IUST Iran University of Science and Technology

Iranian Journal of Electrical and Electronic Engineering

Journal Homepage: ijeee.iust.ac.ir

A Deep Learning Model for YOLOv9-based Human Abnormal Activity Detection: Violence and Non-Violence Classification

Sirajus Salehin*, Shakila Rahman** (C.A.), Mohammad Nur*, Ahmad Asif*, Mohammad Bin Harun*, and Jia Uddin** (C.A.)

Abstract: Abnormal activity detection is crucial for video surveillance and security systems, aiming to identify behaviors that deviate from normal patterns and may indicate threats or incidents such as theft, vandalism, accidents, and aggression. Timely recognition of these activities enhances public safety across various environments, including transportation hubs, public spaces, workplaces, and homes. In this study, we focus on detecting violent and non-violent activities of humans using a YOLOv9-based deep learning model considering the above issues. A diverse dataset has been built of 9,341 images from various platforms, and then the dataset has been pre-processed, i.e., augmentation, resizing, and annotating. After pre-processing, the proposed model has been trained which demonstrated strong performance, achieving an F1 score of 95% during training for 150 epochs. It was also trained for 200 epochs, but early stopping was applied at 148 epochs as there was no significant improvement in the results. Finally, the results of the YOLOv9-based model have been analyzed with other baseline models (YOLOv5, YOLOv7, YOLOv8, and YOLOv10) and it performed better compared with others.

In Press

Keywords: Abnormal activity detection, Deep Learning, YOLOv9-based model, Real-time object detection.

1 Introduction

Abnormal activity detection is one of the most important aspects of video surveillance and security systems. This task is designed to identify behavior that falls outside normal behavior patterns and may be a sign of threat or an incident. Abnormal activities are actions that are not usually observed, but for specific contexts, it can be theft, vandalism, accidents, aggressive actions, and so on. Recognizing these activities in a timely manner is very important for the safety and security of the public at various facilities: transport, public, work and home.

In recent years, the integration of sophisticated deep learning models in abnormal activity detection has become increasingly popular because of its accuracy and speed. The YOLO series has been one of the most prominent models concerning real-time object detection, and YOLOv9, the latest installment of the series, maintains high performance on the task. The model is an advanced option for analyzing a variety of video data and recognizing abnormal activities with superior quality and speed [1]. The desire to improve quality of life and safety

Iranian Journal of Electrical & Electronic Engineering, YYYY.
Paper first received DD MONTH YYYY and accepted DD MONTH YYYY.

*Department of Computer Science and Engineering, American International University, Bangladesh, Dhaka, 1229, Bangladesh
E-mails: iftymdss@gmail.com, mdnur701@gmail.com, asif141201@gmail.com, mohd.binharun788256@gmail.com

** Department of Computer Science and Engineering, American International University, Bangladesh, Dhaka, 1229, Bangladesh
E-mail: shakila.rahman@aiub.edu
** Artificial Intelligence and Big Data Department, Woosong University, Daejeon 34606, Republic of Korea
E-mail: jia.uddin@wsu.ac.kr
Corresponding Authors: Shakila Rahman, Jia Uddin.

in a variety of sectors is the driving force behind the relevance of abnormal activity detection. For example, older people who frequently live alone are more susceptible to situations that they may not recognize because of physical or mental limitations [2]. Conventional techniques for keeping an eye on these people have mostly depended on vision-based systems, which come with a few drawbacks, including complicated computations, inconsistent lighting, and privacy issues [3]. However, while they have their own set of drawbacks, other strategies including ambient sensing technology and wearable sensors also provide intriguing answers [4].

Utilizing sophisticated models to increase detection efficiency and accuracy has become a growing area of interest in latest research. For instance, research has been done on Pyroelectric Infrared (PIR) sensors because of their non-intrusive nature and ability to withstand changes in the environment, which makes them a good choice for monitoring at residence [5]. Similar to this, convolutional neural networks (CNNs) are becoming more and more popular for use in image-based applications because of how well they can capture and process visual input [6]. With its high accuracy and efficiency, the YOLO (You Only Look Once) model, especially in its most recent iterations such as YOLOv9, marks a substantial leap in real-time object recognition and categorization [7]. The goal of the proposed work is to use YOLOv9 to overcome the difficulties in identifying abnormal behavior that is both violent and non-violent. With its exceptional performance in object identification tasks, YOLOv9 is an ideal choice to address the difficulties of distinguishing different human activities from video data. This strategy differs from earlier approaches that mostly relied on outdated YOLO models or other models with lower real-time processing capacities [8].

Our study promises to improve the accuracy of abnormal activity detection while providing timely notifications by integrating YOLOv9, which is important for security and healthcare applications. Our research stands apart from previous studies by using the most recent developments in deep learning with an all-encompassing method for identifying a wide range of human behaviors. In contrast to previous research that frequently concentrates on certain activities or utilizes small datasets [9], our method makes use of cutting-edge technology and a diversified dataset to provide more reliable and broadly applicable conclusions. This thorough technique seeks to fill in gaps found in previous research, including the requirement for increased detection accuracy and adaptability to a variety of activity kinds [10].

The main contributions of this paper are as follows:

- This study focuses on human abnormal activities, such as violent and non-violent activities classification for human safety purposes in different areas.
- First, a large custom dataset including a wide range of abnormal behaviors, which strengthens the model's generalizability compared to previous studies is built.
- Second, the custom dataset is preprocessed using some image processing tools such as, data augmentation techniques to enhance the model’s performance and robustness across different environments along with resizing and labeling. Then, after preprocessing proposed YOLOv9-based object detection deep learning model is trained to accurately recognize and classify various types of activities in real-time, significantly enhancing detection speed and accuracy.
- Finally, we provide a detailed analysis of YOLOv9’s effectiveness in detecting abnormal activities, highlighting its superiority over other baseline models such as YOLOv10, YOLOv5, YOLOv7, YOLOv8 and in all cases proposed YOLOv9 performs better.

To improve model performance and generalizability even further, our strategy incorporates sophisticated data augmentation approaches. Together, these contributions raise the bar for abnormal activity identification and offer a useful tool for security and medical applications.

This paper is organized as follows: Section 2 provides a literature review, discussing previous research and existing methods in the field of abnormal activity detection. Section 3 describes the methodology, including data collection, model training, and evaluation processes. Section 4 offers a comprehensive analysis of the experimental results, comparing the performance of YOLOv9 with other models and highlighting its advantages. Finally, Section 5 concludes the paper, summarizing the key findings and contributions of the study.

2 Literature Review

Abnormal human-activity detection has become a critical area of research due to its applications in security and healthcare. The utilization of sensor-based systems, as explored by Jie Yin, Qiang Yang, and Jeffrey Junfeng Pan in [11], presents a novel two-phase approach using wireless sensors to address the challenges of biased data and infrequent abnormal events. This method significantly reduces false positives compared to traditional systems. In video-based detection, Thomas Gatt, Dylan Seychell, and Alexiei Dingli's work in [12] introduces an automated camera-based system for identifying irregular human behaviors. Their research emphasizes the potential of video-generated models to detect rare events that deviate from normal patterns.

In Press

Meanwhile, support vector machines (SVM) have been applied in telehealth and ubiquitous computing environments to recognize abnormal activities, as discussed in [13] by Adithyan Palaniappan et al., demonstrating the role of wearable sensors in health monitoring. Real-time video surveillance is another significant focus area. Xinyu Wu et al. in [14] developed a system for detecting abnormal behaviors, highlighting the importance of real-time processing in effective surveillance. Additionally, modeling group activities involving moving objects, as detailed by N. Vaswani and colleagues in [15], is crucial for advancing detection methodologies in complex environments. Bruno Degardin and Hugo Proença's research in [16] presents an iterative self-supervised learning framework to enhance detection accuracy in human activity analysis, addressing the limitations of current methods. Furthermore, the switching hidden semi-Markov model (S-HSMM) has proven effective for activity recognition, particularly in recognizing human activities with varying durations and states, as explored in [17] by T.V. Duong et al. Location-based approaches for abnormality detection, such as the one proposed by Y. Benezeth and team in [18], incorporate spatial and temporal information to improve accuracy. Efficient and robust skeleton-based methods, as discussed by Amr Elkholy et al. in [19], are vital for enhancing safety in vulnerable populations. Spatio-temporal descriptors have also been proposed to improve detection with lower computational costs, as shown in [20] by Fam Boon Lung and colleagues. Similarly, probabilistic approaches leveraging sensor data for detecting abnormal behavior in activities of daily living (ADLs) are discussed in [21] by M. Garcia-Constantino et al., emphasizing the enhancement of detection capabilities.

In deep learning, Meriem Zerkouk and Belkacem Chikhaoui's work in [22] leverages models to predict abnormal behavior in elderly persons, improving health monitoring in smart environments. Unsupervised learning approaches for anomaly detection in video sequences are explored by Hyunjong Park and colleagues in [23], emphasizing the enhancement of anomaly detection accuracy through diversity in normal patterns. Attention-based convolutional neural networks (CNNs) have been introduced for weakly labeled human activity recognition, as discussed by Kun Wang et al. in [24], focusing on improving accuracy using wearable sensors. Real-time recognition of abnormal human behaviors using surveillance cameras, as presented by Hoang Nguyen Ngoc et al. in [25], addresses challenges related to data scarcity and model complexity.

Hybrid deep-learning-based schemes for detecting suspicious flows in software-defined networks (SDNs) are discussed in [26] by Sahil Garg and colleagues, contributing to improved security in social multimedia environments. Lastly, ensemble techniques for intrusion detection in Internet of Things (IoT) networks are explored by Nour Moustafa et al. in [27], addressing the challenges posed by botnet attacks in IoT environments. These studies collectively advance the field, offering innovative solutions for safer and more secure environments.

3 Methodology

We have discussed the study's execution process in this section. Here, we are utilizing YOLOv9 as a framework for the detection and localization of abnormal and normal activity. To find and pinpoint the possible site, the architecture automatically retrieved several abnormal and normal detecting features from an input image.

The specific steps of the methodology used to identify abnormal behavior using the YOLOv9-based deep learning model are presented in the following subsections. Data collection, pre-processing, resizing, labeling, model training, and evaluation are all covered by this methodology.

3.1 Data Collection

The information mining for detecting abnormal activities must start by acquiring the data which is prior important in the development of a deep learning model. In this study, data were gathered from a collection of sources to perform a full representation of the data. These sources included social networking sites, other online video libraries, and publicly accessible web databases. The purpose in the data collection process was to include a wide variety of environments like transportation, roads, residential buildings, offices, shopping malls, sports venues, and educational institutions to cover as much environment diversity and generalization as possible.

We took the videos and converted them into a series of static images to remove the dynamic parts. That is a crucial step, as it will allow the model to see a wide range of different situations that have been recorded in the videos . The dataset also added images of normal and abnormal activities which were extracted from the internet, other than just video frames. By collecting data from many subjects following several scenarios in a variety of activities , this method covers various cases where the deliberate human is not masked by his activity. Statistics of the compiled dataset are shown in Table 1., and the sample input images are shown in Figure 1.

Table 1. Dataset Statistics

Activity Type	Quantity
Violence	6300
Non-violence	3041
Total	9341

In Press

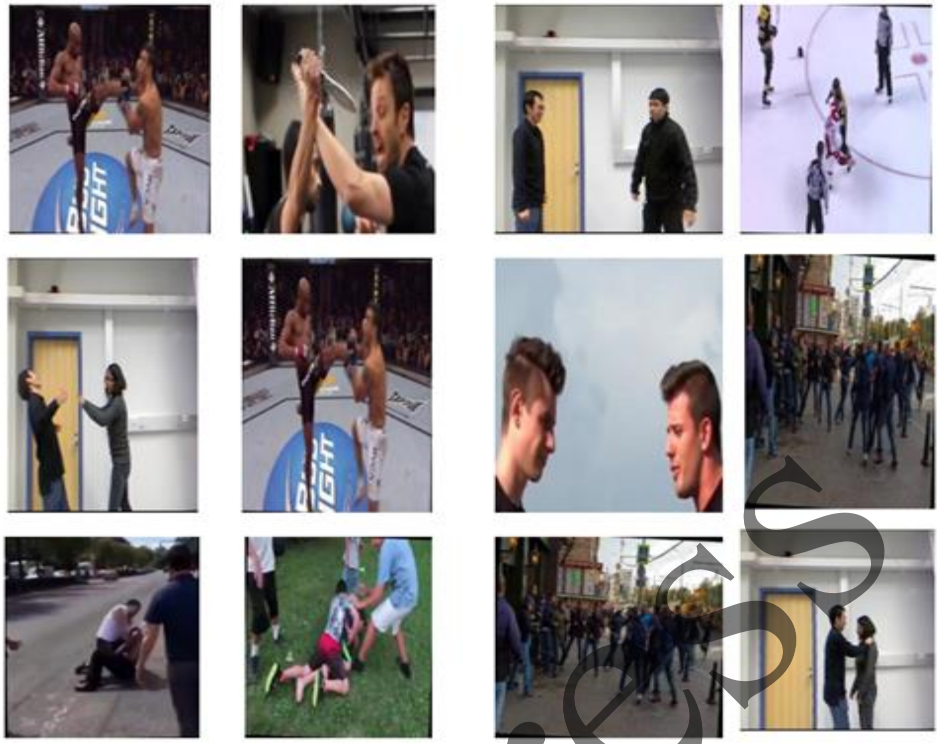


Fig. 1 Sample input data of violence and Nonviolence activities.

In Press

3.2 Data Preprocessing

Data preprocessing plays a vital role in improving the quality of the dataset and bolstering the model's resilience. The preprocessing steps incorporate data augmentation techniques to enhance the variability and diversity of the dataset, thereby strengthening the model's capacity to generalize to unfamiliar data.

Data Augmentation: One of the most prevalent methods to increase dataset artificial diversity and variability is data augmentation; this improves both machine learning (ML) and artificial intelligence (AI) models, making them more robust in performance. Since this dataset was constructed from data collected from various sources, with varied sizes and resolutions, data augmentation techniques were applied to make this dataset representative of real-world scenarios. The section that follows elaborates on the different techniques of data augmentation used in this research.

Case1- Crop and Rotational Augmentation: The data being collected from different resources; some prerequisite adjustments of crop and rotation were required to regularize the dataset. The cropping was done to get only the portions of the images relevant for classification purposes. Random rotations were done on an image to retain variation in the images as they would appear in different orientations in real life. This was done by rotating certain images at random between -20 to +20 degrees. This will make the dataset more representative of real conditions, generalizing the model better in various positions of images or performing better during inference and test periods.

Case2- Adjusting Brightness: Some images in this dataset are very dimly lit. Random brightness changes from -20% to +20% were done to help this model perform better. The method makes necessary images brighter and more visible, thus making them good for analysis and use in model training. Increased visibility means features in the images become more distinct, which is important for a model to learn effectively.

3.3 Image Resizing and Labeling

For consistent input to the YOLOv9-based model, uniform image dimensions are essential. Consequently, every image was resized within the normal 400×300-pixel size. Consistent input to the model is ensured by this scaling, which is necessary for efficient training and inference.

After the resizing process, every image was labeled with the sort of activity it represented, making a distinction between normal and abnormal activity. Labeling involved creating bounding boxes around identified actions and labeling each box with the relevant class using the

Roboflow annotation tool.

Labeling Process: After the images were resized, they needed to be labeled using defined categories. The labeling procedure involves locating and marking different things in the images. We used the Roboflow annotation tool, which is recognized for its effectiveness and simplicity, for this purpose.

The images were divided into two different classifications, including "Violence" and "Non-Violence". Bounding boxes were used to precisely annotate each image, designating the existence and placement of these objects. Each bounding box was given a class ID that corresponded to the object it represented, and they were all color-coded for simple identification. The example labeled photos in Figure 2 provide an illustration of this labeling procedure. Classes: Two primary classes were created from the images: "Violent" and "Nonviolent". Bounding Boxes: Activities within the pictures have been separated by bounding boxes, each labeled with the appropriate class.

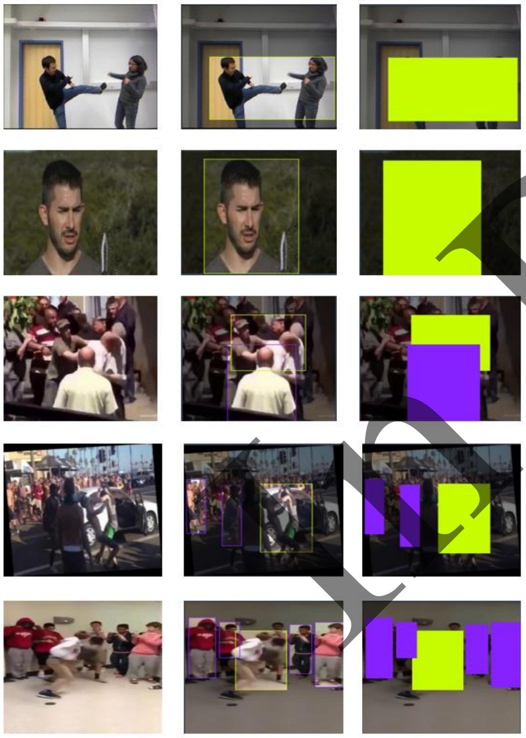


Fig. 2 Sample labeled data

3.4 Model Training

For this study's purpose, the YOLOv9-based model was used because of its excellent detection efficiency and accuracy. Several convolutional layers compose the model architecture, which is intended to efficiently identify and categorize activities in pictures.

Training Setup:

Data Split: 10% of the dataset was set aside for validation, 20% for testing, and 70% for training. With enough data left over for objective testing and validation, this split guarantees that the model gets trained on most of the available data.

Hyperparameters: Including learning rate, batch size, and number of epochs were first defined and then refined through iterative testing and validation to get optimal performance.

Training Procedure:

Initialization: Trained characteristics from a sizable dataset were used to initialize the YOLOv9-based model. By using prior knowledge, this transfer learning strategy accelerates the training process.

Validation: To maintain surveillance on the model's development and avoid overfitting, the validation dataset was used to periodically assess the model's performance during the training phase.

The workflow processing steps are shown in Figure 3. The process of the workflow, we collected the dataset, the images or videos data was annotation to label to the relevant activities. Then we resized all the data and augmented it to increase its inconsistency. To make out all the data label, trained the YOLOv9-based model using the dataset. After training the model we used to predict violent and nonviolent of unseen data. This process will be granted to ensure that the training YOLOv9-based deep learning model is recognized as the violence detection system.

In Press

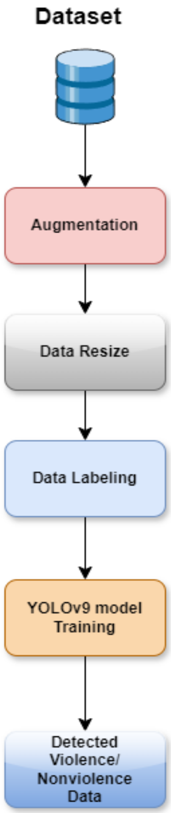


Fig. 3 Workflow diagram of the proposed system.

3.5 Model Architecture

Architecture of Yolov9-based model in Figure 4 is the modern deep learning model for real-time object detection. To increase accuracy, efficiency, and robustness this design incorporates various improvements while building the qualities of its predecessors. A comprehensive overview of the Yolov9-based model architecture, Input Layer, in this layer the model gets the images from the input data. And the consistency of the images is resized into the specific dimension. In the Backbone network, using the Convolutional Neural Network (CNN) extract the characteristics of the images. CNN is used specially to create the balance and depth of the backbone of the model. By the combination of features from many levels the PAN Layer improves the features maps produced by the FPN. The PAN improves the information flow across the network by including shortcut connects and additional convolutional layers. YOLOv9 depends on the FPN, that is built to handle multi-scale feature extraction. Detection Head is responsible for providing object detection prediction. To predicting scores, bounding boxes and class probabilities Anchor boxes are used by the YOLOv9 model.

The YOLOv9 model processes input images that represent violent or non-violent events, identifying and categorizing specific actions or objects within these images. Convolutional layers are responsible for extracting spatial features from the input image or feature maps by applying convolution operations using a set of learned filters, thereby generating feature maps. Residual blocks consist of a series of convolutional layers connected by a shortcut link that bypasses certain layers. This shortcut connection preserves information from earlier layers, helping to prevent the vanishing gradient problem and enabling the training of deeper networks. The Path Aggregation Network (PAN) layer enhances feature representation by aggregating characteristics from multiple levels of the backbone network. By combining information from different network layers, PAN helps the model capture both low- and high-level features more effectively. Similarly, the Feature Pyramid Network (FPN) is a structure that creates feature pyramids, allowing the model to recognize objects at different scales and handle variations in object sizes. FPN accomplishes this by combining features from various levels of the backbone network to generate multi-scale feature maps. Detection layers in YOLOv9 are responsible for identifying objects within the image. The model employs multiple detection layers—small, medium, and large—to accurately detect objects of varying sizes. Once objects are detected, bounding boxes are drawn around them, indicating their locations within the image. Each bounding box is accompanied by a score that reflects the model's confidence in the detection, with higher scores indicating greater certainty. Finally, the detected objects are classified into predefined categories, or classes, such as "violent" or "non-violent," based on the context of human behavior detection.

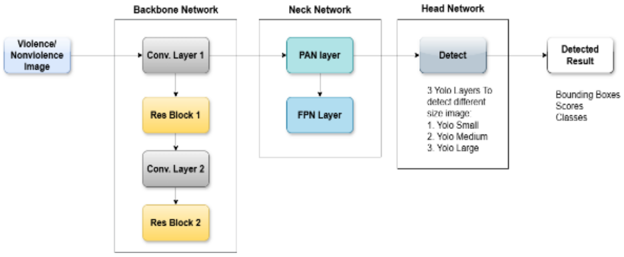


Fig. 4 Proposed Yolov9-based Architecture

4 Results

Various training settings and hyperparameters were applied during the training of the YOLOv5, YOLOv8, and YOLOv9 models for our study on detecting abnormal activities, specifically violence and non-violence. Table 2 outlines the parameters and their respective values. The models were trained with a learning rate of 0.01 and a weight decay of 0.001. We considered two primary classes (violent and non-violent) and included a null class for images without context, though this null class was not treated as a primary class.

The dataset used for training the YOLOv5, YOLOv8, and YOLOv9 models comprised approximately 9,341 images, categorized into two classes: violent action and non-violent action. These images were sourced from Google, GitHub, and Roboflow. The training process for each model involved learning to predict bounding boxes and class labels for objects within these images, adjusting the model's parameters (weights) based on the provided dataset. Training was conducted for 50, 100, 150, and 200 epochs. However, for the 200-epoch run, training automatically stopped at 148 epochs due to an early stopping mechanism that activated when no improvement was observed for a series of consecutive epochs. This early stopping step helped to prevent overfitting and optimize the training process. The batch size was consistently set to 16 for all training sessions. The size of the images used during training was standardized to 640x640 pixels.

Table 2. Parameters used to train YOLOv9-based model

Parameter	Value
Batch size	16
Number of epochs	150
Optimizer	AdamW
Pre-trained	COCO model
Learning rate	0.01
Weight decay	0.001
Patience	50

In Press

4.1 Model Evaluation

The evaluation of the trained YOLOv9-based model showed strong performance in detecting abnormal activities, particularly in distinguishing between violent and non-violent actions. The model evaluation was implemented on a Python platform with CUDA 12.0 and NVIDIA-SMI 525.85.12, utilizing a GTX 1650 Ti GPU with 4 GB of graphics memory and 16 GB of RAM. The YOLOv9-based model, with its 384-layer architecture and 25,320,790 parameters, operated at 102.3 GFLOPs, showcasing its computational efficiency.

To assess the model's effectiveness, various metrics were employed, including mean Average Precision (mAP) and class-specific precision and recall values. The YOLOv9-based model achieved an overall mAP50 of 0.602 and an mAP50-95 of 0.349. Specifically, for non-violent actions, the model demonstrated a precision of 0.63, a recall of 0.688, and an mAP50 of 0.65. For violent actions, the model achieved a precision of 0.59, a recall of 0.558, and an mAP50 of 0.554.

The evaluation process confirmed the model's capability to efficiently differentiate between the two classes of actions, with a particular strength in detecting non-violent activities. Additionally, the inference speed of the model was measured, with preprocessing taking 0.2 ms, model inference 9.6 ms, and postprocessing 2.0 ms per image. The parameters of our proposed model YOLOv9 is shown in Table 3.

Table 3. Model parameters

Parameter	Value
Model layers	384
Model parameters	2,55,32,790
Gradients	2,55,30,758
GFLOPs	102.3

4.2 Result Analysis

The performance analysis of the various YOLO models (YOLOv5, YOLOv7, YOLOv8, YOLOv9, and YOLOv10) reveals distinct strengths and weaknesses in their capabilities for obstacle detection in UAV-aided data collection systems.

YOLOv5 exhibited effective learning with steadily decreasing training and validation losses across different metrics, indicating a robust model. The precision, recall, and mean Average Precision (mAP) values consistently improved, suggesting that YOLOv5 generalizes well to unseen data. However, there were instances of confusion between nonviolent actions and the background, particularly when classifying violent actions. The overall accuracy was solid, but the model's ability to distinguish between different types of actions, especially violent actions, required further enhancement.

YOLOv7 displayed promising performance but encountered fluctuations in objectness and classification losses. These fluctuations indicate that while the model has potential, it might benefit from further tuning and optimization. The confusion matrix showed that YOLOv7 had a moderate ability to classify actions, particularly struggling with violent actions. The overall metrics, such as accuracy, precision, recall, and F1-score, were average, highlighting the need for improvement in this version.

YOLOv8 demonstrated strong performance with high accuracy, precision, and recall. The model's losses decreased consistently, and the metrics improved steadily, showing that the model learned effectively and generalized well to new data. Despite its high performance, YOLOv8 had some misclassifications, particularly in distinguishing nonviolent actions from violent ones. The confusion matrix indicated that while the model was good at identifying the majority of actions, it occasionally misclassified nonviolent actions as violent actions, suggesting a potential area for improvement.

In Figure 5, it can be seen that YOLOv9 showed good performance overall for 150 epochs, with decreasing losses and increasing precision, recall, and mAP values. However, the slightly higher validation loss hinted at potential overfitting, indicating that the model might be tailoring too closely to the training data and might not generalize as well to new data. The confusion matrix analysis revealed that YOLOv9 was effective in identifying violent actions but had a moderate rate of misclassification, particularly between nonviolent actions and the background.

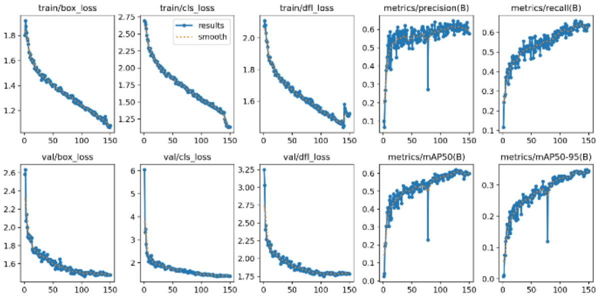


Figure 5. Training curve based on YOLOv9 for 150 epochs.

In Figure 6, it can be seen that for 150 epochs YOLOv10 showed that it is a promising model but still has considerable room for improvement. The model's losses decreased, and its evaluation metrics, such as precision and recall, increased. However, the model's training curves indicated that while it was learning effectively, it might be nearing a plateau in performance. The confusion matrix highlighted that YOLOv10 performed decently but struggled with misclassifications, particularly between nonviole

In Press

nt and violent actions.

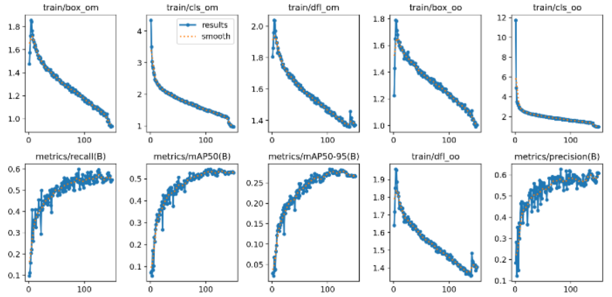


Figure 6. Training curve based on YOLOv10 for 150 epochs.

The testing performance of the proposed violent and non-violent action detection based on various YOLO models (v10, v9, v8, v7, and v5) was compared to assess their effectiveness. The performance analysis was conducted across all classes using our custom dataset, as shown in Table 4. During the training phase, the mean Average Precision at 0.5 Intersection over Union (mAP@0.5) was monitored to evaluate the model's learning capability, with higher values indicating better learning. Additionally, the F1-score, calculated using the formula,

$$F1score = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \qquad (1)$$

In all cases, the YOLOv9-based model demonstrated superior performance, with the highest mAP@0.5 and F1-score among the tested models, reflecting its robustness in detecting both violent and non-violent actions. Furthermore, the model complexities, including the number of trainable parameters for each YOLO version, were also analyzed. YOLOv7, with the highest number of trainable parameters, exhibited lower generalization capacity compared to the other models, contributing to its relatively lower performance in certain cases.

Table 4. Testing performance of YOLOv9, v10, v8, v5 and v7

Model	Epoch	Class	Trainable Parameters	F1score	mAP@0.5	mAP@0.95
Proposed YOLOv9	100	All	25.3M	0.82	0.50	0.26
Proposed YOLOv9	150	All	25.3M	0.95	0.65	0.35
YOLOv10 [33]	100	All	2.7M	0.75	0.52	0.27
YOLOv10 [33]	150	All	2.7M	0.78	0.55	0.30
YOLOv8 [32]	100	All	11.1M	0.77	0.53	0.30
YOLOv8 [32]	150	All	11.1M	0.78	0.52	0.31
YOLOv5 [31]	100	All	7.2M	0.82	0.42	0.22
YOLOv5 [31]	150	All	7.2M	0.83	0.50	0.25
YOLOv7 [30]	100	All	37.2M	0.77	0.32	0.14
YOLOv7 [30]	150	All	37.2M	0.78	0.38	0.17

4.3 Visualization

The model training was carried out over 150 epochs, with each epoch representing a full pass through the training dataset. Throughout this process, the model's parameters were updated based on calculated loss and gradients. Training concluded at 148 epochs as the most optimal results were achieved at 98 steps. Due to the patience parameter being set to 50, the training process halted after the results remained consistent for 50 consecutive steps. The entire training session took approximately 3 hours, although this duration could vary depending on the computational resources and hardware used.

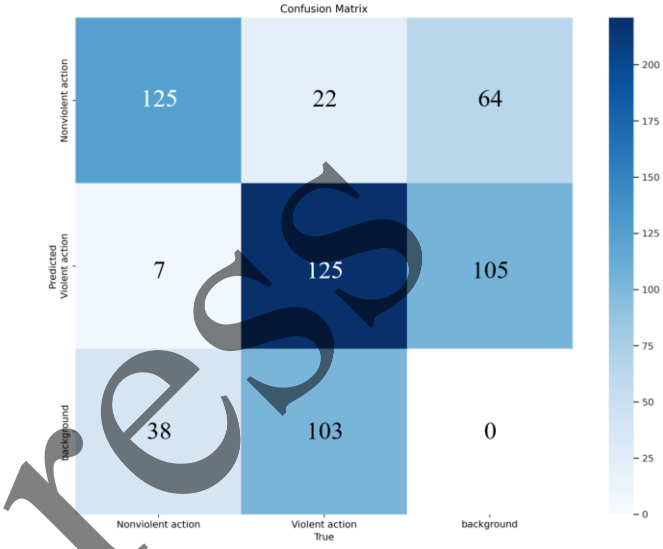


Figure 7. Confusion matrix of YOLOv9-based model for 150 epochs.

The detection metrics for violent and non-violent actions using the YOLOv9 presented in Figure 7. The confusion matrix analysis reveals that the classification model effectively identifies violent actions, with 221 true positives and only 7 false negatives, indicating strong recall. However, the model struggles with precision [28,29], as it misclassifies 64 nonviolent actions as violent (false positives). This suggests a challenge in distinguishing nonviolent actions from the background, resulting in a higher false positive rate. Despite this, the model excels in correctly identifying violent actions with minimal false negatives. For violent action detection, YOLOv9 exhibited the highest performance for violent actions with a Precision of 92%, Recall of 99%, and F1-score of 95%. Its performance for non-violent actions was also notable, with a Precision of 99%, Recall of 71%, and F1-score of 82%. YOLOv10 achieved a Precision of 64%, Recall of 98%, and F1-score of 78%. Meanwhile, for non-violent actions, it recorded a Precision of 92%, Recall of 68%, and F1-score of 78%. YOLOv8 showed strong results for violent actions, with a Precision of 98%, Recall of 64%, and F1-score of 78%, while its non-violent action

In Press

detection mirrored the performance of YOLOv10. The YOLO v5 model yielded a Precision of 72%, Recall of 99%, and F1-score of 85% for violent actions, with consistent non-violent action detection metrics across the models. Lastly, YOLOv7 demonstrated similar performance to YOLOv10 for both violent and non-violent actions, with Precision, Recall, and F1-score values of 64%, 98%, 78% for violent actions, and 92%, 68%, 78% for non-violent actions. These results highlight the effectiveness of the YOLOv9 model in violent action detection across the tested models. Based on the analysis of the models:

YOLOv9: Shows good performance with decreasing losses and increasing precision, recall, and mAP1. However, it has a slightly higher validation loss, indicating potential overfitting2.

YOLOv5: Demonstrates effective learning with decreasing losses and increasing precision, recall, and mAP. It generalizes well to unseen data4.

YOLOv7: Shows promise but has fluctuations in losses, indicating room for improvement.

YOLOv8: Performs well with high accuracy, precision, and recall [34, 35], but struggles with some misclassifications.

YOLOv10: It’s a new model so it doesn't outstand other models that dramatically but despite the promises it shows there is still a lot of room to improve [36].

YOLOv9 performs well but has some overfitting issues. YOLOv5 and YOLOv8 show strong performance with better generalization. Therefore, YOLOv9 is not necessarily the best but overall, it is a better option compared to others. Some of the detected images for violent and non-violent actions using the YOLO models (v10, v9, v8, v7, and v5) are presented in Figure 8:

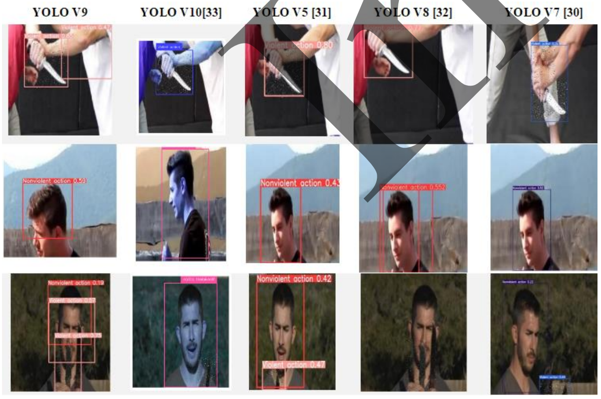


Figure 8. Testing performance for random data

5 Conclusion

In this study, a YOLOv9-based abnormal activity detection system is proposed for real-time surveillance and security applications. This work focused on detecting violent and non-violent activities using a dataset of approximately 9341 images. The system processes these images through various image processing tools, including data augmentation, resizing, and labeling. The YOLOv9 model was trained to identify and classify abnormal activities, with training conducted across 50, 100, 150, and 200 epochs, stopping automatically at 148 epochs due to early stopping mechanisms. The best results were achieved with an mAP50 of 0.65, an mAP50-95 of 0.35, a precision value of 99% for violent actions, 92% for non-violent-action and an F1 score of 95% and 92% accordingly. The evaluation results indicate that YOLOv9 performs better than YOLOv7, YOLOv8, YOLOv5, and other traditional models, particularly in distinguishing between violent and non-violent actions.

While YOLOv9 demonstrated strong performance, slight overfitting was observed, and the model faced some challenges in distinguishing non-violent actions from the background. Additionally, the study highlighted that YOLOv9 showed improved generalization and reliability compared to YOLOv7 and YOLOv10, but further optimization is required to enhance its performance and reduce misclassification rates.

Seven types of abnormal activities were considered in this work. In a real-world scenario, abnormal activities can vary greatly, which presents a limitation of this study. Generally, this project aims to enhance real-time surveillance by accurately detecting and responding to abnormal activities. Future work will focus on expanding the dataset to include more diverse and challenging scenarios and integrating the YOLOv9 model with other sensory data, such as audio and biometric information. Additionally, we plan to explore machine learning-based approaches, such as reinforcement learning, to improve the system's adaptability and response to a wider range of abnormal activities.

In Press

Acknowledgment

This research is funded by Woosong University Academic Research 2024.

Conflict of Interest

The author declares no conflicts of interest.

References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” In the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788.

[2] X. Luo, H. Tan, Q. Guan, T. Liu, H. H. Zhuo, and B. Shen, “Abnormal Activity Detection Using

Pyroelectric Infrared Sensors,” Sensors, vol. 16, no. 6, pp. 822, Jun. 2016.

[3] C. Bahhar, A. Ksibi, M. Ayadi, M. M. Jamjoom, Z. Ullah, B. O. Soufiene, and H. Sakli, “Wildfire and Smoke Detection Using Staged YOLO Model and Ensemble CNN,” Electronics, vol. 12, no. 1, pp. 228, Jan. 2023.

[4] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, “YOLO-Based Deep Learning Model for Pressure Ulcer Detection and Classification,” Healthcare, vol. 11, no. 9, pp. 1222, Apr. 2023.

[5] R. Vrskova, R. Hudec, P. Kamencay, and P. Sykora, “A New Approach for Abnormal Human Activities Recognition Based on ConvLSTM Architecture,” Sensors, vol. 22, no. 8, pp. 2946, Apr. 2022.

[6] A. Mehmood, “Abnormal Behavior Detection in Uncrowded Videos with Two-Stream 3D Convolutional Neural Networks,” Applied Sciences, vol. 11, no. 8, pp. 3523, Apr. 2021.

[7] S. Rahman, J. H. Rony, J. Uddin, and M. A. Samad, “Real-Time Obstacle Detection with YOLOv8 in a WSN Using UAV Aerial Photography,” Journal of Imaging, vol. 9, no. 10, pp. 216, Oct. 2023.

[8] H.-T. Duong, V.-T. Le, and V. T. Hoang, “Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey,” Sensors, vol. 23, no. 11, pp. 5024, May 2023.

[9] M.-t. Fang, Z.-j. Chen, K. Przystupa, T. Li, M. Majka, and O. Kochan, “Examination of Abnormal Behavior Detection Based on Improved YOLOv3,” Electronics, vol. 10, no. 2, pp. 197, Jan. 2021.

[10] S. Liu, X. Wang, H. Ji, L. Wang, and Z. Hou, “A Novel Driven Abnormal Behavior Recognition and Analysis Strategy and Its Application in a Particular Vehicle,” Symmetry, vol. 14, no. 10, pp. 1956, Sep. 2022.

[11] J. Yin, Q. Yang, and J. J. Pan, “Sensor-Based Abnormal Human-Activity Detection,” IEEE transactions on knowledge and data engineering, vol. 20, no. 8, pp. 1082-1090, 2008.

[12] T. Gatt, D. Seychell, and A. Dingli, “Detecting Human Abnormal Behaviour Through a Video Generated Model,” In Proceeding of the 11th International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 264-270. IEEE, 2019.

[13] A. Palaniappan, R. Bhargavi, and V. Vaidehi, "Abnormal Human Activity Recognition Using SVM Based Approach," In Proceeding of the 2012 international conference on recent trends in information technology, pp. 97-102. IEEE, 2012.

[14] X. Wu, Y. Ou, H. Qian, and Y. Xu, "A Detection System for Human Abnormal Behavior," In Proceeding of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1204-1208. IEEE, 2005.

[15] N. Vaswani, A. K. Roy-Chowdhury, and R. Chellappa, "Shape Activity: A Continuous-State HMM for Moving/Deforming Shapes with Application to Abnormal Activity Detection," IEEE Transactions on Image Processing 14, no. 10, pp. 1603-1616, 2005.

[16] B. Degardin and H. Proença, "Human Activity Analysis: Iterative Weak/Self-Supervised Learning Frameworks for Detecting Abnormal Events," In Proceeding of the 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1-7. IEEE, 2020.

[17] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model," In Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 838-845. IEEE, 2005.

[18] Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal Events Detection Based on Spatio-Temporal Co-Occurrences," In Proceeding of the IEEE conference on computer vision and pattern recognition, pp. 2458-2465. IEEE, 2009.

[19] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen, and E. Saba, "Efficient and Robust Skeleton-Based Quality Assessment and Abnormality Detection in Human Action Performance," IEEE journal of biomedical and health informatics 24, no. 1, pp. 280-291, 2019.

[20] F. B. Lung, M. H. Jaward, and J. Parkkinen, "Spatio-Temporal Descriptor for Abnormal Human Activity Detection," In Proceeding of the 14th IAPR International Conference on Machine Vision Applications (MVA), pp. 471-474. IEEE, 2015.

[21] M. Garcia-Constantino, A. Konios, I. Ekerete, S.-R. G. Christopoulos, C. Shewell, and C. Nuge, "Probabilistic Analysis of Abnormal Behaviour Detection in Activities of Daily Living," In Proceeding of the IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 461-466. IEEE, 2019.

[22] M. Zerkouk and B. Chikhaoui, "Spatio-Temporal Abnormal Behavior Prediction in Elderly Persons Using Deep Learning Models," Sensors, vol. 20, no. 8, pp. 2359, 2020.

[23] H. Park, J. Noh, and B. Ham, "Learning Memory-Guided Normality for Anomaly Detection," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14372-

In Press

14381. 2020.

[24] K. Wang, J. He, and L. Zhang, "Attention-Based Convolutional Neural Network for Weakly Labeled Human Activities' Recognition with Wearable Sensors," IEEE Sensors Journal, vol. 19, no. 17, pp. 7598-7604, 2019.

[25] H. N. Ngoc, N. X. Nhat, T. H. Bui, D. H. Hung, and S. Q. H. Truo, "An Efficient Approach for Real-Time Abnormal Human Behavior Recognition on Surveillance Cameras," In Proceeding of the IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1-6. IEEE, 2023.

[26] S. Garg, K. Kaur, N. Kumar, and J. J. P. C. Rodrigues, "Hybrid Deep-Learning-Based Anomaly Detection Scheme for Suspicious Flow Detection in SDN: A Social Multimedia Perspective," IEEE, 2019. IEEE Transactions on Multimedia 21, no. 3, pp. 566-578, 2019.

[27] N. Moustafa, B. Turnbull, and K.-K. R. Choo, "An Ensemble Intrusion Detection Technique Based on Proposed Statistical Flow Features for Protecting Network Traffic of Internet of Things," IEEE Internet of Things Journal 6, no. 3, pp. 4815-4830, 2018.

[28] R. Siddiqua, S. Rahman, and J. Uddin, "A Deep Learning-based Dengue Mosquito Detection Method Using Faster R-CNN and Image Processing Techniques," Annals of Emerging Technologies in Computing, vol. 5, no. 3, pp. 2021, Jul. 2021.

[29] S. B. Hasan, S. Rahman, M. Khaliluzzaman, and S. Ahmed, "Smoke Detection from Different Environmental Conditions Using Faster R-CNN Approach Based on Deep Neural Network," In Cyber Security and Computer Science, Springer, Cham, pp. 705-717, 2020.

[30] A. A. S. Chan, M. F. L. Abdullah, S. M. Mustam, F. A. Poad, and A. Joret, "Face Detection with YOLOv7: A Comparative Study of YOLO-Based Face Detection Models," In the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miri Sarawak, Malaysia, Jan. 17-19, 2024. IEEE Xplore, Mar. 26, 2024.

[31] J. H. Kim, N. Kim, Y. W. Park, and C. S. Won, "Object Detection and Classification Based on YOLO-V5 with Improved Maritime Dataset," Journal of Marine Science and Engineering, vol. 10, no. 3, pp. 377, Mar. 2022.

[32] M. J. A. Daasan and M. H. I. B. Ishak, "Enhancing Face Recognition Accuracy through Integration of YOLO v8 and Deep Learning: A Custom Recognition Model Approach," In the Proceedings of the Conference, pp. 242-253, Oct. 2023.

[33] A. S. Geetha, M. A. R. Alif, M. Hussain, and P. Allen, "Comparative Analysis of YOLOv8 and YOLOv10 in Vehicle Detection: Performance Metrics and Model Efficacy," in Recent Developments in Intelligent Transportation Systems (ITSs), 2nd Edition, Vehicles, vol. 6, no. 3, pp. 1364-1382, Aug. 2024.

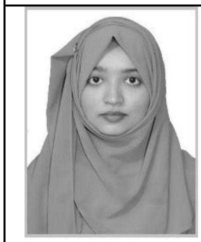
[34] R. Shakila, S. M. H. Jamee, J. K. R., J. S. Juthi, S. A. Aziz, and J. Uddin. "Real-time smoke and fire detection using you only look once v8-based advanced computer vision and deep learning." Int J Adv Appl Sci, vol. 13, no. 4, pp. 987-999, Dec. 2024.

[35] S., Shatabdi, S. Sharmin, M. D. Hasan, S. Rahman, and J. Uddin. "Congenital Anomalies Detection in Medical Images using a YOLOv8-based deep learning model," International Journal of Computing and Digital Systems, vol. 17, no. 1, pp. 1-15, Oct 2024.

[36] M.N. Alam, I. Hasnine, E.H. Bahadur, A.K.M. Masum, Jia Uddin. "DiabSense: Early Diagnosis of Non-Insulin-Dependent Diabetes Mellitus Using Smartphone-Based Human Activity Recognition and Diabetic Retinopathy Analysis with Graph Neural Network", Journal of BigData, vol. 11, no. 1, pp. 103, July 2024.

In Press

Biographies of Authors:

	Sirajus Salehin has a consistent academic record and is a recent graduate student at American International University-Bangladesh (AIUB) studying computer science and engineering. Deep learning and machine learning are two of his research areas. He can be contacted by email: iftymdss@gmail.com.
	**Shakila Rahman** is a lecturer in the Department of Computer Science and Engineering at American International University-Bangladesh (AIUB). She received an M.Sc. in AI & Computer Engineering from the University of Ulsan, South Korea, and completed a B.Sc. in Computer Science & Engineering from IIUC, Bangladesh. Her research interests include machine learning, artificial intelligence, image processing, optimization algorithms, and wireless sensor networks. She can be contacted by email: shakila.rahman@aiub.edu.
	Mohammad Nur has a consistent academic record and is a recent graduate student at American International University-Bangladesh (AIUB) studying computer science and engineering. Deep learning and machine learning are two of his research areas. He can be contacted by email: mdnur701@gmial.com.
	Ahmad Asif has a consistent academic record and is a recent graduate student at American International University-Bangladesh (AIUB) studying computer science and engineering. Deep learning and machine learning are two of his research areas. He can be contacted by email: asif141201@gmail.com.
	Mohammad Bin Harun has a consistent academic record and is a recent graduate student at American International University-Bangladesh (AIUB) studying computer science and engineering. Deep learning and machine learning are two of his research areas. He can be contacted by email: mohd.binharun788256@gmail.com.
	Jia Uddin received Ph.D. in Computer Engineering from the University of Ulsan, Korea, in January 2015. He is an Assistant Professor in AI and Big Data Department, Endicott College, Woosong University, South Korea and was an Associate Professor in Computer Science and Engineering Department at BracU, Bangladesh. His research interests include fault diagnosis, computer vision, and multimedia signal processing. He can be contacted by email: jia.uddin@wsu.ac.kr.

In Press