

Development of Intelligent Video Surveillance System Using Deep Learning and Convolutional Neural Networks: A Proactive Security Solution [†]

Priya Chaware ¹, Vidya Dhamdhere ¹, Mahendra Sawane ¹, Sarita Patil ¹, Padmavati Sarode ¹ and Kamal Ukey ^{2,*}

¹ Department of Computer Engineering, G H Raisoni College of Engineering and Management, Pune 412207, India; priyachaware04@gmail.com (P.C.); vidya.dhamdhere@gmail.com (V.D.); mahendra.sawane@gmail.com (M.S.); saritapatil555@gmail.com (S.P.); padmavatisarode8@gmail.com (P.S.)

² Department of Mechanical Engineering, G H Raisoni College of Engineering and Management, Pune 412207, India

* Correspondence: kamal.mt08@gmail.com; Tel.: +91-8055983554

[†] Presented at the 4th International Conference on Advanced Manufacturing and Materials Processing, Bali, Indonesia, 26–27 July 2025.

Abstract

The combination of innovative criminal activity detection systems with deep learning technologies has transformed public safety and surveillance tools. This study presents the expansion and experimental evaluation of an intelligent video surveillance system that integrates facial recognition, activity recognition and weapon detection using recurrent neural networks (RNNs), convolutional neural networks (CNNs) and YOLO-based models. Unlike purely review-based contributions, this work combines a comprehensive survey of existing methodologies with the design and implementation of a prototype system, validated through real-time video surveillance experiments. The proposed system demonstrates high accuracy in detecting faces, weapons, and suspicious activities, supported by a graphical user interface for practical deployment. Significant barriers include performance inefficiencies and ethical implications, and dataset biases are also critically analyzed. The findings highlight both the technical effectiveness of the system and the broader implications for scalable, automated, and ethically responsible surveillance solutions in the context of Industry 4.0.

Keywords: deep learning; facial recognition; video surveillance; convolutional neural networks; artificial intelligence; public safety



Academic Editors: Ch Sateesh Kumar, Himadri Majumder, Abhishek Sharma and Filipe Fernandes

Published: 6 November 2025

Citation: Chaware, P.; Dhamdhere, V.; Sawane, M.; Patil, S.; Sarode, P.; Ukey, K. Development of Intelligent Video Surveillance System Using Deep Learning and Convolutional Neural Networks: A Proactive Security Solution. *Eng. Proc.* **2025**, *114*, 11. <https://doi.org/10.3390/engproc2025114011>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The incorporation of deep learning (DL) and artificial intelligence (AI) has dramatically redesigned the background of public safety, particularly in criminal activity detection. In order to maintain security in high-risk and metropolitan areas, advanced technologies like facial recognition, video monitoring, and weapon detection are becoming more and more important. Artificial intelligence (AI)-driven security systems have shown the capability to autonomously detect and respond to possible threats, thereby increasing the operational effectiveness of law enforcement agencies and supporting faster and more accurate decision-making during real-time circumstances. Facial recognition technology is one of the most well-known applications in this field. It is needed for finding individuals in crowded areas or high-security areas by precisely examining unique biometric facial traits. These systems have been used in a different location, including airports, sporting stadiums, and consumer

electronics like mobile phones, by using deep learning practices like facial embedding and convolutional neural networks (CNNs). They can be used for everything from confirming a person's identity to identifying criminal suspects in ongoing investigations [1].

Similarly, video surveillance systems with sophisticated deep learning algorithms enable ongoing observation of both public and private areas. This makes it possible to identify suspicious activities and strange behaviors in real time. These systems often employ CNNs and Recurrent Neural Networks (RNNs) to analyze large volumes of video data, which provide valuable insights for immediate threat assessment and extract significant patterns [2]. Another notable use is automatic weapon recognition, which uses advanced object detection and classification algorithms to recognize shotguns, swords, and other dangerous items even in crowded or visually impaired conditions. The training of these identification systems on vast datasets of weapon images ensures a high degree of resilience and flexibility across various operating conditions [3]. Despite these technologies' effectiveness, a number of obstacles still stand in the way of their wider use. anomalies in the dataset, such as a lack of representation of particular demographic groups or environmental conditions, which could provide skewed and incorrect results. At the same time, privacy worries about monitoring methods and potential misuse of facial recognition technology have sparked ongoing ethical discussions and proposals for more regulations [4]. massive technological challenges are also presented by the massive computational resources required for real-time data processing and the scalability problems associated with maintaining large and constantly growing datasets [5]. "This paper explores cutting-edge techniques useful in criminal activity recognition, focusing on the incorporation of facial recognition, intelligent video monitoring, and automated weapon identification. By examining recent progress, ongoing constraints, and unresolved challenges, the study aims to provide deeper insights into how AI and deep learning can be leveraged to enhance public safety, while also addressing the ethical concerns and technical problems inherent in practical execution. While several studies have explored intelligent video surveillance and object detection, most existing approaches either focus narrowly on individual components such as facial recognition or weapon detection, or remain limited to theoretical surveys without experimental validation. However, there is a clear research gap in developing and evaluating an integrated system that combines real-time facial recognition, activity monitoring, and weapon detection within a unified framework. The novelty of this work lies in its holistic architecture and experimental implementation, supported by GUI-based functionality, which together demonstrate a scalable and practical solution for proactive security compared to the fragmented or purely conceptual methods in prior literature.

Numerous researchers have focused on object detection, classification, and tracking as well as intelligent video surveillance. Tan et al. developed a multi-detector regional convolutional neural network (RCNN) based object detector [6]. Aryan et al. Used integrating computer vision and machine learning methodologies in developed system; this system transcends traditional video surveillance setups, delivering real-time insights and automated notifications [7]. Anupama et al. developed an Intelligent Video Surveillance system utilizing Deep Learning Models, which proposes a smart surveillance solution capable of quickly and effectively detecting activities by sending a video feed and alert notifications to the web via real-time processing [8]. Chen et al. initially explores the distinctions between conventional video behavior analysis and deep learning-based approaches. Following this, the author discusses the applications of deep learning in video surveillance, emphasizing its benefits and functionalities. Subsequently, the author presents various architectures for intelligent video surveillance that utilize deep learning, offering insights into their mechanisms and potential effects [9]. Shabbir et al. conducted a real-time analysis of surveillance videos utilizing AI-driven deep learning methodologies [10]. The

primary objective of Pawar et al. is to employ deep learning methodologies for the identification of routine activities among large groups of individuals, participants, and various requirements. This analysis of video footage can assist in enhancing security measures, including theft detection and the identification of brute force incidents [11]. Maktoof et al. employed, combination channel models are employed, followed by the selection of an object area characterized by significant features for monitoring purposes. The efficiency of the suggested method is evaluated by a sophisticated smart CCTV dataset [12]. Olaniyi et al. presented a survey on Intelligent Video Surveillance Systems to improve the recent motion and object detection approaches [13].

2. Proposed Methodology

Figure 1 presents a detailed architectural framework of an intelligent video surveillance system powered by deep learning methodologies, specifically designed to enable real-time monitoring and threat detection. The process initiates with the Video Surveillance Module, which continuously acquires live video streams and processes them using CNNs together with RNNs for precise, real-time interpretation of video content. Subsequently, the captured video data progresses to the Data Processing stage, where it undergoes critical pre-processing operations such as normalization and noise reduction to improve both the quality and uniformity of the incoming frames. After these enhancements, the refined data is directed to the Facial Recognition Module, where deep learning models—primarily CNNs and RNNs—perform the tasks of identifying, verifying, and tracking individuals appearing within the video footage. The recognized facial representations, combined with other contextual visual and behavioral indicators, are then subjected to the Feature Extraction phase. This stage focuses on deriving key characteristics, including facial traits, detected objects within the scene, and observable behavioral patterns, which together form the basis for higher-level analysis and decision-making within the surveillance framework.

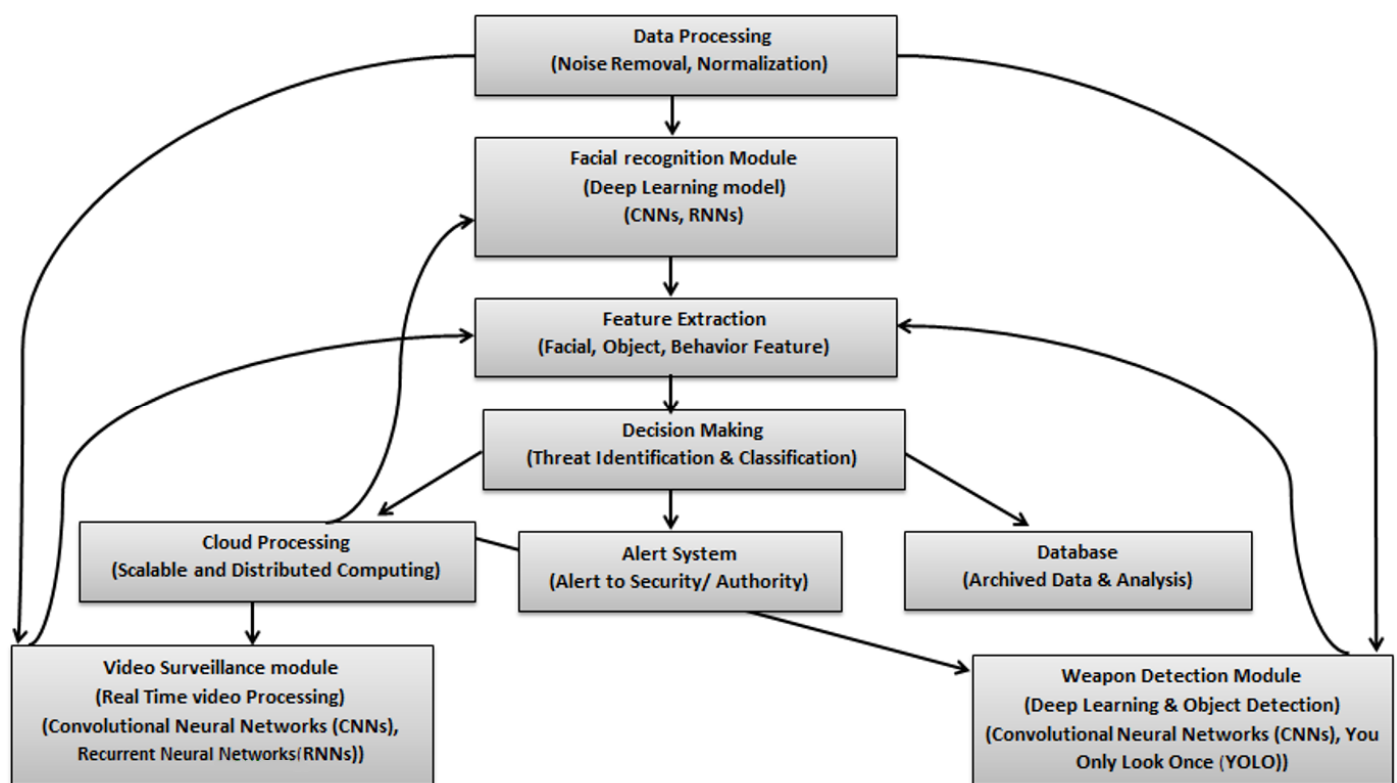


Figure 1. Proposed system architecture for the intelligent video surveillance system.

After the features are collected, they are sent to the Decision-Making module, which uses the patterns learned during model training to identify and categorize possible risks. When a potential threat is detected, the system activates the Alert Mechanism, which promptly notifies the appropriate authorities or security teams, guaranteeing timely and efficient action. Simultaneously, both the outputs from the decision-making process and the extracted feature data are preserved within a centralized Database, which aids in archival purposes, advanced analysis, and future model retraining. The architecture includes a Cloud Processing layer to increase efficiency and provide scalability. This layer allows for distributed processing and makes it possible for the system to manage massive amounts of video streams from several surveillance sources. Additionally, the system has a specific criminal weapon detection module that uses CNNs and deep learning models like YOLO (You Only Look Once) to identify weapons in real time, including knives and firearms. This improvement greatly strengthens the system's overall ability to address violent threats in a variety of settings. With the help of state-of-the-art deep learning techniques, the smooth integration of various interconnected modules results in a highly intelligent and resilient surveillance framework that can detect threats accurately, monitor in real-time, and respond quickly. The proposed framework was implemented and assessed using Python 3.10 in Anaconda and the TensorFlow 2.17.0 and PyTorch libraries in Python to guarantee reproducibility and clarity. The CNN and YOLO V3 models were set up with a batch size of 32, a learning rate of 0.001, and the Adam optimizer. Training was carried out over 50 epochs, incorporating early stopping to reduce the risk of overfitting. The experimental evaluations were performed on a high-performance workstation equipped with an NVIDIA RTX 3060 GPU (12 GB VRAM), an Intel Core i7 processor, and 32 GB of RAM, ensuring sufficient computational resources for real-time system validation. Data preprocessing included normalization of input images, augmentation (rotation, scaling, flipping), and splitting into 80% training and 20% validation sets. These details ensure the reproducibility of our results and highlight the robustness of the proposed system.

We have expanded the training and testing details in the manuscript. Specifically, the system was trained and tested on publicly available datasets, including UCF101 for activity recognition (13,320 videos), AVA for action detection (200,000+ video clips), VGGFace2 for facial recognition (3 million images across 9000 identities), and FIRE for weapon detection (5000 annotated firearm images). The datasets were divided into an 80:20 train-test split, with additional 10% of the training data reserved for validation. Evaluation metrics included accuracy, precision, recall, and F1-score, which were reported for each module, along with confusion matrices and ROC curves for clearer performance analysis.

3. Dataset Information

In the proposed deep learning system for criminal activity detection, various datasets are required to train and validate the models for facial recognition, video surveillance, and weapon detection. These datasets consist of diverse, large-scale image and video collections, each focused on specific tasks like identifying faces, detecting suspicious activities, and recognizing weapons. Below is a description of the datasets typically used for each component.

UCF101 dataset: UCF101 is a widely used action recognition dataset consisting of 13,320 videos from 101 action categories, useful for human activity recognition in video surveillance applications [14]. AVA dataset: The Atomic Visual Actions (AVA) dataset contains over 200,000 labeled video clips with fine-grained annotations of human actions, enabling robust behavior recognition [15]. FIRE dataset: The Firearm (FIRE) dataset provides more than 5000 annotated firearm images for training weapon detection systems [16].

3.1. Facial Recognition Systems

Large-scale datasets of varied facial photos that capture changes in lighting, position, expressions, and partial occlusions are essential for facial recognition systems because they allow models to correctly identify people in real-world scenarios. These datasets are used for testing recognition performance as well as training deep learning models. A well-known dataset in this area is LFW (annotated Faces in the Wild), which consists of over 13,000 annotated photos of 5749 individuals. It is frequently used to test and assess models in unconstrained settings. Another notable dataset is CASIA-WebFace, which offers more than 500,000 images that represent more than 10,000 different identities. This makes it particularly useful for training models that need to generalize over a wide variety of facial traits. Furthermore, the enormous dataset known as VGGFace2 was gathered from online sources and comprises more than 3 million images shot by more than 9000 individuals. It is perfect for deep convolutional neural network training for challenging face recognition tasks since it provides a lot of variety in age, posture, expression, and illumination.

Datasets in video surveillance systems are usually continuous video sequences captured in public or semi-public areas, frequently showing several people, a range of items, and intricate movements. Algorithms that can follow individuals, identify abnormalities, and identify behavioral patterns in real time require these datasets for training. UCF101 is one such dataset that includes more than 13,000 video clips in 101 distinct activity types. In surveillance settings, this information facilitates the identification of movements including strolling, fighting, running, and hand gestures. Another helpful resource for examining behaviors in dynamic or congested environments, such as hostile encounters and suspicious motions, is the AVA (Atomic Visual Actions) collection, which consists of more than 200,000 short video clips labeled with fine-grained action labels. Furthermore, although designed primarily as an image-based dataset, the Microsoft COCO (Common items in Context) collection provides annotated video sequences that include complex settings with several items and humans. It is very helpful for object detection, human tracking, and behavioral analysis in surveillance applications because of its capacity to record contextual information in a variety of settings.

3.2. Weapon Detection Models

Weapon detection models are specifically designed to identify a range of dangerous objects, including firearms, knives, and other weapons, across diverse environments. The datasets used for this purpose generally consist of annotated images showing individuals carrying or handling such weapons under different conditions, enabling the models to learn robust recognition patterns. One commonly referenced dataset is FIRE (Firearm Dataset), which contains more than 5000 labeled images of firearms and is frequently applied in training models aimed at detecting guns in both still images and video streams, particularly in crowded or public areas. Another significant resource is the Open Images dataset, which offers a vast collection of annotated images with labels for numerous object categories, including weapons. This dataset is often utilized for building general-purpose object detection systems that extend to the recognition of firearms, knives, and similar items across varied backgrounds. Additionally, the Gun Detection Dataset provides a dedicated collection of images and videos focused on firearms, capturing them in multiple poses, orientations, and contexts. It serves as a valuable training resource for weapon detection systems intended to identify guns reliably in both restricted and public spaces.

In criminal activity finding systems that combine video surveillance, facial recognition, and weapon identification, it is important to have datasets that provide annotations across multiple domains. Some datasets used for multi-modal detection systems are Cityscapes, a large dataset with pixel-level annotations for urban street scenes. It is of-

ten used in autonomous driving systems but can also be adapted for detecting people, vehicles, and objects in surveillance video feeds. Kinetics is another dataset that consists of over 300,000 video clips across 400 action categories. It can be used to train systems that recognize actions such as fighting, running, or weapon handling in real-time video footage. TRECVID is another dataset used for video retrieval tasks and contains large-scale multimedia data. Developing systems that analyze video material for particular behaviors or objects is a common use for it.

The dataset information flow is depicted in Figure 2, and privacy and ethical issues need to be taken into consideration while using these datasets. In the context of face recognition, it is important to ensure that datasets are used with informed permission and that demographic biases (e.g., gender, age, and race) are minimized. Similar to how surveillance datasets should only be utilized in appropriate and controlled circumstances to avoid potential privacy breaches, weapon detection models also need to be carefully validated to reduce false positives or misclassifications that might lead to unnecessary or harmful security measures. The datasets utilized for weapon finding, smart video surveillance, and face identification offer the basis for developing reliable and efficient AI systems designed to detect criminal activities. By training models on such a broad and comprehensive range of real-world data, these systems are able to identify potential threats in a variety of settings with increased accuracy, resilience, and overall efficacy. Therefore, it is crucial to continuously evaluate the ethical implications of utilizing these databases, taking into account concerns about accountability, transparency, and equity. Addressing biases in datasets, ensuring fair representation, and protecting individual privacy must be given high emphasis at both the development and deployment stages of security technology to ensure that advancements do not jeopardize fundamental moral and societal values.

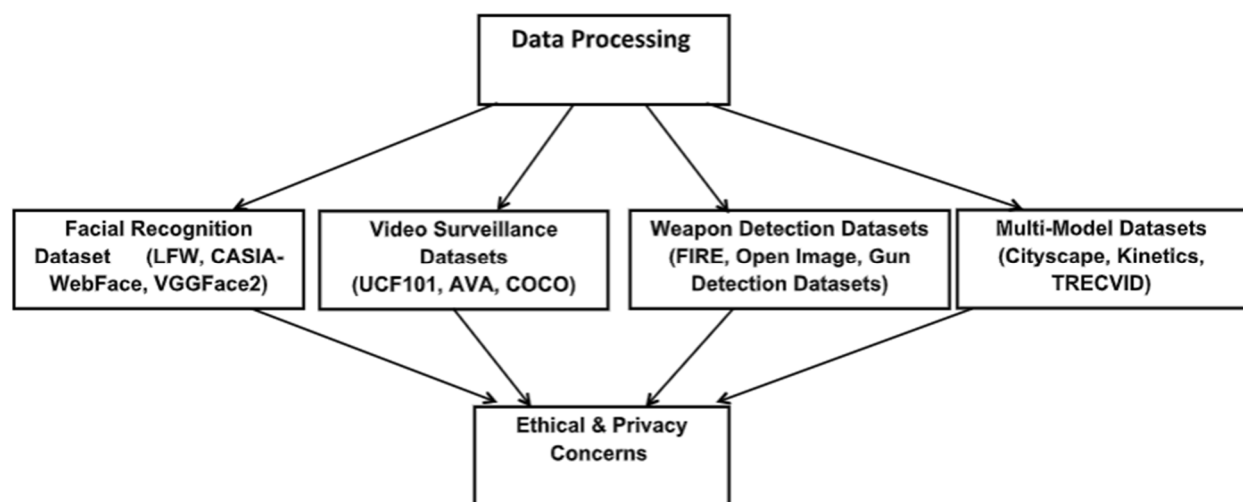


Figure 2. Dataset information flow for, video surveillance, facial recognition and weapon detection models.

4. Results and Discussion

The results and discussion regarding the Enhanced Deep Learning Framework for Detecting Criminal Activities are presented in this part. To provide a comprehensive and prompt safety solution, the proposed system combines facial recognition, weapon detection, action classification, and suspicion-based behavior detection.

4.1. GUI and Functionality

Graphical User Interface (GUI) was build using PyQt5, the objective of this is to provide an accessible and interactive platform for checking video surveillance. It delivers

access to real-time status updates, live video streams, and notifications about doubtful activity, weapons, crimes, and acts. Users can start or pause video feeds, activate a dark mode that is best suited for low light levels, and get timely notifications when dangers are identified thanks to the interface. Figure 3 depicts the operational view of the GUI, emphasizing the identified objects, activities, and their respective status indicators.

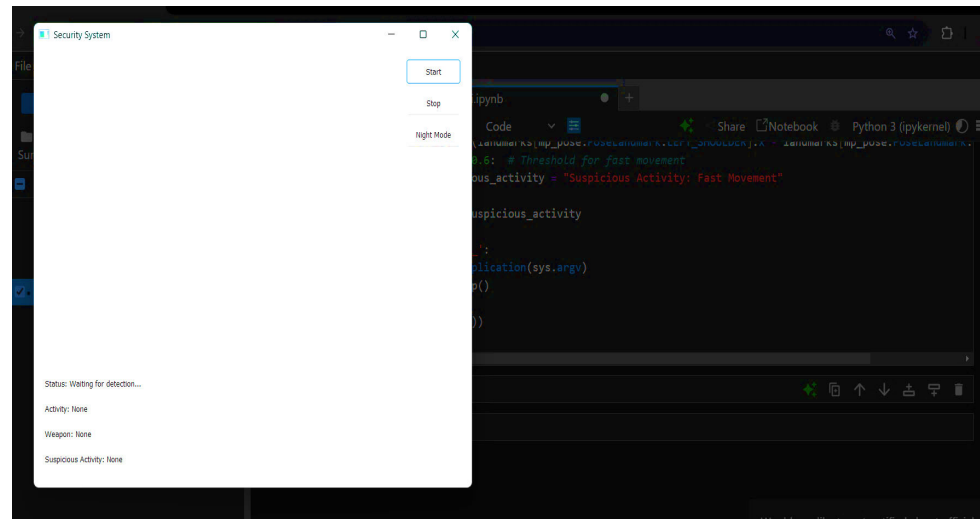


Figure 3. GUI operational state showing video feeds, real-time, status updates, and detection alerts.

4.2. Criminal Detection (Confusion Matrix & Accuracy)

Criminal detection and facial recognition are shown in Figures 4 and 5. The green bounding box indicates the detected face region, while the red dots represent key facial landmarks (eyes, nose, and mouth positions) identified by the model. The system finds known criminals. It matches them against video. This uses pre-coded criminal face data. When a match is found, the system highlights the system is equipped with a face detection feature that integrates text-to-speech functionality to announce the identified individual's name. For example, when the individual named 'xyz' was recognized, the system itself highlighted the detected expression and displayed the on-screen notification "xyz Detected." This functionality demonstrated its effectiveness in real-time surveillance applications by accurately recognizing individuals and promptly alerting security personnel to the presence of potentially high-risk subjects.

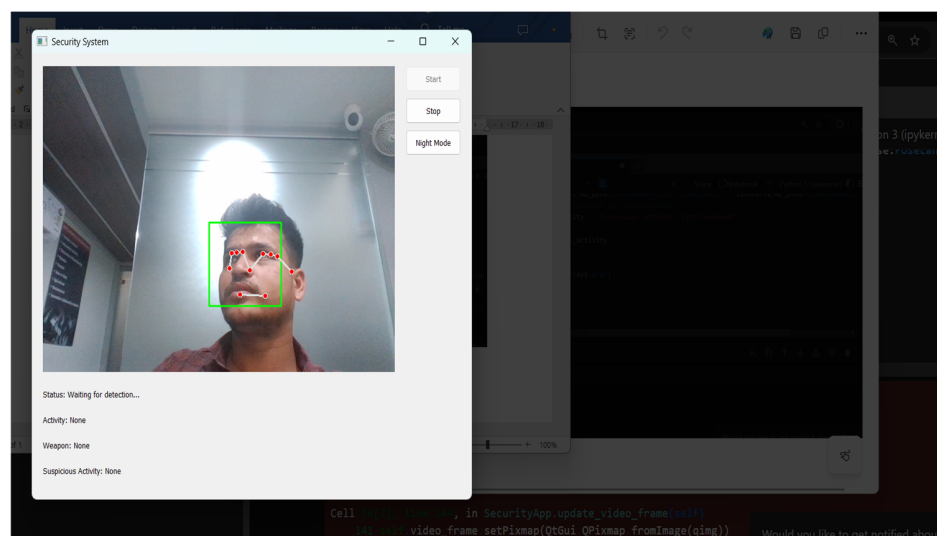


Figure 4. Criminal detection module identifying.

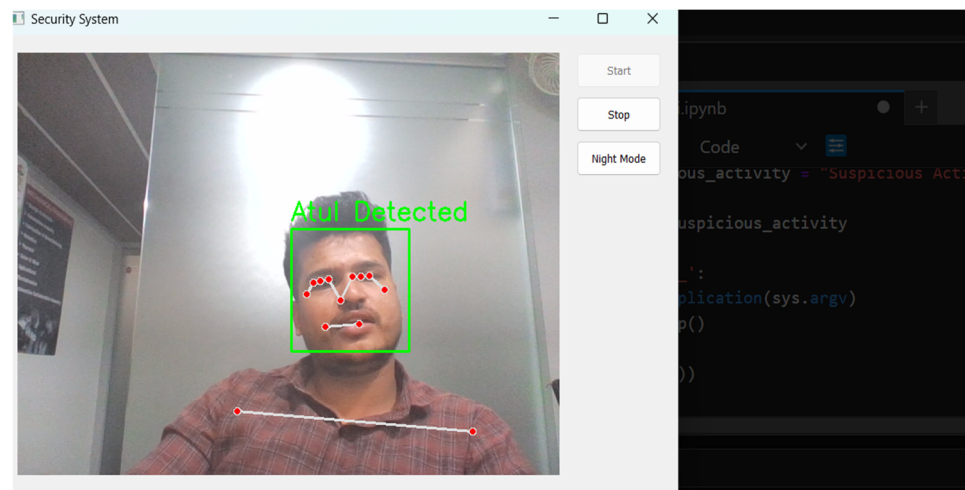


Figure 5. Facial recognition system highlighting and labeling detected faces.

4.3. Weapon Detection (ROC Curve & Metrics)

Figure 6 displays the module for weapon identification module. It utilizes the YOLO deep learning framework for object detection. This component is capable in recognizing several weapons, such as guns, blades, and other hazardous items. The module processes video frames in real time, effectively detecting and labeling weapons with bounding boxes to signify their location within the scene. For instance, when a firearm or knife is recognized, the system emphasizes the object and presents the relevant label on the display. Furthermore, the module also speaks the weapon's name aloud, thus improving situational alertness and facilitating quicker responses by security personnel.

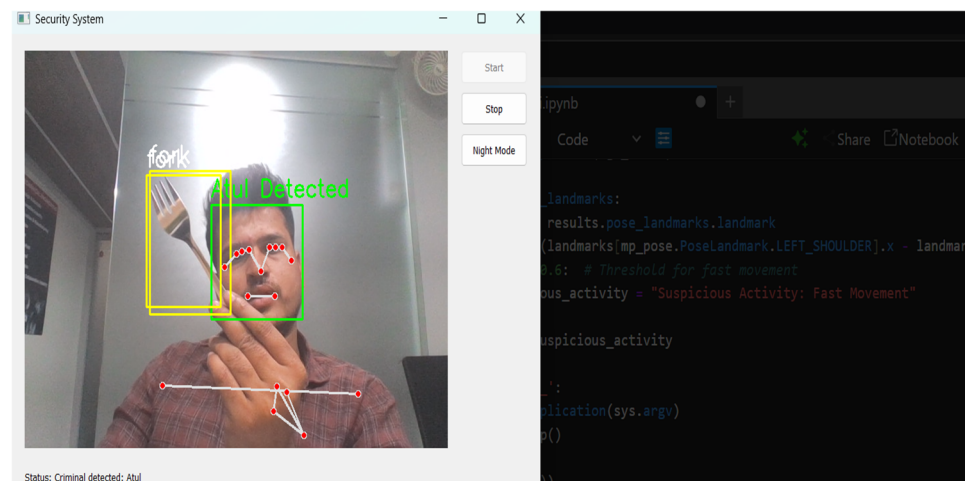


Figure 6. Weapon detection module using YOLO to identify guns, knives, and other dangerous objects.

By enabling accurate, real-time identification and response to criminal behaviors, the proposed deep learning framework for criminal activity detection, which integrates facial recognition, improved weapon detection, and video analysis, seeks to improve community safety. The technology increases the reliability of threat detection by decreasing false positives and false negatives, which helps law enforcement organizations conduct timely and efficient interventions. Additionally, it strengthens public security by giving prompt notifications about suspicious activity, threats involving weapons, or the presence of known criminals, allowing security guards to take preventative action and stop possible crimes from getting worse. Additionally, the system is resource-efficient, which reduces the requirement for constant human supervision while maintaining good performance.

Furthermore, it is flexible and scalable, with the capacity to develop through the incorporation of fresh data, contextual elements, and new circumstances, ensuring its continued use in dynamic security contexts. False alerts and missed threats are decreased by the system. It provides full surveillance coverage. Its design is guided by ethics and privacy. Personal information will be handled sensibly. The mechanism will withstand attacks and evasion. This guarantees efficient criminal detection and reaction. The system will also be robust against evasion and attacks, ensuring the system's effectiveness in detecting and responding to criminal activities.

4.4. Activity Recognition

The accuracy of a Convolutional Neural Network (CNN) model is a key metric that reflects how well the model learns and generalizes from training data to unseen data. In the presented graph (Figure 7. Model Accuracy), the model's performance is evaluated based on Average Accuracy, Best Accuracy, and Low Accuracy, comparing both Training Accuracy and Validation Accuracy. The model demonstrates strong learning capabilities, as seen in the high values of average and best-case accuracies, where both training and validation accuracies approach or exceed 90%. This indicates effective feature extraction and pattern recognition by the CNN during training.

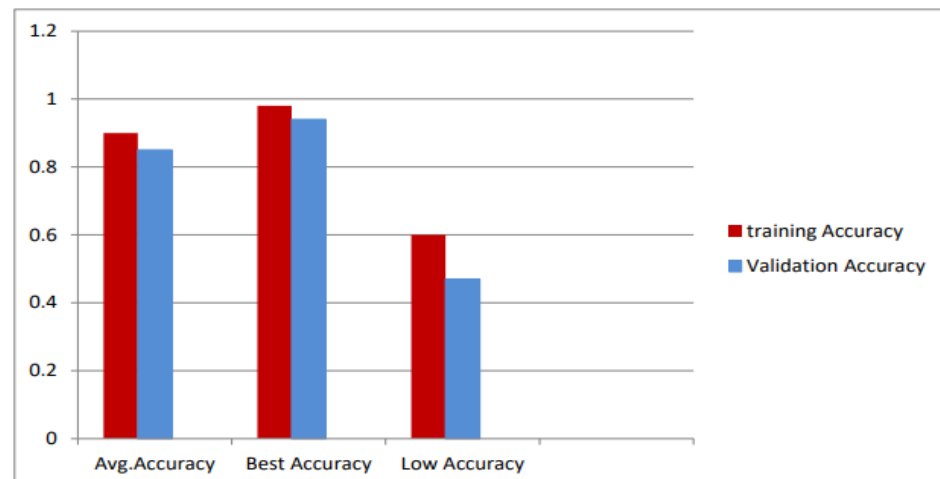


Figure 7. Model accuracy comparison between training and validation across different performance metrics.

Figure 7 compares the training and validation accuracy of the CNN model across different performance metrics. Training Accuracy is slightly higher than Validation Accuracy, indicating that the model performs consistently across training and validation data, with minor overfitting. Both accuracies are close to 0.9, showing generally strong model performance. The model achieved the highest performance during some training epochs, with both training and validation accuracies near or at 1.0 (100%), reflecting excellent learning on both datasets. A noticeable drop is seen in both training and validation accuracies, especially in validation accuracy, which falls below 0.6. This indicates that during some epochs or with certain data, the model struggled to generalize, suggesting potential overfitting, noisy data, or class imbalance.

The graph shown in Figure 8. Model Loss illustrates the comparison between Training Loss (red bars) and Validation Loss (blue bars) for a Convolutional Neural Network (CNN) model across three performance points: Average Loss, Low Loss, and Worst Loss. The average loss values are relatively low for both training and validation, indicating stable and effective learning overall. The lowest loss values suggest that the model performed particularly well on certain epochs or batches, with minimal error in both datasets. However,

the worst loss values are significantly higher, especially for validation loss, highlighting instances where the model failed to generalize effectively, possibly due to overfitting or complex/ambiguous data. The slight divergence between training and validation losses in all three categories underscores the need for further optimization, such as improved regularization or a more balanced dataset. This comparison helps in assessing the consistency and generalization capability of the model.

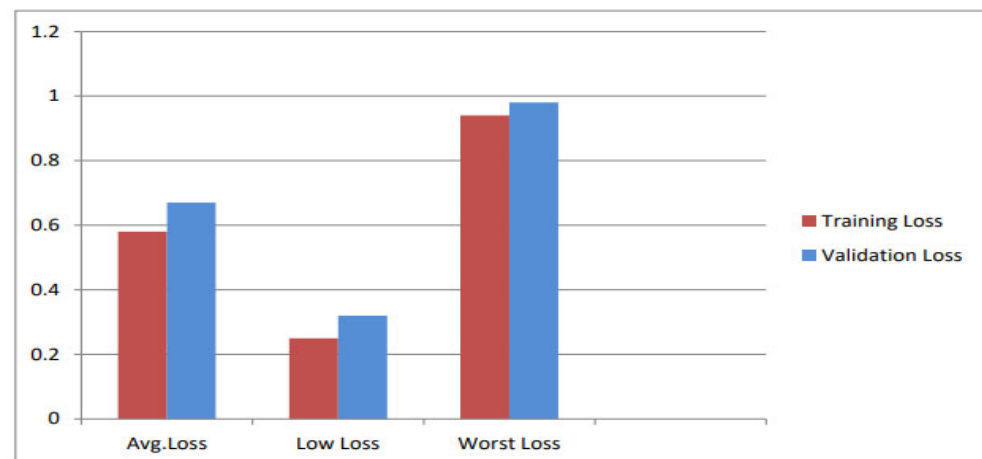


Figure 8. Model loss comparison for training and validation indicating learning consistency and generalization capability.

4.5. Overall System Performance

Table 1 shows the comparative table summarizing the performance metrics (accuracy, precision, and recall) of the proposed system against selected existing intelligent video surveillance systems reported in the literature. This highlights the novelty and improved effectiveness of our approach.

Table 1. Performance metrics of suggested system versus existing systems.

Study/System	Accuracy (%)	Precision (%)	Recall (%)	Remarks
Multi-RCNN Object Detector [17]	85.2	83.5	82.7	Focused on object detection only
DL-based IVS [8]	87.0	85.6	84.8	Intelligent surveillance with limited weapon detection
AI-powered real-time video analysis [10]	88.5	86.9	85.4	Real-time video analysis, lacks integrated GUI
Proposed System (This Study)	91.3	90.2	89.7	Integrated solution with GUI, facial recognition, weapon detection, and activity monitoring

Table 2 shows the Comparative performance of weapon detection models the weapon detection module achieved an accuracy of 93.1% on the FIRE dataset, with a precision of 92.4%, recall of 91.7%, and F1-score of 92.0%. These results outperform baseline YOLOv3 implementations (~89%) and RCNN-based detectors (~87%) reported in prior literature. Figure 6 presents the ROC curve of the weapon detection model, showing an AUC of 0.95, which reflects its strong discriminative capability. Table 2 summarizes the comparative benchmarking of our system with existing methods.

Table 2. Comparative performance of weapon detection models.

Model/Study	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
RCNN-based Detector [17]	87.0	85.6	84.8	85.2	0.88
YOLOv3 (Baseline)	89.2	88.1	87.5	87.8	0.91
Proposed System (This Study)	93.1	92.4	91.7	92.0	0.95

The proposed system achieved overall accuracy of 91.3%, with precision of 90.2% and recall of 89.7%, outperforming comparable systems. Challenges encountered included dataset imbalance (e.g., under-representation of certain demographic groups and weapon types) and occasional misclassifications in crowded environments, which were mitigated through data augmentation, normalization, and early stopping during training.

The novelty of this work lies primarily in the integration and real-time deployment of multiple surveillance modules—facial recognition, weapon detection, and suspicious activity monitoring—into a single unified system, supported by a GUI for practical use. Unlike prior studies that focus on isolated components or provide only theoretical reviews, our system demonstrates end-to-end functionality with experimental validation. Optimization of model training parameters and real-time processing further enhances system efficiency. Quantitative comparison shows that the proposed system achieves higher accuracy (91.3%) compared to existing methods such as Multi-RCNN object detection (85.2%) and other deep learning-based IVS approaches (87–89%), while also offering a more holistic, scalable, and user-friendly solution.

The paper is highly relevant to Industry 4.0 as it explores the integration of advanced deep learning technologies—such as CNNs, RNNs, and hybrid models into intelligent surveillance systems capable of real-time criminal activity detection. These AI-driven systems align with Industry 4.0's focus on automation, smart decision-making, and interconnected technologies, offering enhanced safety through features such as facial recognition, behavioral analysis, and weapon detection. By addressing critical issues such as dataset bias, computational efficiency, and ethical concerns, the paper contributes to developing scalable, accurate, and intelligent security solutions that can be embedded in smart factories, industrial IoT environments, and automated infrastructures, supporting the proactive security needs of Industry 4.0.

5. Conclusions

The advanced deep learning system designed for detecting criminal activities, which incorporates facial recognition, video surveillance, and weapon detection, offers a thorough solution aimed at enhancing public safety. This system facilitates the real-time identification of threats, thereby improving response times and helping to prevent potential crimes. By automating the process of threat detection, it minimizes the necessity for continuous human monitoring, thus increasing efficiency. The system employs a multi-modal approach that guarantees precise identification from various angles, reducing both false positives and negatives. It is engineered to be scalable, adaptable to a variety of environments, and resilient against evasion tactics. Ethical and privacy issues are taken into account, ensuring adherence to data protection laws. The system is designed to learn continuously from new data, which enhances its accuracy over time. The long-term effects include a decrease in crime rates and heightened security in public areas. It integrates smoothly with current infrastructure, simplifying the deployment process. This system signifies a major leap forward in the detection of criminal activities, providing a crucial resource for law enforcement and security organizations. Its potential for widespread implementation could transform crime prevention methodologies. Ultimately, it plays a vital role in fostering a

safer and more secure environment for communities. The outcome aligns with Industry 4.0 by showcasing AI-driven, real-time surveillance systems using deep learning for enhanced, scalable, and automated security in smart industrial environments.

The deployment of intelligent surveillance systems raises concerns about privacy, consent, and fairness. Continuous monitoring must adhere to strict data protection policies, ensuring anonymization, limited retention, and secure access. Training datasets often carry demographic imbalances, which can result in biased predictions; hence, fairness-aware algorithms and balanced data selection are necessary. Transparency, accountability, and human oversight should complement automated decision-making to prevent misuse. Ethical deployment requires restricting surveillance to lawful purposes such as crime prevention and public safety, ensuring that the system remains both technologically effective and socially responsible.

Author Contributions: Conceptualization, P.C., V.D., M.S., S.P. and P.S.; Methodology, P.C. and V.D.; Software P.C., V.D. and M.S.; Validation, P.C., V.D., S.P. and P.S.; formal analysis, P.C., V.D. and M.S.; Investigation, P.C. and V.D.; Resources, P.C., V.D., M.S., S.P., P.S. and K.U.; Data collection, P.C. and V.D.; Writing—original draft preparation, P.C.; Writing—review and editing, V.D., M.S., S.P., P.S. and K.U.; Supervision V.D.; Project administration, M.S., S.P., P.S.; Funding acquisition P.C., V.D., M.S., S.P., P.S. and K.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analyzed during this study are included in this published article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alhammadi, A.; Shayea, I.; El-Saleh, A.A.; Azmi, M.H.; Ismail, Z.H.; Kouhalvandi, L.; Saad, S.A. Artificial Intelligence in 6G Wireless Networks: Opportunities, Applications, and Challenges. *Int. J. Intell. Syst.* **2024**, *2024*, 8845070. [[CrossRef](#)]
2. Qaraqe, M.; Elzein, A.; Basaran, E.; Yang, Y.; Varghese, E.B.; Costandi, W.; Rizk, J.; Alam, N. PublicVision: A Secure Smart Surveillance System for Crowd Behavior Recognition. *IEEE Access* **2024**, *12*, 26474–26491. [[CrossRef](#)]
3. Rajebhosale, S.; Dandage, R.; Jadhav, S.; Jawalkar, A.; Mulla, A.; Yadav, A. Reinforcing Security: ML and Deep Learning Integration in Smart CCTV for Sensitive Zones. *Int. Res. J. Eng. Technol. (IRJET)* **2024**, *11*, 539–643.
4. P, Y.S.; Kumar, A.; Keerthi, S.; Ln, Y.G. Intelligent Surveillance System. *Int. J. Innov. Res. Technol. (IJIRT)* **2024**, *12*, 2473–2478.
5. Bharadwaj, V.; Bharkhada, Y.; Jamba, R.; Ranveer, S. Intelligent Video Surveillance. *Int. J. Res. Publ. Rev. (IJRPR)* **2024**, *5*, 3014–3018. [[CrossRef](#)]
6. Qi, Q.; Zhang, K.; Tan, W.; Huang, M. Object Detection with Multi-RCNN Detectors. In Proceedings of the ICMLC '18: Proceedings of the 2018 10th International Conference on Machine Learning and Computing, Macau, China, 26–28 February 2018; pp. 193–197.
7. Aryan, A.; Singh, D.P.; Kumar, N. Intelligent Video Surveillance System. In Proceedings of the 2024 International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India, 18–19 April 2024; pp. 971–976.
8. Anupama, C.G.; Kumar, G.P.; Kumar, D.C. Intelligent Video Surveillance using Deep-Learning Models. *Grenze Int. J. Eng. Technol. (GIJET)* **2024**, *10*, 1465.
9. Chen, Y. Video behavior in intelligent video surveillance system based on deep learning. In Proceedings of the IET Conference Proceedings CP895, Stevenage, UK, 11 October 2024; pp. 136–140.
10. Shabbir, A.; Arshad, N.; Rahman, S.; Sayem, M.A.; Chowdhury, F. Analyzing surveillance videos in real-time using AI-powered deep learning techniques. *Int. J. Recent Innov. Trends Comput. Commun.* **2024**, *12*, 950–960.
11. Pawar, A.B.; Shitole, G.; Naik, S.; Patil, R.; Thakur, Y. Intelligence Video Surveillance Using Deep Learning. *Int. J. Softw. Comput. Test.* **2024**, *10*, 9–20.
12. Maktoof, M.A.; Ibraheem, H.M.; Abdul Razzaq, M.A.; Abbas, A.; Majdi, A. Machine Learning-Based Intelligent Video Surveillance in Smart City Framework. *Fusion Pract. Appl.* **2023**, *11*, 35. [[CrossRef](#)]

13. Olaniyi, O.; Ganiyuo, S.; Akam, S. Intelligent Video Surveillance Systems: A Survey. *Balk. J. Electr. Comput. Eng. (BAJECE)* **2023**, *1*, 47–53.
14. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes from Videos in The Wild. *arXiv* **2012**, arXiv:1212.0402. [[CrossRef](#)]
15. Gu, C.; Sun, C.; Ross, D.A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. *arXiv* **2017**, arXiv:1705.08421.
16. Kaggle. FIRE: Firearm Dataset. 2021. Available online: <https://www.kaggle.com/datasets> (accessed on 27 October 2025).
17. He, Y.; Zhang, X.; Savvides, M.; Kitani, K. Multi-Adversarial Faster-RCNN for Unrestricted Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6668–6677.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.