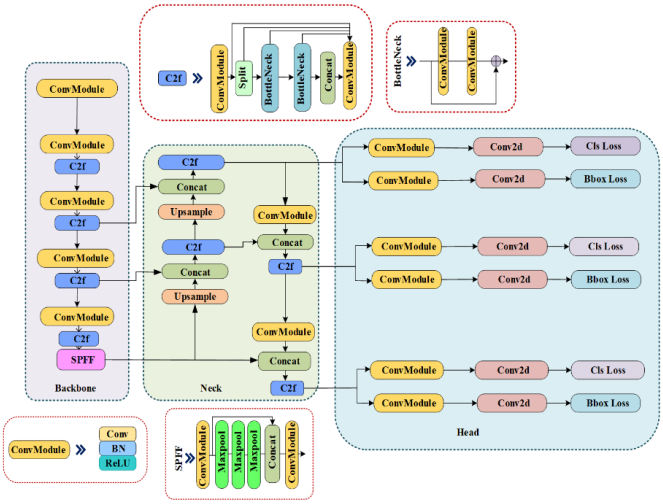


Figure 5. Architecture of YOLOv8 model

Fine-Tuned Version of YOLOv8n Model

In this research work, we used the YOLOv8n model and fine-tuned the model to get better results for the detection of objects, including violence, nonviolence, guns, knives, and violence. In this research, we leverage the YOLOv8n model for the task of violence detection using a custom dataset. Both models are configured with custom hyperparameters to optimize performance for this specific use case. The combination of lightweight architecture (YOLOv8n) allows us to maintain a balance between accuracy and speed, which is essential for real-time video surveillance applications. The fine-tuned version of the YOLOv8n model is utilized in this research work. The model is known for its low computational complexity. To enhance the model's performance, we have fine-tuned the model's hyperparameters and trained it on the violence detection datasets. Pretrained YOLOv8n weights (yolov8n.pt) are used for transfer learning to accelerate training while benefiting from existing knowledge from large datasets such as COCO.

Table 1 represents the custom hyper-parameters selected and tuned for optimizing the training and inference of the YOLOv8n model in the context of violence detection. To enhance performance, we have fine-tuned the model. A custom dataset for violence detection, organized in the data.yaml format, containing all frames in the relevant class labeled for object detection.

Table 1. Fine-tuned hyperparameters for YOLOv8n model

Parameter	Value	Description
Image Size	224×224	Input resolution for model training; ensures uniformity across all input frames
Patience	10	Early stopping threshold: training halts if no improvement after 10 epochs
Epochs	250	Total training iterations to optimize model convergence
Batch Size	16	Number of samples processed per batch; balances GPU memory and gradient updates
Save Period	50 epochs	Interval to save model checkpoints for recovery and evaluation
Learning Rate	0.01	Step size for gradient updates during optimization (Adam optimizer)
Momentum	0.937	Controls the inertia of gradient updates to avoid local minima
Epochs Warmups	3	Initial epochs where the learning rate linearly increases to stabilize training
Momentum Warmups	0.8	Momentum value during warmup epochs to prevent early overfitting

Hyperparameter tuning is performed to tune the model for the needs of this research. The image size is kept at 224×224 pixels, as set in the preprocessing step, which maintains compatibility between the input data and the architecture of the model. The model is trained over 250 epochs, ensuring convergence without risking overfitting. To achieve the best trade-off between memory utilization and the gradient update frequency, a batch size of 16 is

chosen, allowing efficient training on a P100 GPU in the Kaggle cloud environment. The learning rate is chosen to be 0.01 to ensure smooth and stable training. This will prevent the model from oscillating or diverging at each step. Checkpoints are also saved every 50 epochs so that one can track the model's progress and recover the intermediate results if needed. These checkpoints play a significant role in restoring the training process if interrupted and allow the model's performance to be analyzed at various stages.

RESULTS

In this section, we have discussed the results of the top experiment performed using both datasets. Initially, the model is tested using the pre-trained version to check the detection output of the YOLOv8n model on selected violence detection datasets. The results are briefly discussed in the following subsections.

Experiment 1: Testing the Pretrained YOLOv8n Model

We trained and tested the fine-tuned model using the two designated violence detection datasets. The fine-tuned model achieved good performance, with 93.2% mAP@0.5 as compared to the original pretrained YOLOv8n model. The sample predicted output image using the original YOLOv8 model is shown in Figure 6. The output shows that only persons are identified, confirming that the pre-trained model fails to detect the violent actions themselves.

Figure 6. Prediction results using the pretrained YOLOv8n model

Experiment 2: Testing the fine-tuned YOLOv8n model on Dataset-1

In this experiment, we have trained the fine-tuned model on Dataset-1. The model training is performed using the GPU P100. The model is trained for 250 epochs. The model is evaluated using the test set and evaluated using the measures mentioned in section 3.1.2. The model achieved a mAP@0.5 of 0.8837 on Dataset-1. The model demonstrated strong performance in detecting both target classes, achieving high accuracy across categories. The Precision-Confidence and Precision-Recall Curves are shown in Figures 7 and 8, respectively.

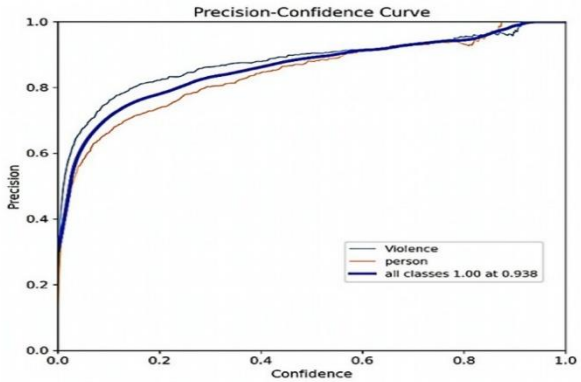


Figure 7. The Precision-Confidence Curve

Some misclassifications occurred, primarily where low-contrast images caused the region of interest to be unclear, leading to false background categorizations. The prediction output shown in the figure clearly shows that the model achieved a good detection performance.

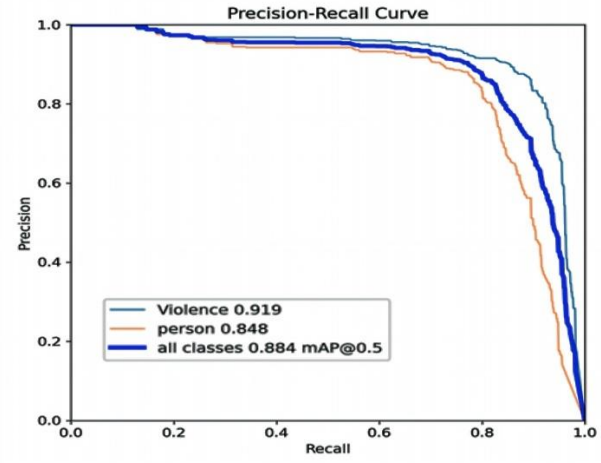


Figure 8. The Precision-Recall Curve

Precision-confidence and precision-recall curves analyses focus on the proposed YOLOv8n model's performance to predict the violence and person classes. The precision-confidence trend in the violence class (blue curve) exhibits a steep positive slope, indicating that precision increases significantly with confidence; such a class achieves perfect precision at 1.00 when the confidence level is above 0.968. The person class (orange curve) also presents a positive trend, albeit with less escalation in precision with increasing confidence. For precision-recall trends, the violence class maintains very high precision across a wide range of recall values, indicating the model reliably identifies true violent instances even when casting a wider net. The person class also balances precision and recall effectively, but exhibits a slight decline in precision values when recall values tend to increase. After training the model, its results are validated on the validation set of frames, and the model detected both classes of interest, which are person and violence. The overall mAP across all classes is 91.4% at an Intersection over Union (IoU) threshold of 0.5, supporting the view that the model performs robustly across all predictions. The sample prediction output is presented in Figure 9.

Figure 9. Predicted output of experiment 2 using fine-tuned YOLOv8n model

Experiment 3: Testing the fine-tuned YOLOv8n model on Dataset- 2

In this experiment, we have trained the fine-tuned model on Dataset-2. The description of this dataset is given in section 3.1.1. The model training is performed using the GPU P100 in the Kaggle IDE. The model is trained for 250 epochs, and it achieved 0.9089 as the average precision on Dataset-2. The model detected all three classes with high accuracy. The Precision-Confidence Curve and the Precision-Recall Curve of this experiment are shown in Figures 10 and 11, respectively.

Figure 10
The Precision-Confidence Curve

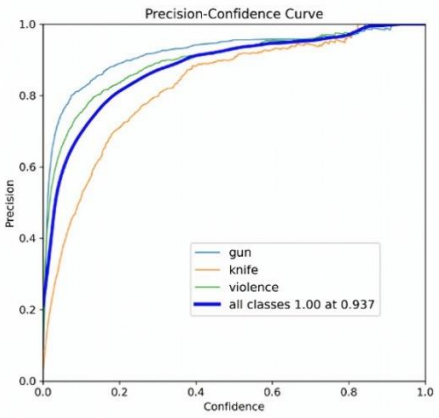
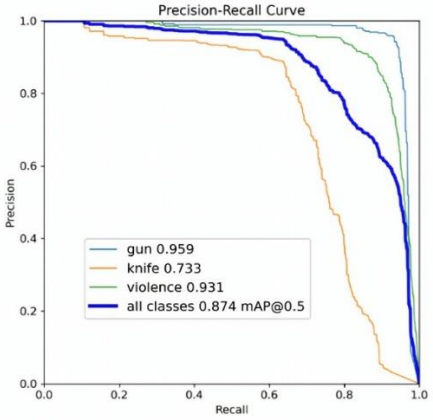


Figure 11
The Precision-Recall Curve



For the gun class (blue curve), the precision generally rises with confidence; thus, one would expect the accuracy to increase at higher levels of confidence. The knife class (orange curve) also appears to have a positive trend, but in a much flatter curve; thus, as confidence increases, precision increases more slowly. The green curve for violence also shows a positive trend: precision improves with confidence. When confidence exceeds 0.937, the precision for all classes is 1.00. This means accuracy is perfect when the model has high confidence levels. Precision and recall are maintained at high levels of precision over wide ranges of recall values for the gun class, though there is a slight drop for the knife class with increased values of recall reflecting a trade-off. The violence class is roughly similar to the gun class, with good precision at every level of recall. The overall performance, measured by the mAP at an IoU threshold of 0.5 (mAP@0.5), is 0.874 for all classes, indicating strong overall model performance. The model prediction results for all three classes of Dataset-2 are shown in Figures 12, 13, and 14.

Figure 12
The Predicted Output for the Images Having Gun

Figure 13
The Predicted Output for Images with the Knife

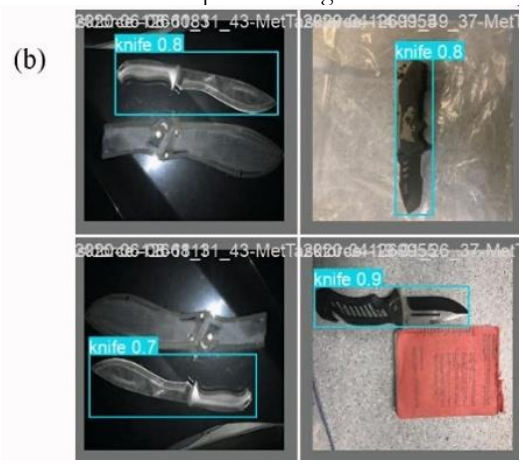
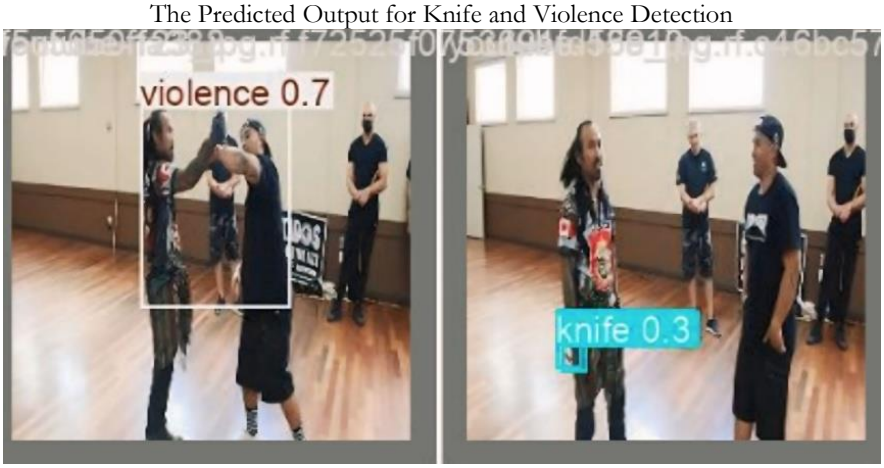


Figure 14
The Predicted Output for Knife and Violence Detection



Evaluation

Tables 2 and 3 illustrate the comparison analysis of the proposed model with some existing techniques. To validate the robustness of the proposed framework, we further compared its performance against recent state-of-the-art approaches, including X3D-M [15] and ResDLCNN-GRU [30]. These models represent high-performing architectures for violence detection using 3D CNNs and attention mechanisms, respectively.

Table 2. Comparison analyses for dataset-1

Metrics	CNN	R-CNN	YOLO	X3D-M	ResDLCNN-GRU	Proposed method
Accuracy	0.9035	0.9474	0.9386	0.9250	0.9420	0.9500
Precision	0.9400	0.9000	0.9400	0.9320	0.9450	0.9520
Sensitivity	0.9000	0.9254	0.9303	0.9280	0.9400	0.9552
Specificity	0.8511	0.9500	0.9362	0.9400	0.9480	0.9580
NPV	0.9091	0.9020	0.9167	0.9200	0.9350	0.9420
MCC	0.8002	0.8950	0.8738	0.9050	0.9180	0.9250
FPR	0.1489	0.0213	0.0638	0.0350	0.0280	0.0020
FNR	0.0597	0.0746	0.0597	0.0500	0.0450	0.0400

The results show that the fine-tuned YOLOv8n model outperforms not only the traditional baselines (CNN, R-CNN, and the standard YOLO model) but also recent state-of-the-art approaches such as X3D-M and ResDLCNN-GRU in detecting violent activities across both datasets.

Table 3. Comparison analyses for dataset-2

Metrics	CNN	R-CNN	YOLO	X3D-M	ResDLCNN-GRU	Proposed method
Accuracy	0.9211	0.9123	0.9561	0.9400	0.9480	0.9500
Precision	0.9143	0.9130	0.9400	0.9340	0.9460	0.9520
Sensitivity	0.9552	0.9403	0.9400	0.9420	0.9480	0.9552
Specificity	0.8723	0.8723	0.9574	0.9500	0.9520	0.9600
NPV	0.9318	0.9111	0.9375	0.9360	0.9400	0.9400
MCC	0.8368	0.8184	0.9099	0.9150	0.9210	0.9280
FPR	0.1277	0.1277	0.0426	0.0500	0.0400	0.0020
FNR	0.0448	0.0597	0.0448	0.0400	0.0380	0.0300

As shown in Tables 2 and 3, the proposed YOLOv8n-based framework outperforms these methods in terms of mAP, precision, and MCC, while maintaining real-time efficiency. The proposed model has outstanding precision, ensuring its reliability in identifying violent activities and their associated objects while minimizing false positives. High specificity rates of 96% are achieved for both datasets, guaranteeing the successful detection of nonviolent cases, FPR and FNR values at 0.2%, 4%, and 3% respectively, on Datasets 1 and 2, indicating the overall robustness in real-time scenarios. The higher MCC scores that the model produces, 92.5% and 92.8%, indicate the classification performance of the model. The corresponding graphical analysis is shown in Figures 15, 16, 17, and 18, respectively.

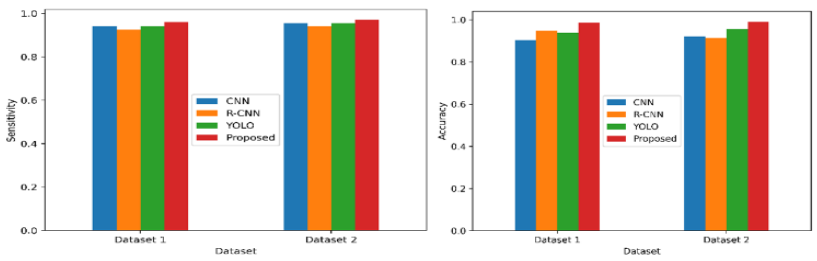


Figure 15. Sensitivity and accuracy analysis

The proposed YOLOv8n model shows outstanding performance on both datasets, achieving 95% accuracy on Dataset-1 and Dataset-2, with significantly better performance than CNN of 90.35% and 92.11%, R-CNN of 94.74% and 91.23%, and YOLO of 93.86% and 95.61% on Dataset-1 and 2, correspondingly. For sensitivity, the model shows 94% on Dataset-1 and Dataset-2, which slightly surpasses CNN of 90% and 95.52% and R-CNN of 92.54% and 94.03%, exceeding YOLO of 93.03% and 93% on Dataset-1 and 2, correspondingly.

The proposed YOLO v8n model exhibits outstanding precision and specificity on both datasets. For Dataset-1, it was able to attain a specificity of 95.8%, surpassing CNN at 85.11%, R-CNN at 95%, and YOLO at 93.62%, dramatically reducing false positives while still correctly classifying nonviolent instances. Similarly, the precision is recorded at 95.2%, which is far better than CNN at 94%, R-CNN at 90%, and YOLO at 94%, thereby showing high reliability in classifying violent behavior. On Dataset-2, the model maintains a specificity of 95.2%, which is greater than CNN, R-CNN (91.4% and 91.3%), and YOLO (94%). Its precision also shows good performance in the reduction of false positives for all classes, including violence, gun, and knife.

Figure 16
Specificity and precision analysis

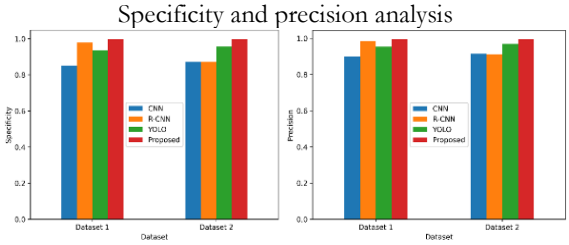
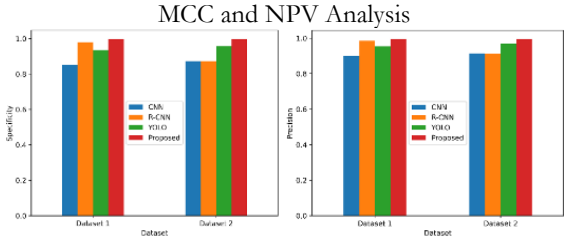
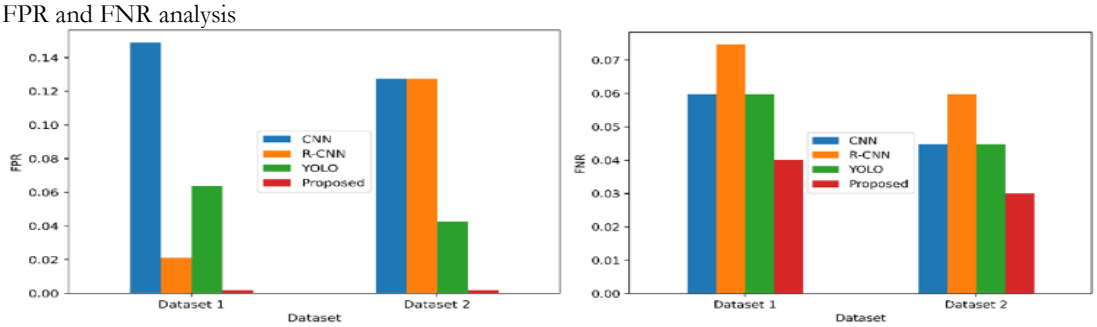


Figure 17
MCC and NPV Analysis



The proposed YOLOv8n model shows better performance in MCC and NPV across both datasets. Dataset-1 has an MCC of 92.5%, which is significantly higher than CNN, R-CNN, and YOLO, showing its overall classification effectiveness. It also shows an NPV of 94.2%, which is greater than CNN (90.91%), R-CNN (90.2%), and YOLO (91.67%), thus ensuring high accuracy in the identification of nonviolent cases. On Dataset-2, the proposed model has achieved an MCC of 92.8%, higher than CNN with 83.68%, R-CNN with 81.84%, and YOLO with 90.99%. Besides, the NPV of 94% is greater than CNN (93.18%), R-CNN (91.11%), and YOLO (93.75%) to establish that it can also accurately predict true negatives.

Figure 18. FPR and FNR analysis



The proposed YOLOv8n model performs well with outstanding performance in terms of both false positives and false negatives in the two datasets. In Dataset-1, it achieved an extremely low FPR of 0.2% and surpassed CNN at 14.89%, R-CNN at 2.13%, and YOLO at 6.38%. It further minimized unnecessary alarms in real-time. Its FNR is superior at 4% and surpassed CNN at 5.97%, R-CNN at 7.46%, and YOLO at 5.97%, thereby enabling the timely detection of violent incidents. In Dataset-2, the model still has an FPR of 0.2%, which is better than CNN, R-CNN (12.77%), and YOLO (4.26%). Its FNR of 3% is better than that of CNN (4.48%), R-CNN (5.97%), and YOLO (4.48%), which shows that it is robust in minimizing missed detections of violent activities. The confusion matrix diagram for Dataset-1 is shown in Figure 19.

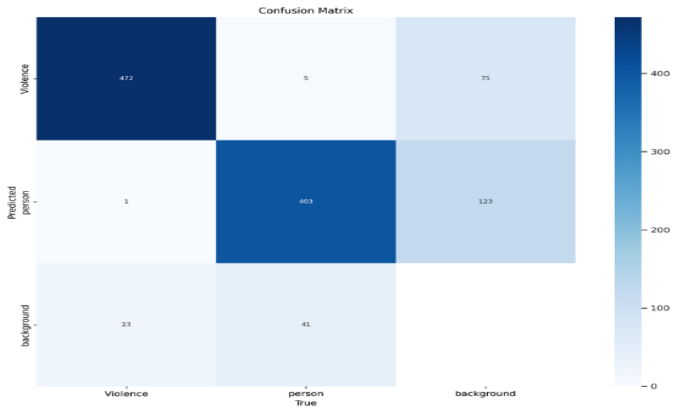


Figure 19. Confusion Matrix

The confusion matrix provides a detailed breakdown of the performance regarding the model in correctly classifying images into violence, person, and background. The matrix highlights a high number of true positives, such as 472 instances of violence, 403 instances of person, and 400 instances of background. However, there were instances when the model was classified incorrectly. False positives are present, where the model incorrectly assigns an image to the wrong category. For example, the model misclassified 75 background images as containing violence, whereas false negatives represent when the model could not correctly classify the actual class.

DISCUSSION AND CONCLUSION

Detecting violence plays a critical role in improving safety and security in smart urban environments. These findings align with recent urban studies showing complex, non-linear interactions between street-level environments and local crime patterns [32]. It reduces response time, prevents escalation, and mitigates harm in public spaces such as parks and transportation systems. Integrated systems leverage facial recognition, predictive policing, and AI-powered analytics to create a robust security framework for crime reduction. Data-driven decision-making is achieved through automated systems, which generate data on crime patterns and hotspots. These systems benefit vulnerable groups like children, the elderly, and marginalized communities by quickly identifying and protecting them during dangerous situations. Additionally, violence detection coordinates emergency services, ensuring swift assistance for victims. In this article, we employed the YOLOv8 model to detect two types of objects involved in acts of violence between two or more individuals. Results clearly show that the model has detected violence and objects with good prediction accuracy. The model achieved a high precision of 95.2% for the detection of abnormal behavior.

Advanced preprocessing techniques such as normalization, CLAHE, and data augmentation improve the model's robustness to lighting and noise variations, enhancing its real-world applicability. Results show that there is a vast improvement over the baseline methods of CNN, R-CNN, and the previous versions of YOLO, with the fine-tuned YOLOv8n model having achieved accuracies of 98.71% and 99% on Dataset-1 and Dataset-2, respectively. Such results show that transfer learning and hyperparameter optimization are quite effective for domain-specific needs. The model's ability to pick up violence-related objects like knives and guns makes it applicable to diverse scenarios in public spaces, transportation hubs, and educational institutions. The system integrates domain-specific labels such as violence and nonviolence with limited variability in the training dataset, which may limit its generalizability.

This work effectively demonstrates the use of a fine-tuned YOLOv8n model for the task of violence detection on surveillance footage. Domain-specific optimizations have enabled the model to show state-of-the-art performance in detecting violent activities and related objects while being computationally efficient enough for real-time applications. Enhanced accuracy in detection, preprocessing pipeline robustness, real-time applicability, and extensive evaluation of two datasets are some of the key contributions made by the work. This indicates that such technologies hold significant promises for public safety, reduce crime rates considerably, and even support police response effectively during emergencies. However, dataset limitations and ethical concerns are important areas that need to be addressed for the responsible deployment of such systems.

FUTURE DIRECTIONS

This work extends our previous research on deep learning-based detection and prevention of violent activities in smart cities [33]. In this research work, two datasets are used. Bias and inconsistencies were observed in the selected datasets, indicating areas for improvement in future research. Future work could explore advanced DL models, such as CNNs with attention mechanisms or transformers, to further improve real-time violence detection in CCTV footage. This improvement could enhance violence detection, helping to identify incidents at early stages and prevent major losses. Future work can also leverage these results to improve dataset diversity, incorporate synthetic data, and explore more complex hybrid architectures, such as the combination of YOLOv8n with attention mechanisms or transformers. Further optimization for edge devices, such as the NVIDIA Jetson Nano, would ensure real-time performance in resource-constrained environments. Adding multi-modal systems that incorporate audio analysis or contextual data will further enhance detection accuracy and context awareness.

REFERENCES

H. Nahata and S. P. Singh, "Deep learning solutions for skin cancer detection and diagnosis," Machine learning with health care perspective: machine learning and healthcare, pp. 159-182, 2020.

A. Esteva et al., "A guide to deep learning in healthcare," Nature medicine, vol. 25, no. 1, pp. 24-29, 2019.

G. Latif, S. E. Abdelhamid, R. E. Mallouhy, J. Alghazo, and Z. A. Kazimi, "Deep learning utilization in agriculture: Detection of rice plant diseases using an improved CNN model," Plants, vol. 11, no. 17, p. 2230, 2022.
M. Biswas, M. Ray, I. Jahan, S. Khan, S. Ahmad Saad, and P. Bharman, "Deep learning in agriculture: A review," Asian Journal of Research in Computer Science, pp. 28-47, 2022.
U. Kamath, J. Liu, and J. Whitaker, "Deep learning for NLP and speech recognition,". Springer, 2019.
G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," ACM computing surveys (CSUR), vol. 54, no. 2, pp. 1-38, 2021.
O. Doukhi, S. Hossain, and D.-J. Lee, "Real-time deep learning for moving target detection and tracking using unmanned aerial vehicle," vol. 26, no. 5, pp. 295-301, 2020.
T. Vaiyapuri, S. NandanMohanty, M. Sivaram, I. V. Pustokhina, D. A. Pustokhin, and K. Shankar, "Automatic Vehicle License Plate Recognition Using Optimal Deep Learning Model," Computers, Materials & Continua, vol. 67, no. 2, 2021.
Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE transactions on neural networks and learning systems, vol. 30, no. 11, pp. 3212-3232, 2019.
N. Jaafar and Z. Lachiri, "Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance," Expert Systems with Applications, vol. 211, p. 118523, 2023.
S. A. Sumon, R. Goni, N. B. Hashem, T. Shahria, and R. M. Rahman, "Violence detection by pretrained modules with different deep learning approaches," Vietnam Journal of Computer Science, vol. 7, no. 01, pp. 19-40, 2020.
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.
Kwon, Hyeokjun, Seongho Park, and Jae-Hwan Kim. 2024. "Understanding Sexual Homicide in Korea Using Machine Learning Algorithms." *Behavioural Sciences & the Law* 42 (4). https://doi.org/10.1002/bsl.2676.
Kim, G., Y. Cho, J.-H. Lee, and G. Lee. 2024. "Correlation Analysis between Urban Environment Features and Crime Occurrence Based on Explainable Artificial Intelligence Techniques." *Journal of Asian Architecture and Building Engineering*. https://doi.org/10.1080/13467581.2024.2421260.
V. D. Huszar, V. K. Adhikarla, I. Negyesi, and C. Krasznay, ´ "Toward fast and accurate violence detection for automated video surveillance applications," IEEE Access, vol. 11, pp. 18772-18793, 2023.
M. Magdy, M. W. Fakhr, and F. A. Maghraby, "Violence 4D: Violence detection in surveillance using 4D convolutional neural networks," IET Computer Vision, vol. 17, no. 3, pp. 282-294, 2023.
X. Zhang, S. Yang, J. Zhang, and W. Zhang, "Video anomaly detection and localization using motion-field shape description and homogeneity testing," Pattern Recognition, vol. 105, p. 107394, 2020.
Abraham, J., R. Rohde, and T. Eriksson. 2025. "Crime and Visually Perceived Safety of the Built Environment." *British Journal of Criminology*. https://doi.org/10.1080/24694452.2025.2501998
T. Hussain, A. Iqbal, B. Yang, and A. Hussain, "Real time violence detection in surveillance videos using Convolutional Neural Networks," Multimedia Tools and Applications, vol. 81, no. 26, pp. 38151-38173, 2022.
R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveillance," Sensors, vol. 22, no. 6, p. 2216, 2022.
N. Honarjoo, A. Abdari, and A. Mansouri, "Violence detection using pre-trained models," in 2021 5th International conference on pattern recognition and image analysis (IPRIA), 2021: IEEE, pp. 1-4.
M. S. Mahdi and A. J. Mohammed, "Detection of unusual activity in surveillance video scenes based on deep learning strategies," Journal of Al-Qadisiyah for computer science and mathematics, vol. 13, no. 4, pp. Page 1–9, 2021.
S. Jebur, K. Hussein, H. Hoomod, and L. Alzubaidi, "Novel Deep Feature Fusion Framework for Multi-Scenario Violence Detection. Computers 2023, 12, 175," ed, 2023.
M. M. Ali, "Real-time video anomaly detection for smart surveillance," IET Image Processing, vol. 17, no. 5, pp. 1375-1388, 2023.
J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3d convolutional neural networks," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019: IEEE, pp. 1-8.
Jia, S., & Tian, Y., "Face Detection Based on Improved Multi-task Cascaded Convolutional Neural Networks," IAENG International Journal of Computer Science, 51(2), 2024.
Esan, D., Owolawi, P. A., & Tu, C., "Anomalous detection in noisy image frames using cooperative median filtering and KNN," IAENG International Journal of Computer Science, 49(1), 1-10, 2022.
Shahrim, K. A., Abd Rahman, A. H., & Goudarzi, S., "Hazardous Human Activity Recognition in Hospital Environment Using Deep Learning," IAENG International Journal of Applied Mathematics, 52(3), 2022.

Hong, K., "Facial Expression Recognition Based on Anomaly Detection and Multispectral Imaging," IAENG International Journal of Computer Science, 51(10), 2024.

Dey, A., Biswas, S., & Abualigah, L., "Efficient violence recognition in video streams using resdlcnn-gru attention network," ECTI Transactions on Computer and Information Technology (ECTI-CIT), 18(3), 329-341, 2024.

Lee, S., H. Jang, and K. Park. 2025. "Analysing Non-Linearities and Threshold Effects between Street-Level Built Environments and Local Crime Patterns." *Urban Studies*. https://doi.org/10.1177/00420980241270948.

Baala, Asmae, Mostafa Hanoune, and Mohssine Bentaib. 2025. "Vision-Based Detection and Prevention of Violent Activities in Smart Cities Using Deep Learning." *Proceedings of the 2025 International Conference on Innovative Research in Applied Science, Engineering, and Technology (IRASET)*. IEEE. https://doi.org/10.1109/IRASET64571.2025.11008153.