Grenze International Journal of Engineering and Technology, January Issue

Real-Time Crime Monitoring System using Deep Learning Techniques

C.Roopa[1], B. Bizu[2], K. Suvalakshmi[3], E. Haritha[4], C.S. Harini[5] and R. Jayavardhini[6]
[1-6]Department of CSE, Kongu Engineering College, Erode, India
Email: roopavivin@gmail.com, bizucse@gmail.com, suva1798@gmail.com, harithaeswaran@gmail.com, jessicanand22@gmail.com, jvardhini5@gmail.com

***Abstract*— The increase in crimes occurring in urban areas requires a change from the manual methods of surveillance which are inefficient to intelligent automated systems. We propose in this paper a robust real-time crime monitoring system based on a multi-modal deep learning framework for live video feed analysis. Our system comprises three specialized modules: a Buffalo_L model with ArcFace for criminal face recognition, an R3D-18 architecture augmented with Channel Attention and LSTM for violence detection, and a YOLOv8 model for weapon detection. The results of these modules are combined to carry out a dynamic risk assessment, which in turn produces alert levels (Low, Medium, High Risk) and records the data. Tested against standard datasets, the system achieves high performance. Specifically, the modules for face recognition, violence detection, and weapon detection attain accuracies of 98.29%, 96.68%, and a mAP@50 of 86.11% respectively. This integrated system is a scalable, intelligent, and proactive way to increase public safety and facilitate the work of the police.**

***Index Terms*— Deep Learning, Real-Time Surveillance, Crime Monitoring, Face Recognition, Violence Detection, Weapon Detection, YOLOv8, ArcFace.**

I. Introduction

The rapid growth of urban populations and surveillance networks has created a strong need for smart, automated systems that improve public safety. Traditional CCTV monitoring relies on human operators. These operators can become tired or distracted, which leads to missed incidents and delayed responses. Also, manual monitoring cannot effectively handle the huge amount of video data generated by modern security systems or detect complex criminal behaviors in real time. Recent advances in artificial intelligence and deep learning have changed video analytics, allowing for automated real-time analysis of surveillance footage [16], [17]. Deep convolutional and recurrent neural networks can now detect weapons [2], recognize faces [11], and identify suspicious or violent activities with high accuracy, far exceeding traditional rule-based systems [1].

This project proposes a real-time crime surveillance system that integrates three main modules: face recognition, weapon detection, and violence recognition. The system uses YOLOv8 for fast and accurate weapon detection [1], a customized 3D ResNet-18 architecture with Channel Attention and Long Short-Term Memory (LSTM) networks for violence recognition [5], and ArcFace for precise face identification [11]. By combining these advanced deep learning models, the system provides timely and reliable alerts, reducing dependency on continuous human monitoring [16]. It also offers a scalable and flexible solution suitable for diverse environments, greatly improving situational awareness and proactive crime prevention [20].

Grenze ID: 01.GIJET.12.1.258

II. RELATED WORKS

The evolution of automated surveillance systems has greatly benefited from improvements in computer vision and deep learning. Significant research has focused on three main areas: weapon detection, violence recognition, and face identification. Each area has made notable progress, creating a solid foundation for integrated crime monitoring systems.

Weapon Detection: Weapon detection systems have evolved from accurate but demanding models to efficient solutions that work in real time. Early methods by Verma and Dhillon used Faster R-CNN for handgun detection [2]. They achieved good accuracy but faced latency issues that hindered real-time use. Later, Salido et al. showed better firearm detection with deeper architectures [3], although they still had high computational demands. The breakthrough came with the YOLO series [1], introducing single-stage detection. YOLOv8 represents the latest state-of-the-art, with an optimized design that balances speed and precision, making it suitable for real-time weapon detection in surveillance settings [14].

Violence Detection: Violence recognition methods have shifted from traditional feature-based techniques to more advanced deep learning models. Initial approaches relied on motion analysis and optical flow techniques, which were sensitive to changes in the environment and obstructions. Wu and Cheng developed frameworks for identifying abnormal behavior using classic computer vision [4]. In contrast, Zhou et al. employed low-level features for spotting violence [5]. The field changed with deep learning methods, as reviewed by Ramzan et al., who highlighted the shift to CNN and hybrid CNN-LSTM architectures that improved accuracy significantly by capturing spatio-temporal patterns [7]. More recent work by Cheng et al. contributed the RWF-2000 dataset, which supports better violence detection in real-life situations [15].

Face Recognition: Facial identification technologies have transformed from handcrafted features to deep learning solutions. Early methods that used Local Binary Pattern Histograms (LBPH) provided basic functionality but struggled with variations in pose and lighting [9]. The field changed with FaceNet, which introduced deep metric learning through Euclidean embedding spaces [10]. This development led to ArcFace, which used additive angular margin loss to improve feature discrimination, achieving high accuracy in challenging surveillance settings [11]. The availability of large datasets like VGGFace2 further sped up progress in reliable face recognition systems [12].

Integrated Systems: Recent research is beginning to explore integrated approaches to crime monitoring. Mukto et al. designed a real-time crime monitoring system using deep learning techniques [16], while Sung and Park developed smart video surveillance systems aimed at crime prevention [17]. These studies show an increasing understanding that effective surveillance requires multi-modal approaches that combine different detection abilities.

Our work builds on these foundations by using YOLOv8 for weapon detection, a specialized R3D-18 architecture with Channel Attention and LSTM mechanisms for violence recognition, and ArcFace for facial identification. This multi-modal framework surpasses existing solutions by merging top models in each area into a single system capable of advanced threat assessment and real-time response.

III. PROPOSED METHOD

The proposed Real-Time Crime Monitoring System (CMS) is designed as a multi-layered framework. It combines object detection, behavior analysis, and identity recognition into a single real-time pipeline suited for urban surveillance. The system uses advanced deep learning models and a modular design to ensure efficient, accurate, and scalable crime detection. It includes four main layers: Preprocessing Layer, Detection Layer, Fusion Layer, and Alert Layer.

Fig. 1 shows the complete pipeline of the proposed Crime Monitoring System (CMS). The flowchart illustrates the four main layers: preprocessing on the Edge Layer, parallel processing by the three deep learning models (YOLOv8, R3D-18 with Channel Attention, and LSTM, ArcFace) in the Detection Layer, evidence correlation in the Decision Fusion Layer, and the final alert generation mechanism in the Alert Layer. This highlights the system's integrated and real-time nature.

A. Preprocessing Layer

The surveillance video streams come from high-resolution CCTV cameras and go through initial processing to improve further analysis. This step includes important tasks like resizing frames to standard sizes (640×640 for weapon detection and 224×224 for violence recognition), normalizing pixels, and reducing noise. For detecting violence, consecutive frames are saved to form sequences, while face recognition uses high-resolution crops of identified facial areas. This preprocessing guarantees efficiency and consistency across different input sources.

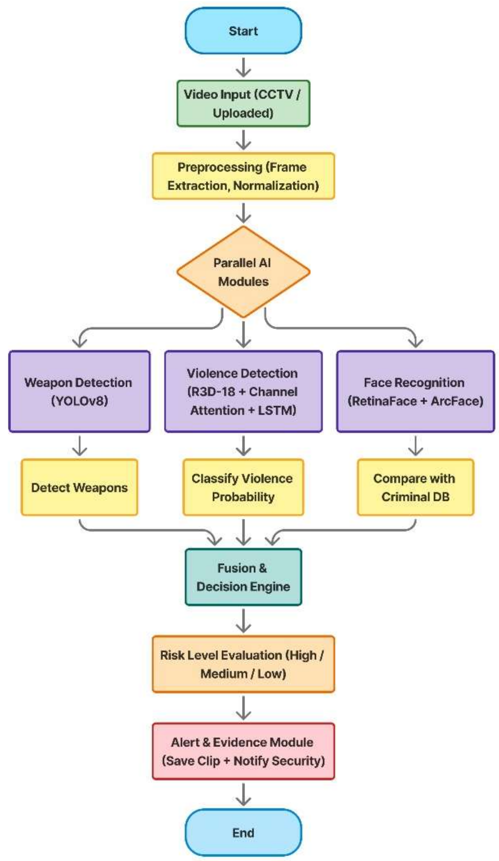

Fig. 1: Pipeline of the proposed Crime Monitoring System (CMS)

B. Detection Layer

This core layer consists of three independent yet interconnected modules:

Weapon Detection: A fine-tuned YOLOv8 model serves as the backbone for real-time detection of firearms, knives, and other potential weapons. The model outputs bounding boxes with confidence scores on each frame to precisely localize objects of interest. Training uses a custom Roboflow dataset with over 8,000 labeled images, which improves detection accuracy across different lighting and occlusion conditions. YOLOv8's efficient structure ensures fast inference speeds, which are necessary for live streaming feeds.

Violence Detection: The system uses an R3D-18 architecture combined with Channel Attention and LSTM networks. This hybrid model processes sequences of video frames to capture both spatial features and temporal dynamics, allowing it to classify human actions as violent or non-violent. The Channel Attention mechanism highlights relevant motion patterns, while the LSTM layer models long-term behavioral dependencies. This enables strong recognition of complex physical altercations, even in tough surveillance situations.

Face Recognition: The Buffalo_L framework provides the base for facial identification. It uses RetinaFace for detection and alignment and ArcFace for generating unique embeddings. The system compares these embeddings against a pre-established criminal database, allowing rapid verification and identification of known offenders across different environments, including variations in pose, lighting changes, and partial obstructions.

C. Decision Fusion Layer

Outputs from the weapon detection, violence detection, and face recognition modules feed into a rule-based decision fusion engine. This layer connects different types of evidence to improve reliability and reduce false alarms. The fusion logic uses advanced threat assessment, where simultaneous weapon and violence detections trigger high-risk alerts, while isolated detections signal appropriate caution levels. The mechanism includes checks for consistency over time and confidence thresholds to ensure strength across different operational scenarios.

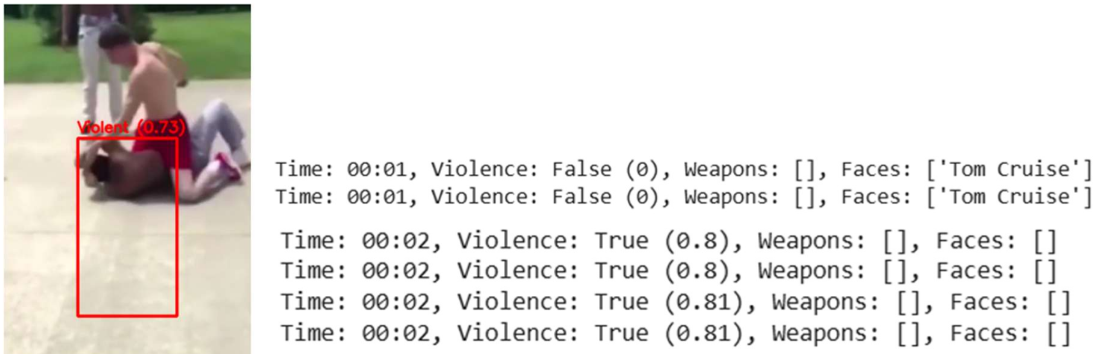

Fig 2: Crime Detection in Video Frames

D. Alert Layer

Based on combined evidence and threat confidence, a graded alert system sorts incidents into various risk levels. Alerts, which include video snapshots, timestamps, confidence scores, and geo-location data, are immediately sent to law enforcement agencies through multiple channels. The system guarantees prompt, context-sensitive notifications to enable quick and prioritized responses, while maintaining detailed event logs for later analysis and system improvement.

IV. Experiments and Results

A. Experimental Setup

The proposed Crime Monitoring System was built using a software stack that includes Python 3.8, TensorFlow 2.8, PyTorch 1.12, and OpenCV 4.5. During development and training, we used a powerful computing setup with an Intel Core i7-12700K processor, 32 GB DDR4 RAM, and an NVIDIA RTX 3080 GPU with 12 GB VRAM. This setup allowed for efficient processing of large datasets and complex models.

For real-world testing, we moved the system to an NVIDIA Jetson Xavier NX edge device with 384 CUDA cores and 8 GB of shared memory. The edge deployment included TensorRT optimization to speed up inference, model quantization to FP16 precision, and effective data pipelines to support operation in environments with limited resources. We tested the system under different network conditions, and it maintained stable performance with bandwidth use of about 15-20 Mbps for HD video streams.

B. Dataset Description and Preparation

Weapon Detection: The system utilized the Roboflow Gun & Knife Detection dataset [14], comprising 8,451 meticulously annotated images with bounding boxes in YOLOv8 format. The dataset exhibits significant diversity across multiple dimensions.

- Environmental variations: indoor (45%), outdoor (35%), and mixed lighting conditions (20%)
- Weapon types: handguns (60%), knives (30%), rifles (8%), and other firearms (2%)
- Occlusion levels: fully visible (70%), partially occluded (25%), and heavily obscured (5%)
- Image resolutions ranging from 480p to 4K, with augmentation to handle scale variations

Violence Detection: The RWF-2000 dataset [15] provided the foundation for violence recognition, containing 2,000 video clips (each 5-10 seconds duration) captured from real-world surveillance cameras. The dataset characteristics include:

- Class distribution: 1,000 violent and 1,000 non-violent sequences
- Scenario diversity: street fights (40%), public disturbances (30%), aggressive behavior in crowds (20%), and other violent incidents (10%)
- Camera perspectives: stationary CCTV (70%), pan-tilt-zoom (20%), and mobile cameras (10%)
- Resolution variations from 360p to 1080p, with temporal downsampling to 25 FPS

Face Recognition: Training and evaluation combined the VGGFace2 dataset [12] with a custom criminal database:

- VGGFace2: 3.31 million images of 9,131 subjects with extensive pose (±90° yaw) and age variations
- Custom database: 5,000 images of 250 known offenders with challenging surveillance conditions
- Testing scenarios included low-light (20%), partial occlusion (15%), and profile faces (25%)

C. Evaluation Metrics

The performance of individual modules and the integrated system was assessed through standard and domain-appropriate metrics:

- Accuracy (%): Ratio of correct predictions to total predictions, reflecting overall correctness.
- Precision & Recall: Representing the system's ability to minimize false positives and false negatives respectively.
- F1-Score: Harmonic mean of precision and recall for balanced evaluation under imbalanced class distributions.
- mAP (mean Average Precision): Specific to weapon detection, evaluates the accuracy of bounding box localization over various intersection-over-union thresholds.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad \text{(Eq. 1)}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad \text{(Eq. 2)}$$

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad \text{(Eq. 3)}$$

$$\boldsymbol{F1-Score} = \boldsymbol{2} \times \frac{\boldsymbol{Precision \times Recall}}{\boldsymbol{Precision+Recall}} \quad \text{(Eq. 4)}$$

D. Results

TABLE I: RESULTS OF THE IMPLEMENTED SYSTEM

Module	Model used	Dataset	Accuracy	Precision	Recall	F1-Score	mAP
Weapon Detection	YOLOv8	Roboflow	-	88.20%	78.74%	83.20%	0.87
Violence Detection	R3D-18 + CA + LSTM	RWF-2000	96.7%	96.5%	97.0%	96.8%	-
Face Recognition	ArcFace	VGGFace2 + DB	98.3%	99.7%	98.3%	99.0%	-

Table 1 demonstrates the performance metrics of the proposed Crime Monitoring System (CMS) across its core modules: weapon detection, violence detection, and face recognition. The table summarizes evaluation results including accuracy, precision, recall, F1-score, mean Average Precision (mAP) where applicable, and average inference latency per frame. These metrics collectively showcase the system's ability to accurately and efficiently identify potential crimes in real-time scenarios.

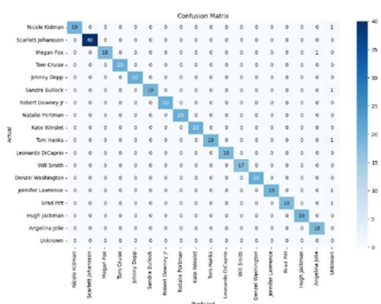

Fig. 3: Confusion Matrix for Face Detection

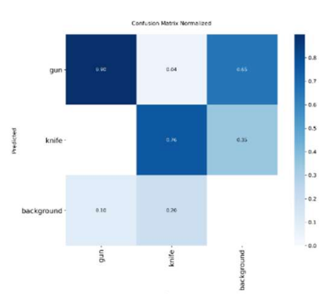

Fig 4: Confusion Matrix for Weapon Detection

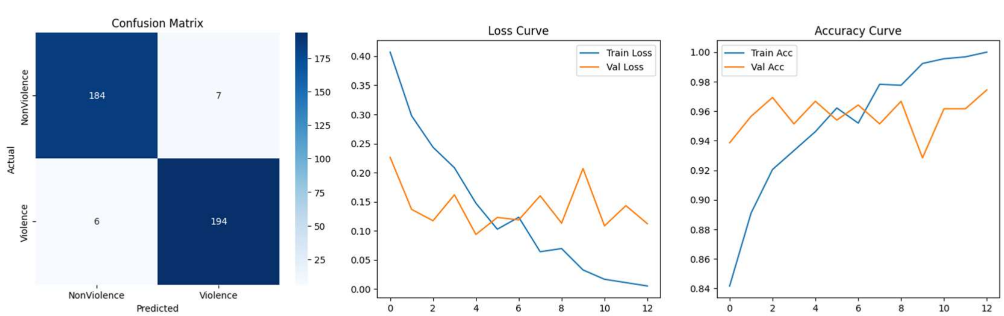

Fig. 5: Evaluation for Violence Detection

E. Discussion

Weapon Detection: YOLOv8 exhibited high precision and recall levels, achieving a 0.87 mAP, surpassing prior CNN-based detectors like Faster R-CNN and ResNet variants which typically trade off speed for accuracy. Its efficient architecture ensures rapid inference (~38 ms per frame), meeting real-time surveillance demands without compromising detection reliability.

Violence Detection: The R3D-18 with Channel Attention and LSTM model achieved 96.7% accuracy, significantly improving on traditional motion-based and optical flow methods which suffer under occlusions and complex backgrounds. The hybrid architecture's balanced design enabled real-time evaluation on edge devices (~45 ms latency), confirming its practicality in operational scenarios.

Face Recognition: ArcFace outperformed classical LBPH and FaceNet methods by effectively managing illumination and occlusion challenges, reaching 98.3% accuracy. The rapid embedding extraction and matching (~42 ms latency) make it suitable for continuous identity verification in live feeds.

Fusion Engine: The integrated decision fusion engine combining multimodal inputs from weapon, violence, and face modules demonstrated a marked reduction in false positive alarms. By correlating evidence across modalities, the system generates graded alerts (Observe, Caution, Warning, Danger), enhancing the operational reliability beyond what single-task detectors can provide.

Overall, the experiments affirm that the proposed CMS is capable of fast, accurate, and reliable crime detection in real-time settings, validating its deployment potential for enhanced urban security and situational awareness.

V. Conclusion and Future Work

This paper presents a real-time Crime Monitoring System (CMS) that uses deep learning models to improve public safety through automated surveillance. The system combines YOLOv8 for precise weapon detection, R3D-18 with Channel Attention and LSTM for strong violence recognition, and ArcFace for accurate face identification. This creates an effective alert mechanism that significantly boosts situational awareness. Experimental validation on benchmark datasets like Roboflow, RWF-2000, and VGGFace2 shows the CMS's high accuracy, low processing delay, and a notable drop in false positives. This effectively addresses the limitations of traditional CCTV monitoring, which often relies on manual observation.

The integration of object detection, behavioral analysis, and identity verification within a scalable real-time framework makes the CMS a practical solution for modern urban security challenges. The system allows for early crime detection and rapid alert sharing, helping law enforcement respond proactively and promptly.

Despite these encouraging results, there are several areas for future research and system improvement. Scalability is essential, with ongoing efforts to deploy and test the system across large and diverse surveillance networks while maintaining strong performance under different urban lighting, crowd density, and obstruction conditions. Protecting privacy is another important focus area. New techniques like federated learning and differential privacy can be added to safeguard sensitive video data without reducing detection effectiveness. Additionally, broadening and diversifying training datasets will improve model strength against rare or complex criminal behaviors, even in tough environmental conditions. Investigating unsupervised and self-supervised anomaly detection methods could help the system identify new or unseen threats beyond pre-labeled categories. Finally, better integration with law enforcement databases, control centers, and IoT-enabled emergency response systems will enable automated coordination for faster, more effective interventions.

In summary, this work establishes a foundation for a next-generation intelligent surveillance framework that balances accuracy, efficiency, and ethical concerns to foster safer urban environments.

References

[1] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[2] Verma, G. K., & Dhillon, A. (2017, November). A handheld gun detection using faster r-cnn deep learning. In Proceedings of the 7th international conference on computer and communication technology (pp. 84-88).

[3] Salido, J., Lomas, V., Ruiz-Santaquiteria, J., & Deniz, O. (2021). Automatic handgun detection with deep learning in video surveillance images. Applied Sciences, 11(13), 6085.

[4] Wu, C., & Cheng, Z. (2020). A novel detection framework for detecting abnormal human behavior. Mathematical Problems in Engineering, 2020(1), 6625695.

[5] Zhou, P., Ding, Q., Luo, H., & Hou, X. (2018). Violence detection in surveillance video using low-level features. PLoS one, 13(10), e0203668.

[6] Papadopoulos, K. (2020). From Dense 2D to Sparse 3D Trajectories for Human Action Detection and Recognition.

[7] Ramzan, M., Abid, A., Khan, H. U., Awan, S. M., Ismail, A., Ahmed, M., ... & Mahmood, A. (2019). A review on state-of-the-art violence detection techniques. IEEE Access, 7, 107560-107575.
[8] Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., & Gong, B. (2021). Movinets: Mobile video networks for efficient video recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16020-16030).
[9] Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. IEEE transactions on pattern analysis and machine intelligence, 28(12), 2037-2041.
[10] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).
[11] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4690-4699).
[12] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018, May). Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018) (pp. 67-74). IEEE.
[13] Singh, S. Incorporating Mtcnn and Deepface: A Novel Hybrid Deep Learning Method for Human Face Detection and Recognition.
[14] Omiotek, Z. (2025). Effectiveness of Modern Models Belonging to the YOLO and Vision Transformer Architectures in Dangerous Items Detection. Electronics, 14(17), 3540.
[15] Cheng, M., Cai, K., & Li, M. (2021, January). RWF-2000: An open large scale video database for violence detection. In 2020 25th International conference on pattern recognition (ICPR) (pp. 4183-4190). IEEE.
[16] Mukto, M. M., Hasan, M., Al Mahmud, M. M., Haque, I., Ahmed, M. A., Jabid, T., ... & Islam, M. (2024). Design of a real-time crime monitoring system using deep learning techniques. Intelligent Systems with Applications, 21, 200311.
[17] Sung, C. S., & Park, J. Y. (2021). Design of an intelligent video surveillance system for crime prevention: applying deep learning technology. Multimedia Tools and Applications, 80(26), 34297-34309.
[18] Irfanullah, Hussain, T., Iqbal, A., Yang, B., & Hussain, A. (2022). RETRACTED ARTICLE: Real time violence detection in surveillance videos using Convolutional Neural Networks. Multimedia Tools and Applications, 81(26), 38151-38173.
[19] Xue, R., Chen, J., & Fang, Y. (2020, October). Real-time anomaly detection and feature analysis based on time series for surveillance video. In 2020 5th International Conference on Universal Village (UV) (pp. 1-7). IEEE.
[20] Maktoof, M. A., HM, I., Abdul Razzaq, M. A., Abbas, A., & Majdi, A. (2023). Machine Learning-Based Intelligent Video Surveillance in Smart City Framework. Fusion: Practice & Applications, 11(2).