

# Real-time crime monitoring system using deep learning for weapon, behavior, and facial detection

S. Arul<sup>1</sup> and P. Kavitha<sup>2,\*</sup>

<sup>1</sup>Department of Computer Science, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Pallavaram, Chennai, India

<sup>2</sup>Department of Computer Applications, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Pallavaram, Chennai, India

\*E-mail: pkavikamal@gmail.com

Received for publication  
May 28, 2025.

## Abstract

This study presents an efficient crime monitoring system (CMS) that leverages the UCF crime dataset, which contains actual surveillance footage of various criminal activities. The CMS integrates real-time video surveillance with deep learning (DL) models for crime detection, including weapon identification, misbehavior detection, and facial recognition. It utilizes hyper capsule networks (H-CapsNets) for identifying visible and concealed firearms, 3D convolutional neural network (C3D) for detecting anomalous behavior, and deep convolutional neural network (CNNs) for facial feature extraction and matching with a suspect database. Real-time detection is achieved with low response times—generating alerts in <1.5 s for weapon detection and 2.2 s for misbehavior detection, ensuring rapid intervention. Data augmentation techniques, including rotation, scaling, flipping, and panning, enhance model generalization, resulting in 93.5% accuracy, 92.8% precision, 91.7% recall, and 92.2% F1-score. The CMS demonstrates excellent real-time performance with low false positive (FP, 3.1%) and false negative (FN, 2.7%) rates, showcasing its potential for practical deployment in public safety systems.

## Keywords

crime monitoring system, weapon detection, misbehavior detection, facial recognition, UCF dataset, C3D, CNN

## 1. Introduction

Weapon detection, misbehavior detection, and facial recognition-related violence pose a significant threat to public safety, especially in urban areas. Traditional closed-circuit television (CCTV) systems rely on manual monitoring, often leading to delayed responses. This study proposes a real-time gun detection system that uses a deep learning (DL) algorithm and convolutional neural networks (CNNs) for accurate and rapid firearm identification in surveillance footage. The system automates weapon detection, issues instant alerts, and supports timely intervention. Integrating with existing CCTV infrastructure enhances situational awareness and response efficiency, contributing to safer environments through intelligent, automated surveillance driven by advanced computer vision technology

[1]. Improving security and stopping criminal activity require close observation of human behavior in public and private areas. The manual observation, a major component of traditional surveillance systems, is laborious and prone to human error. Automated behavior analysis has become more accurate and efficient due to recent DL developments. CNN is one of the examples of DL model that is used in this article to identify and evaluate human behavior in real time. The study aims to enhancing surveillance systems by facilitating the intelligent, real-time identification of anomalous, or questionable activity patterns in intricate settings [2]. Video surveillance has evolved over time, transitioning from analog to digital systems. Due to increased public interest in public safety and crime prevention, image security and surveillance systems have grown significantly, supported by the availability of high-quality,

affordable imaging technology [3]. Security is always a top issue in all fields because of the increased crime rate in congested regions and suspicious isolated places. Computer vision has significant uses in abnormal detection and monitoring to address various issues. Owing to the increasing need for safety, security, personal property protection, and intelligence monitoring, video surveillance systems that can recognize and comprehend scenes and unusual events are crucial [4]. Facial expression-based emotion recognition has drawn a lot of interest. People's widespread feelings of dread and insecurity highlight the urgency of proactively addressing such dangers. Handguns, knives, firearms, and other portable weapons are often used in numerous terrible crimes that endanger public safety [5]. The monitoring system's original investment and ongoing maintenance costs are substantial. These expenses are mostly related to the required human and hardware resources. The surveillance team is a crucial but sometimes disregarded component of surveillance systems [6]. Although manual monitoring is laborious and prone to human mistakes, surveillance cameras have become essential for maintaining public safety. As artificial intelligence advances, DL methods provide a potent way to automate identifying illegal activity. CNNs are used in various DL models. It is possible to analyze video frames in real time and identify suspicious behaviors accurately. This approach enhances surveillance systems by providing timely alerts, reducing response time, and improving overall security [7]. Real-time surveillance systems have become essential in ensuring public safety, particularly in dynamic and densely populated environments. This study suggests a DL-based method for the spatiotemporal detection of abnormal human behavior. By integrating CNNs, the system effectively captures spatial and temporal patterns in video data [8]. Thus, the development of an automated border surveillance system that can perform the surveillance job without requiring human involvement is urgently needed [9]. CCTV cameras are considered one of the most critical security-related criteria and are an essential part of the solution to this problem. The main goals of CCTVs, which are currently installed in every public space, are security detection, safety, and criminal investigation. CCTV footage is the most crucial evidence in courts [10].

### a. Objective

- To create a real-time crime monitoring system (CMS) that combines facial recognition, behavior analysis, and weapon detection to identify and react to crimes via video surveillance automatically.

- To utilize spatial hierarchies in video frames to effectively identify visible and concealed firearms by using the hyper capsule network (H-CapsNet).
- To extract spatiotemporal information from video sequences using a 3D convolutional neural network (C3D) to detect unusual or suspicious behavior.
- To reliably identify offenders, compare photographed pictures to a saved database using deep CNNs for facial recognition.

### b. Contribution work

- The study presents a comprehensive CMS architecture that combines facial identity, misbehavior recognition, and weapon detection to enable real-time detection and warning in public surveillance systems.
- It uses an H-CapsNet to identify both visible and concealed weapons. This network uses the spatial hierarchies in video frames to provide a more reliable and precise identification.
- Spatiotemporal data in movies are efficiently captured and analyzed using a C3D to detect aggressive and aberrant human behaviors.

The remaining portion of the study is divided into significant sections, which are described as follows: Section II examines the current research efforts in real-time crime monitoring systems using deep learning for weapon, behavior, and facial detection used by different authors. Section III explains the workflow of the suggested approach in the proposed methodology. Section IV presents the findings, analysis, and performance data. Section V presents the conclusion.

## II. Related Work

Mukto et al. [11] introduced a novel and efficient CMS designed to leverage video surveillance for real-time crime detection and immediate notification to law enforcement authorities. The system was developed to overcome typical human limitations, such as inattentiveness, delayed responses, and oversight in recognizing criminal behavior. By integrating DL models with advanced image processing techniques, the CMS effectively utilized the capabilities of CCTV networks. The framework operated in three critical stages: facial recognition for suspect identification, violence detection for recognizing aggressive actions, and weapon detection to flag potential threats.

Panda et al. [12] focused on classifying cyberattacks using the recent, noisy, and unbalanced Internet of Things (IoT)-Botnet dataset known as UNSW-NB15 Dataset (UNSW-NB15). They use four contemporary

methodologies to validate the ideas that they have suggested. A semi-naïve Bayesian averaged two-dependence estimator (A2DE) and Hierarchical Graph Convolution (HGC), which combine genetic algorithms with K-means clustering, enhance the chi-square automated interaction detection decision tree method. Additionally, two DL models are used: a CNN-based classifier and a deep multilayer perceptron. The goal of the project is to enhance the precision and effectiveness of cyberattack detection in intricate IoT systems.

Vuyyuru et al. [13] proposed an inventive response to the rising rates of crime worldwide that is an automated drone-based street crime detection system. The technology analyses drone photos using sophisticated CNN models. Images are broken down into base and detail layers using the embedding bilateral filter technique to improve detection accuracy. The proposed fusion model, Information Retrieval (IR) with attention-based ConvViT, successfully combines Inception-V3, ResNet-50, and Convolution Vision Transformer. This integration captures both shape and texture features efficiently, making the system robust for real-time crime detection in dynamic urban environments.

To determine the network systems anomaly, we used CNNs. Akhtar and Fen [14] presented a revolutionary binary and multiclass classification model. The NSL-KDD Dataset (NSLKDD) dataset was utilized in this instance. Our model uses a CNN to perform both binary and multiclass classification. We developed a DL-based DoS detection model for both datasets.

Nazir et al. [15] proposed a method that utilizes YOLOv5 for object detection, where the obtained temporal characteristics are bounding box coordinates. The topic is presented as a time-series classification job, and Deep Sort is used to monitor individuals across a video clip. The modern robust temporal feature magnitude technique, which uses the inflated 3D ConvNet, is contrasted with the suggested method for pre-processing. Experimental evaluations were conducted using the widely recognized UCF crime dataset to assess performance. The method aims to enhance crime detection accuracy by leveraging both spatial and temporal features.

Sahay et al. [16] used DL architectures to suggest a unique method for crime scene video surveillance systems that identify violence in real time. Here, the objective is to collect crime scene footage from real-time surveillance systems and apply the spatio-temporal technique to extract features using a deep reinforcement neural network-based classification algorithm. The characteristics were extracted and classified after the incoming footage was processed

and converted into video frames. Its goal is to identify signs of violence and animosity immediately so that anomalies may be differentiated from regular trends.

Ullah et al. [17] presented an intelligent anomaly detection system based on deep features that can operate well in surveillance networks with less temporal complexity. The suggested architecture uses a multi-layer bidirectional long short-term memory model to process a series of video frames after first extracting spatiotemporal information from the frames. This model has been pre-trained with a CNN to enhance its ability to recognize anomalous events. As a result, the system can reliably classify ongoing normal or abnormal activities in complex surveillance environments, making it highly suitable for smart city applications.

Raza et al. [18] developed a predictive system for maternal health issues using artificial neural networks and real-world medical data. They proposed that a Dual-Tree Bidirectional Long-Term Convolutional Network (DT-BiLTCN) is a new DL architecture that combines temporal convolutional networks, decision trees, and a bidirectional long short-term memory network. A total of 1218 samples from hospitals, community clinics, and maternity health facilities were gathered using an IoTs-based risk monitoring system to create the dataset. Using the synthetic minority oversampling technique, the data's class imbalance is addressed. This approach enhances prediction accuracy and supports proactive maternal healthcare through advanced DL techniques.

Verma et al. [19] reviewed standard methods for detecting and identifying suspicious activities, covering both supervised and unsupervised machine learning (ML) techniques previously used by researchers. These approaches range from modeling individual behavior to analyzing crowded environments and primarily rely on classifiers such as ANN, SVM, and Hidden Markov Models. The study also discusses system models designed to distinguish between normal and abnormal human behaviors, along with various feature selection and detection methods employed in earlier research to improve detection accuracy.

Palanivinayagam et al. [20] focused on vulnerability analysis by extracting key variables, such as time zones, crime areas, and crime frequency, to increase ML algorithm accuracy. Two common datasets were used to evaluate the suggested approaches. The outcomes showed that their feature-building strategy greatly improved the ML models' performance. The efficiency of the suggested method in crime data analysis was demonstrated by the notable achievement of the maximum accuracy of 97.5% utilizing the Naïve Bayes algorithm on the San Francisco dataset.

Alzahrani and Bamhdi [21] presented a robust approach to identifying botnet assaults on devices connected to the Internet. The CNN model with a lengthy short-term memory processing mechanism was imaginatively integrated with four various types of security cameras to identify two common and important IoT hazards. Real-time lab-connected camera equipment in IoTs settings provided the datasets, which included typical malicious network traffic.

Zahrawi and Shaalan [22] provided research based on real-time object detection systems to automatically detect firearms in video surveillance systems. We provide an early weapon identification framework using cutting-edge and real-time object recognition techniques such as You Only Look Once (YOLO) and single shot multi-box detector (SSD). To use the model in practical applications, we also considered drastically lowering the number of false alerts. The approach works well with inside security cameras in gas stations, supermarkets, banks, shopping centers, and other establishments.

Zhang et al.'s [23] primary contribution is creating a fraud detection system that combines a DL architecture with a sophisticated feature engineering approach based on homogeneity-oriented behavior analysis. To evaluate the efficacy of the suggested framework, we conducted a comparison analysis using an actual dataset from one of China's biggest commercial banks. The trial's findings demonstrate the practicality and efficacy of our proposed strategy for identifying credit card fraud.

Akhtaret et al. [24] employed the CNN-LSTM technique, a hybrid DL strategy, to identify botnet assaults. Without the administrator's permission, malicious software gets installed on a computer system or infiltrates it. Numerous viruses are used by cybercriminals to accomplish their malevolent goals. The increasing quantity of harmful applications has prompted the development of a novel DL system. The system employs Natural Language Processing (NLP) approaches as a baseline, records local spatial correlations by combining CNNs and Long Short-Term Memory (LSTM) neurons, and learns from subsequent long-term dependencies.

Chinedu et al. [25] examined the advancements made over the past 10 years in applying ML models to support the development of clever solutions to reduce the threat posed by cybercrimes. Their study focuses on published resources from reputable databases and adopts an inquisitive perspective. It downplays the alleged benefits of some reportedly clever anti-cybercrime tactics while highlighting their application and potential [25-30].

Kuklin et al. [31] proposed a model to enhance the reliability of biometric authentication systems by addressing data drift—a common issue that reduces model accuracy over time. Their framework uses dynamic data correction to maintain stable performance, which is particularly relevant to surveillance systems using facial recognition.

Kurdthongmee and Kurdthongmee [32] developed a fast and accurate pupil detection method using semantic segmentation fine-tuned on a lightweight CNN backbone. This approach improves eye tracking and gaze estimation in real-time applications, offering insights into optimizing lightweight models for surveillance and human behavior analysis.

Abumalloh et al. [33] conducted a text mining and qualitative study to explore how individuals experience security breaches and cyber-attacks. Their findings provide a human-centered perspective on cybersecurity risks, which complements technical monitoring systems like CMS by highlighting the need for user awareness and trust.

Previous studies on crime detection and video surveillance systems have made notable contributions as in Table 1, but suffer from several critical limitations that hinder their practical deployment:

- **Limited real-time performance:** Many traditional models (e.g., CNN-only or YOLO-based systems) suffer from latency issues due to sequential processing, which delays alert generation during critical incidents [11, 22].
- **Weak multimodal integration:** Several existing systems focus on either weapon detection, misbehavior analysis, or facial recognition in isolation, lacking a unified architecture to handle multifaceted crime detection concurrently [15, 16].
- **Inadequate handling of spatiotemporal features:** Models that rely solely on 2D CNNs or classical motion analysis (e.g., optical flow) often fail to detect subtle, evolving human behavior patterns over time [17].
- **Poor detection of concealed objects:** Conventional object detection frameworks like YOLO and SSD perform well on visible weapons but struggle with partially concealed weapons due to limited hierarchical spatial encoding [22].

To address these gaps, our proposed CMS integrates multiple advanced DL models within a unified, real-time pipeline:

- The H-CapsNet outperforms conventional object detectors by capturing spatial hierarchies and detecting both visible and concealed firearms with high precision.

**Table 1: Summary of Existing studies.**

| References No. | Dataset used                     | Algorithm used       | Result achieved                                                                                                      |
|----------------|----------------------------------|----------------------|----------------------------------------------------------------------------------------------------------------------|
| [26]           | AWID dataset                     | ML algorithm         | With a detection rate of 99.75% and an accuracy of 99.45%, the model performed better than the others                |
| [27]           | CICD DoS2019 dataset             | DL algorithm         | F1-score and accuracy rate >98%                                                                                      |
| [28]           | UCF crime dataset                | ML algorithm         | The average of 98.0%. In the meantime, our PBVAD-MIM approach yielded an average success rate of 80.7% for the tests |
| [29]           | CIC-DDoS 2019 dataset            | ML and DL algorithms | Obtained a 99.50% accuracy rate with a delay                                                                         |
| [30]           | Historical daily weather dataset | ML and DL algorithms | Accuracy rate of 96.65% and 84.0%                                                                                    |

DL, deep learning; ML, machine learning; PBVAD-MIM, Priority-Based Vulnerability Assessment and Detection – Machine Intelligence Model.

- The C3D effectively captures spatiotemporal patterns from video streams, enabling accurate detection of anomalous or aggressive behavior in dynamic environments.
- A CNN-based facial recognition module ensures rapid and accurate identification of individuals using robust feature embeddings and real-time matching with suspect databases.
- Unlike fragmented previous approaches, the proposed system combines weapon detection, behavior analysis, and facial recognition within a single framework, allowing seamless and synchronized monitoring.
- Extensive testing on the University of Central Florida (UCF) crime dataset shows superior performance metrics (e.g., 93.5% accuracy, 92.2% F1-score) and response times <2.2 s, making it suitable for real-world deployment.

### III. Proposed Methodology

The suggested CMS uses three different methods to identify crimes in real time, as shown in Figure 1. It uses an H-CapsNet to detect firearms by examining video frames and determining which weapons are hidden or visible. Misbehavior can be identified using a C3D, which can locate odd behaviors. Deep CNNs use face recognition to find and identify suspects. The CMS makes efficient real-time crime monitoring and response possible by combining these elements. The system uses supervised learning to differentiate between benign and dangerous activity. Techniques for data augmentation improve the generalization and dependability of models. Users' situational awareness increases with CMS, enabling prompt emergency reactions. The system's potential for practical uses is demonstrated through testing and validation on the UCF crime dataset. All things considered, the CMS provides a strong option for improving public safety via clever video surveillance. It is a valuable tool for law enforcement because of its possibilities.

#### a. Dataset (UCF crime dataset)

Weapon detection, misbehavior detection, and face identification are three tasks that may be performed using this dataset, which includes all CMSs in one group and all routine activities in another. Our testing results show that our MIL strategy for weapon detection, misbehavior detection, and face identification generates a significant performance boost compared with the state-of-the-art techniques. We also provide results on detecting abnormal behavior from several recent DL baselines.

#### b. Weapon detection

Several classes are involved in the task of detecting weapons. The model must not only detect the presence of a weapon but also accurately identify its type. The model is trained using images of various weapons, such as handguns, from a weapons database. This helps determine whether the weapon being detected is deadly or not. CMS prioritizes searching for firearms because it can ultimately determine whether a crime is occurring, a H-CapsNet object identification method for weapons discovery. From model training to application development, users can quickly transition with the help of the H-CapsNet. Objects of interest in an image are predicted and recognized by H-CapsNet promptly and efficiently.

$$F_{frame} = I_f \in \mathbb{R}^{H \times W \times C} \quad (1)$$

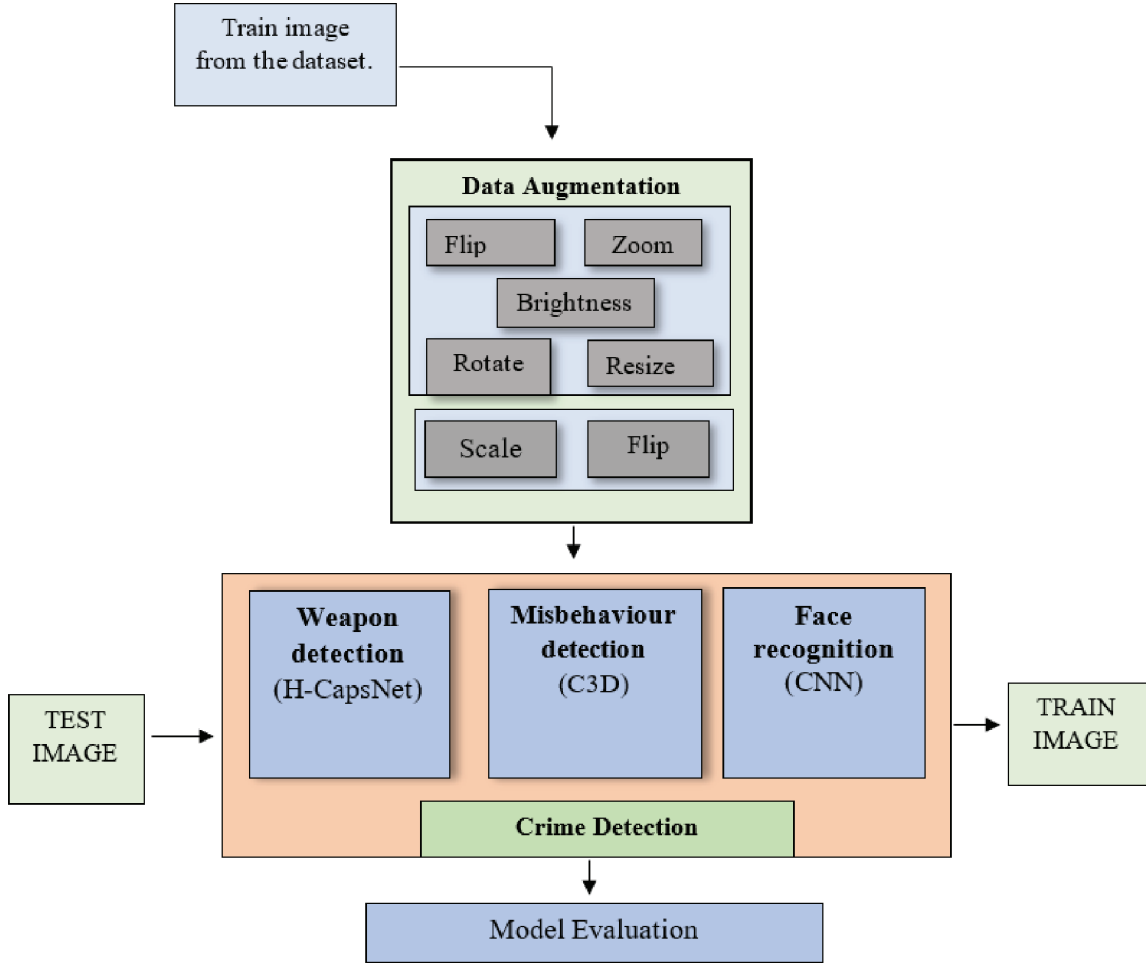


Figure 1: Proposed methodology block diagram.

where

- $F_{frame}$  – Input frame features

$$u_i = W_i \cdot X + b_i \quad (2)$$

- $u_i$  – Output vector of the  $i^{th}$  capsule in the primary layer
- $W_i, b_i$  – Weights and bias matrices
- $X$  – Flattened or pooled input feature vector

An important part of the capsule network's routing procedure is the total input to higher capsule. To provide the input for a higher-level capsule, predictions from lower-level capsules must be combined. Each prediction vector  $\hat{u}_{j|i}$  from a lower capsule is weighted by its coupling coefficient  $c_{ij}$ , which reflects the strength of the connection. The final score is obtained by summing these weighted predictions.

$$s_j = \sum_i c_{ij} \hat{u}_{j|i} \quad (3)$$

where

- $c_{ij}$  – Coupling coefficient between capsule  $i$  and  $j$
- $\hat{u}_{j|i}$  – Prediction of capsule  $i$  for capsule  $j$

After integrating inputs from lower-level capsules, the output of a higher-level capsule signifies the capsule's final activation. The probability and instantiation characteristics of the identified feature, such as a weapon, are both encoded in this vector.

$$v_j = \text{squash}(v_j) \quad (4)$$

- $v_j$  – Output vector of capsule  $j$ , encoding both presence and pose
- $\text{squash}(\cdot)$  – Nonlinear activation ensuring the output vector length is between 0 and 1

The length of  $v_j$  represents the probability of the detected entity.

The vector length of the output capsule is used by the final weapon detection decision to assess if a weapon is present in a video frame. After processing, a vector is produced by the weapon-class corresponding capsule  $v_{\text{weapon}}$ , whose length indicates the probability of detection as in algorithm 1

$$w_d = \begin{cases} \text{Detected, if } \|v_{\text{weapon}}\| > \delta \\ \text{Not Detected, otherwise} \end{cases} \quad (5)$$

- $\|v_{\text{weapon}}\|$  – Magnitude (length) of the weapon capsule output.
- $\delta$  – Detection threshold (a value between 0 and 1, e.g., 0.5).
- $w_d$  – Final weapon detection decision.

#### Algorithm 1: Weapon detection algorithm

1. Extract low-level features from  $F_g$  using CNN
2. For each primary capsule  $i$ :  
Compute output vector:  
$$u_i = W_i \cdot X + b_i$$
3. For each higher-level capsule  $j$ :  
For each lower-level capsule  $i$ :  
Compute prediction vector:  
$$\hat{u}_{ji} = w_{ji} \times u_i$$
4. Initialize coupling coefficients  $c_{ij}$  using routing softmax
5. For routing iterations:  
For each capsule  $j$ :  
$$s_j = \sum_i c_{ij} \hat{u}_{ji}$$
  
Compute capsule output:  
$$V_j = \text{squash}(v_j)$$
  
Update  $c_{ij}$  based on agreement between  $v_j$  and  $\hat{u}_{ji}$
6. Select  $v_{\text{weapon}}$  from the capsule corresponding to the weapon class
7. Final Decision:  
If  $\|v_{\text{weapon}}\| > \delta$   
$$w_d = \text{"Detected"}$$
  
Else:  
$$w_d = \text{"Not Detected"}$$

Identifying and classifying weapons based on a reliable H-CapsNet model are key components of crime detection. Using images annotated to the data is essential for H-CapsNet's performance because the finer points of its weapon identification processes are determined by H-CapsNet files, which act as careful configuration files for H-CapsNet.

#### c. Misbehavior detection

The C3D model is a supervised learning model that uses 3D convolution networks to train enormous amounts of image data. Furthermore, because of its spatiotemporal structure and  $3 \times 3 \times 3$  kernel shape, it is quite successful in learning image data. In the past, 2D convolution networks were used to analyze images. A crucial challenge in computer vision is the automated identification of suspicious or aberrant behavior in real-time, whether in social settings or during video surveillance. DL methods, in particular CNNs, which can collect both spatial and temporal characteristics of video input, are among the most successful ways to tackle this issue.

Convolutional 3D networks, or C3D models, are robust supervised learning frameworks that use 3D convolutional operations to evaluate video data. The C3D model extends convolutions into the temporal dimension, allowing it to learn motion patterns and temporal dynamics across successive frames, in contrast to conventional 2D convolutional networks that only analyze spatial information inside a single visual frame. By combining temporal and spatial data into a single DL architecture, the C3D model marks a substantial breakthrough in video-based behavior analysis. For applications requiring automated monitoring, public safety, and anomaly detection, this makes it a top option.

The spatial and temporal structure of the C3D model, which uses 3D Convolution Networks, makes it very effective in learning picture data about misbehavior. Learning misbehavior patterns in CCTV images is a prerequisite for employing such C3D to identify aberrant conduct in the photographs. Figure 2 illustrates how the C3D model learns and identifies anomalous activity in CCTV footage. To begin with, learning requires learning.

The input video clip representation is the foundational step in misbehavior detection using a C3D. It involves segmenting a continuous video stream into fixed-length clips, typically comprising TT consecutive frames, denoted as  $V_t = \{F_1, F_2, \dots, F_T\}$ ,  $F_i \in \mathbb{R}^{H \times W \times 3}$ .

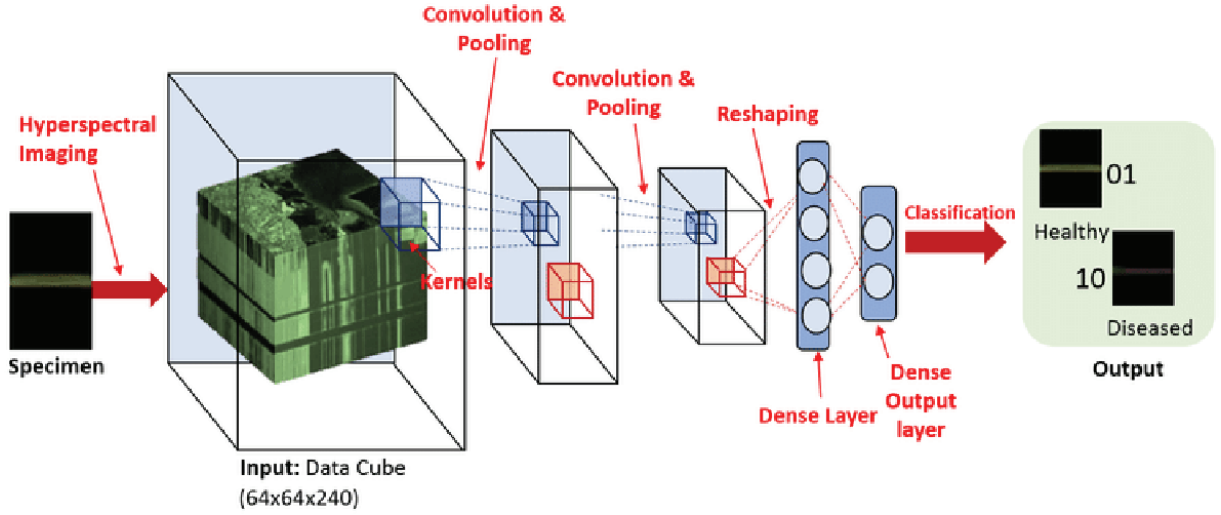


Figure 2: C3D architecture. C3D, 3D convolutional neural network.

Each frame  $F_i$  is a color image with dimensions  $H \times W \times 3$ , representing the height, width, and color channels. This sequential structure allows the model to capture both spatial and temporal features, enabling the detection of dynamic and suspicious human behaviors across multiple frames, rather than analyzing isolated still images.

$$V_t = \{F_1, F_2, \dots, F_T\}, F_i \in \mathbb{R}^{H \times W \times 3} \quad (6)$$

where

- $V_t$  – Sequence of video frames
- $T$  – Number of frames
- $H, W$  – Height and width

Frame normalization is a crucial preprocessing step in misbehavior detection that enhances model performance by standardizing the pixel intensity values of each video frame. It involves adjusting the raw pixel

values of a frame  $F_i$  using the formula  $F_i^{norm} = \frac{F_i - \mu}{\sigma}$ .

$$F_i^{norm} = \frac{F_i - \mu}{\sigma} \quad (7)$$

where

- $\mu$  is the mean.
- $\sigma$  is the standard deviation of pixel intensities.

This procedure lessens the impact of changing lighting, shadows, and camera exposure by guaranteeing that all input frames have comparable statistical

characteristics. Normalized frames decrease overfitting and increase convergence speed, which aids in the model's learning.

$$V_{stacked} \in \mathbb{R}^{T \times H \times W \times 3} \quad (8)$$

Temporal stacking of frames is a method used to organize a sequence of video frames into a structured format suitable for input into a C3D. In this process, consecutive frames  $F_1, F_2, \dots, F_T$ s are stacked along a new temporal dimension, forming a 4D tensor  $V_{stacked} \in \mathbb{R}^{T \times H \times W \times 3}$ . This structure preserves both spatial (height and width) and temporal (frame sequence) information. By stacking frames temporally, the network can analyze motion patterns and dynamic behaviors across time, which is essential for detecting misbehavior and recognizing actions in video surveillance footage.

$$Y = (V_{stacked} * K) + b \quad (9)$$

where

- $b$  is the bias.

3D convolution operation is a key process in misbehavior detection using C3D. Unlike 2D convolution, which only captures spatial features, 3D convolution processes both spatial and temporal dimensions simultaneously. It involves sliding a 3D kernel  $K$  over the input video tensor  $V_{stacked}$ , computing features across width, height, and time, producing an output feature map.

#### d. Face recognition (CNNs)

The input image in a face recognition system refers to the facial photograph captured by a camera or extracted from a video frame. It is typically represented as a color image in RGB format with dimensions  $H \times W \times 3H$  times, where  $HH$  and  $WW$  denote the height and width of the image, and the 3 represents the red, green, and blue color channels. This image serves as the primary data source for subsequent processing steps such as face detection, normalization, and feature extraction. The quality and resolution of the input image significantly influence the recognition accuracy.

$$I_f = \text{RGB}(H \times W \times 3) \quad (10)$$

where

- $I_f$  is the input face image with height  $H$  and width  $W$ .

$$I_{\text{gray}} = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B \quad (11)$$

Convert the RGB image to grayscale to simplify feature extraction if required.

$$B_f = f_{\text{det}}(I_f) \quad (12)$$

A face detector  $f_{\text{det}}$  outputs a bounding box  $B_f$  for the detected face region within  $I_f$ .

$$I_n = \frac{I_f - \mu}{\sigma} \quad (13)$$

Pixel values can be normalized by deducting the mean  $\mu$  and dividing by the standard deviation  $\sigma$  to standardize the input

$$V_f = f_{\text{CNN}}(I_n) \quad (14)$$

Pass a normalized face image  $I_n$  through a deep CNN  $f_{\text{CNN}}$  to extract a high-dimensional feature vector  $V_f$ .

$$E = W_e \cdot V_f + b_e \quad (15)$$

Transform  $V_f$  to a fixed-length embedding vector  $E$  using a fully connected layer with weights  $W_e$  and bias  $b_e$ .

$$V_{\text{norm}} = \frac{E}{\|E\|_2} \quad (16)$$

Normalize the embedding  $E$  to unit length (L2 normalization) for consistent distance comparison.

## IV. Result and Discussion

The suggested CMS uses video surveillance to detect and report crimes in real-time efficiently. It combines deep CNN-based facial recognition for suspect identification, H-CapsNets for weapon detection, and 3D CNNs for misbehavior recognition. The system uses supervised learning to differentiate between normal and deviant actions, as demonstrated by tests on the UCF crime dataset. Data augmentation approaches improve generalization. CMS helps create safer, more intelligent public surveillance systems by enabling quick emergency responses and enhancing situational awareness.

### a. Model evaluation

False positive (FP), false negative (FN), true negative (TN), and true positive (TP) are acronyms that indicate the ratio of photos that are accurately recognized as positive, photos that are mistakenly categorized as negative, photos that are mistakenly labeled as negative, and photos that are mistakenly classified as positive.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (17)$$

CMS must have a low false negative rate (FNR) to guarantee system security. FNR is supplied by

$$\text{FNR} = \frac{FN}{TP + FN} \times 100\% \quad (18)$$

F1-score, confusion matrix, and sensitivity, along with other criteria such as recall, accuracy, specificity, and precision, are used to evaluate the efficacy of the neural-based model system. A brief description of these measurements is given below. Accuracy is the percentage of data instances that are properly categorized overall.

Precision indicates what percentage of identifications are correct

$$\text{Precision} = \frac{TP}{TP + FP} \quad (19)$$

This proportion of correctly identified TPs is known as recall.

$$\text{Recall} = \frac{TP}{TP + TN} \quad (20)$$

The accuracy and recall metrics' harmonic mean is used to get the F1 score. Two classifiers are compared using it.

$$\text{F1 Score} = \frac{\text{Precision}}{\text{Precision} + \text{Recall}} \quad (21)$$

The performance analysis of the weapon detection model in Table 1 shows that it is overall very successful. The model's 93.5% accuracy rate indicates that many cases were correctly identified. With a recall of 91.7%, it effectively detects real weapons, while an accuracy of 92.8% suggests its ability to reduce false alarms. Reliability in handling recall and precision is demonstrated by the model's balanced F1 score of 92.2%. False positive rate (FPR) and FNR are low, at 3.1% and 2.7%, respectively, indicating its dependability in lowering incorrect classifications. For real-time weapon detection applications, this makes it a dependable choice.

As seen in Figure 3 and Table 2, the performance analysis of the weapon detection system shows excellent results, with 93.5% accuracy, demonstrating reliable weapon recognition in various settings. The model's 92.8% accuracy indicates that it can reduce false alarms, and its 91.7% recall rate suggests it can detect most real weapons. An F1 score of 92.2% confirms a balanced precision and recall performance. Furthermore, the

system's low FPR and FNR, 3.1% and 2.7%, highlight its resilience and decreased probability of misclassification, making it ideal for security applications.

The photos in Figure 4 show a computer vision and artificial intelligence-based weapon detection system. Using a bounding box marked "pistol," the algorithm detects a pistol being aimed by a human in the first image with 89% confidence. The system demonstrates its resilience in the second image by detecting a guy brandishing a knife even if it is upside down. Both situations demonstrate how well the technology can identify potentially harmful items like knives and firearms in real time. These kinds of systems are essential for public safety and monitoring because they allow law enforcement to react swiftly to any threats in delicate locations such as shopping centers, schools, or public areas.

Figure 5 shows the model's training accuracy evolution during 31 epochs. The accuracy starts at approximately 54% and improves markedly during the first several epochs, which indicates learning is taking place. Fluctuations are observed, but the norm is upwards after the fifth epoch, where it levels off at 85%. In the later epochs, the accuracy improves >90%, indicating good fine-tuning of parameters. The gradual rise and plateau suggest minimal overfitting, and the model converges during training.

Figure 6 displays the training loss across 32 epochs. The graph is declining. Initially, the loss is high (>0.7), but it gradually decreases throughout the epochs. After epoch 10, there are a few minor oscillations, but generally, it is declining and leveling

### Performance Metrics of Weapon Detection System

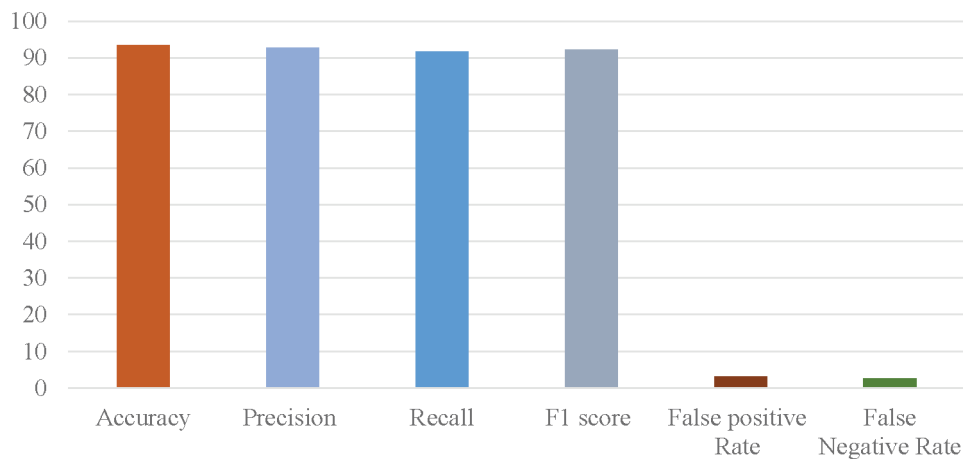


Figure 3: Performance metrics of the weapon detection system.



Figure 4: Weapon detection.

**Table 2: Weapon detection performance metrics**

| Metric    | Value (%) |
|-----------|-----------|
| Accuracy  | 93.5      |
| Precision | 92.8      |
| Recall    | 91.7      |
| F1 score  | 92.2      |
| FPR       | 3.1       |
| FNR       | 2.7       |

FNR, false negative rate; FPR, false positive rate.

off around 0.2. Since the loss is not growing, the model is convergent and not overfitting the architecture and training function well, as evidenced by the diminishing loss.

Figures 7A and 7B show the CNN-based face recognition detection model's accuracy and loss curves, emphasizing how well it learns from photos of firearms and crime scenes. By epoch 10, training and validation losses were decreased to 0.33 and 0.4, respectively, with little overfitting from their initial high values of 0.8 and 0.75. Strong generalization was demonstrated by the accuracy improvements from 85.36% to 91.49% (training) and 83.65% to 91.38% (validation).

A facial recognition system in operation is shown in Figure 8. The same person, identified as "Mukto," is detected and recognized by the algorithm in four distinct frames. Every detection has a green bounding box around it, and the recognition accuracy is shown by the confidence percentage. The results, which vary from 69.70% to 87.46%, demonstrate that the system operates dependably even when head angle and illumination circumstances change. This illustrates how facial recognition algorithms can reliably and instantly identify people. These systems are commonly used to automate very accurate identification verification in surveillance, security, and attendance monitoring applications.

To further evaluate the classification performance of the CMS components, confusion matrices were generated for both the weapon detection and misbehavior detection modules. These matrices help illustrate how well the models distinguish between TPs, TNs, FPs, and FNs. For weapon detection, the confusion matrix shows the system's ability to differentiate between the presence and absence of firearms in real-time surveillance frames. Likewise, the misbehavior detection matrix evaluates the accuracy in identifying suspicious human actions. As shown in Figure 9, the high number of TPs and TNs, along with low misclassification rates, confirms the robustness and reliability of both subsystems.

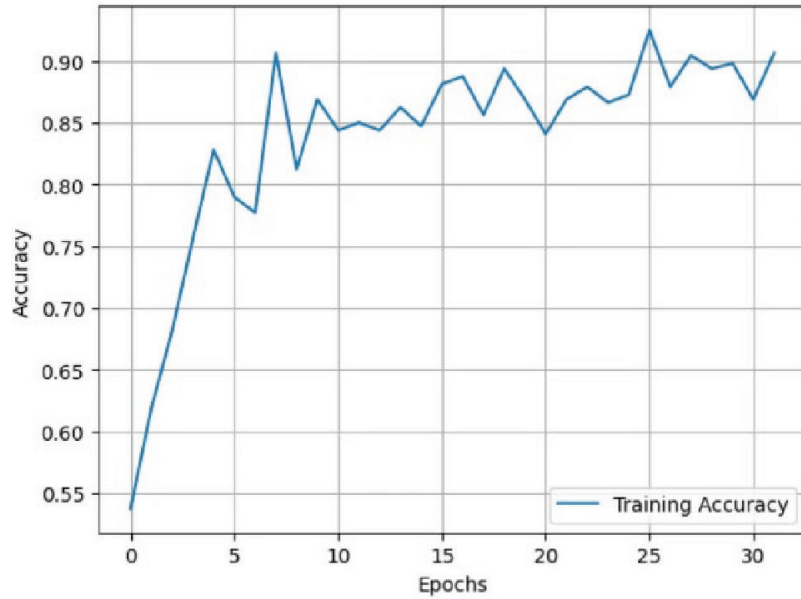


Figure 5: Accuracy plot.

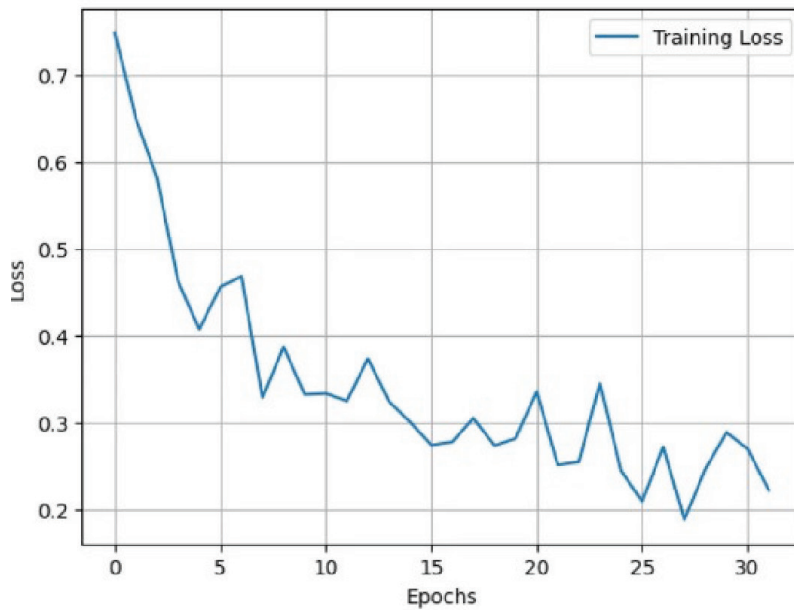


Figure 6: Loss plot comparison.

### a. Ablation study

To evaluate the individual contribution of each module within the CMS, an ablation study was conducted as in Table 3. The study involved selectively disabling each component—H-CapsNet (for weapon detection), C3D (for misbehavior detection), and CNN

(for facial recognition)—and observing the resulting system performance on the UCF crime dataset.

The full CMS configuration outperforms all others, indicating that each module contributes significantly to overall system performance. The largest performance drop is observed when H-CapsNet

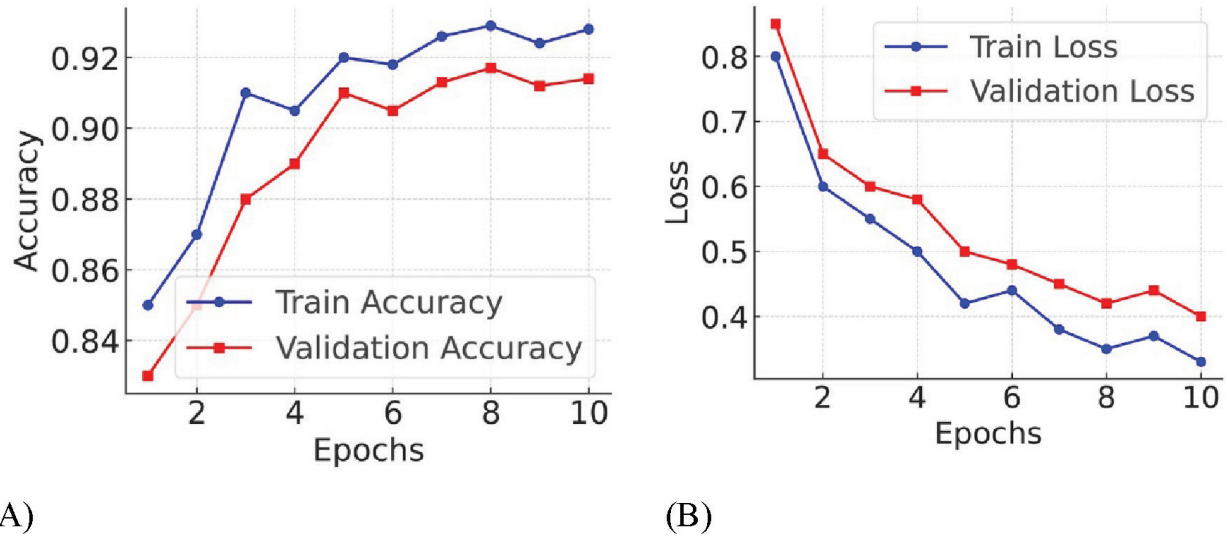


Figure 7: (A) Training and validation accuracy and (B) training and validation loss.

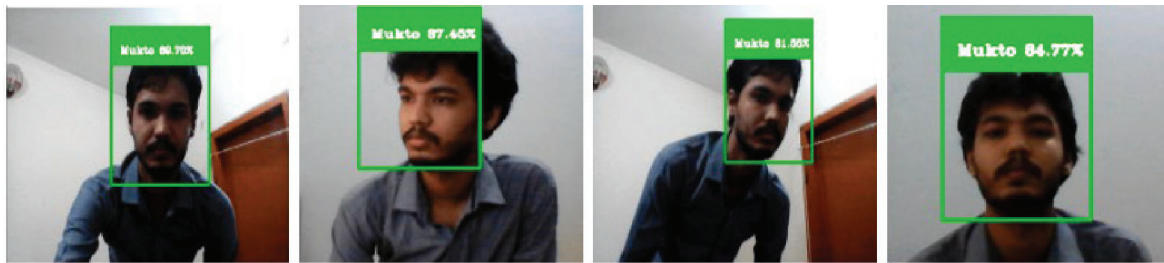
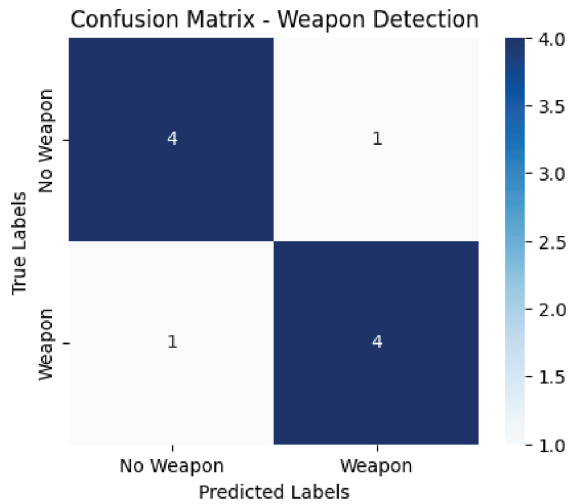


Figure 8: Face recognition.

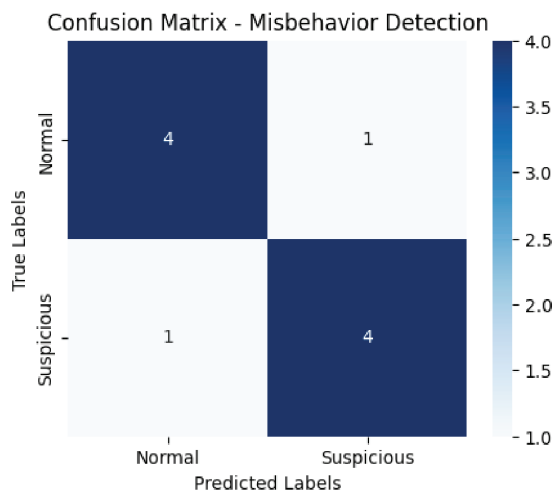
Table 3: Ablation study

| Configuration                    | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|----------------------------------|--------------|---------------|------------|--------------|
| Full CMS (H-CapsNet + C3D + CNN) | 93.5         | 92.8          | 91.7       | 92.2         |
| Without H-CapsNet                | 88.1         | 87.4          | 85.9       | 86.6         |
| Without C3D                      | 89.3         | 88.7          | 86.2       | 87.4         |
| Without CNN (face recognition)   | 90.2         | 89.8          | 88.1       | 88.9         |
| Only H-CapsNet                   | 86.7         | 85.3          | 84.5       | 84.9         |
| Only C3D                         | 84.9         | 83.6          | 82.4       | 83.0         |
| Only CNN                         | 85.5         | 84.8          | 83.3       | 84.0         |

C3D, 3D convolutional neural network; CMS, crime monitoring system; CNN, convolutional neural network; H-CapsNet, hyper capsule network.



(A)



(B)

Figure 9: Confusion matrices for (A) weapon detection and (B) misbehavior detection modules. The diagonal values indicate correct classifications, demonstrating the model's effectiveness in both safety-critical scenarios.

is removed, highlighting the importance of reliable weapon detection in public safety scenarios. The C3D and CNN modules also show notable impacts, confirming that behavior analysis and face recognition are essential for a comprehensive CMS.

## b. Ethical concerns of facial recognition in CMS

While facial recognition enhances the effectiveness of crime monitoring, it also raises important ethical issues that must be addressed:

### b.i. Bias and fairness

Facial recognition systems can exhibit bias against certain demographics, such as people of color, women, or younger and older individuals. This may lead to misidentification or false accusations. To reduce this risk:

- We trained the system on diverse datasets.
- We plan to include fairness checks and ongoing evaluations to monitor bias.

### b.ii. Privacy

Continuous surveillance and face matching may infringe on individual privacy, especially in public spaces. To address this:

- Facial data is processed locally, and sensitive information is not stored permanently.
- The system is intended for use only by authorized law enforcement under regulated conditions.

### b.iii. Potential misuse

There is a risk of the system being misused for unauthorized tracking, profiling, or surveillance without proper oversight. Therefore:

- We recommend strict legal frameworks and transparent governance policies.
- System access should be audited and limited to approve personnel.

## V. Conclusion

The UCF crime dataset, which includes actual surveillance footage of a range of criminal activity and sophisticated DL algorithms, is used in this study to create an efficient and intelligent CMS. The CMS offers a complete solution for real-time crime detection and alerting by combining modules like Deep CNNs for facial recognition, C3D for misbehavior

detection, and H-CapsNet for weapon identification. Further improving model generality and resilience is the use of data augmentation techniques like as rotation, scaling, flipping, and panning. With 93.5% accuracy, 92.8% precision, 91.7% recall, and a 92.2% F1-score, the system performs well and has low rates of FPs and FNs. These findings demonstrate how proactive monitoring and quick reaction systems in the planned CMS might greatly enhance public safety. For large-scale validation, future research can concentrate on growing the dataset, adding more contextual variables, and implementing the system in actual smart city settings.

## References

- [1] H. A. Al-Shammari, A. S. Al-Dahoud and F. A. Hussain, "Real-Time Gun Detection Using Deep Learning and CCTV for Public Safety," *IEEE Access*, vol. 10, 2022.
- [2] Y. Li, K. Xu, and J. Liu, "Human Behavior Analysis Using Deep Learning in Surveillance Systems," *Sensors*, vol. 23, no. 2, 2023.
- [3] Kim, Hyun-Bin, Nakhoon Choi, Hye-Jeong Kwon, and Heeyoul Kim. "Surveillance system for real-time high-precision recognition of criminal faces from wild videos." *IEEE Access* 11,2023.
- [4] Wakhare, Surdi, Kajal Ingale Rutuja Patil, and Vishwadip Kajal Ingale. "Face, Expression and Gesture Recognition and Compilation in Database Using Machine Learning." *International Journal of Scientific Research in Science and Technology*, 2021.
- [5] Raval, Jay, Nilesh Kumar Jadav, Sudeep Tanwar, Giovanni Pau, Fayez Alqahtani, and Amr Tolba. "Criminal emotion detection framework using convolutional neural network for public safety." *Scientific Reports* 15, no. 1, 2025.
- [6] Santos, Tomás, Hélder Oliveira, and António Cunha. "Systematic review on weapon detection in surveillance footage through deep learning." *Computer science review*, Vol. 51,2024.
- [7] Sridevi, Gandla, and G. Kiran Kumar. "Detecting Criminal Activities of Surveillance Videos using Deep Learning." *Journal of Computational Analysis & Applications*, Vol. 33, No. 7,2024.
- [8] Pangavhane, Madhuri, Rahul Patil, Rajesh Bharati, Deepak Gupta, Prashant Ahire, Pramod Patil, Wasudeo Rahane, and Deepak Dharrao. "Real-Time Deep Learning-Driven Surveillance with Spatiotemporal Feature Extraction for Detection of Anomalous Human Behavior Across Dynamic Environments." *International Journal of Safety & Security Engineering*, Vol. 15, no. 1,2025.
- [9] Umoren, Imeh, Saviour Inyang, and Abasiama Silas. "Intelligent Surveillance and Facial Recognition System for Efficient Border Monitoring and Threats Prediction Using Machine Learning Approach." *Researchers Journal of Science and Technology* 1, no. 1,2021.
- [10] Bhatti, Muhammad Tahir, Muhammad Gufran Khan, Masood Aslam, and Muhammad Junaid Fiaz. "Weapon detection in real-time CCTV videos using deep learning." *IEEE Access*, Vol. 9, 2021.
- [11] Mukto, Md Muktadir, Mahamudul Hasan, Md Maiyaz Al Mahmud, Ikramul Haque, Md Ahsan Ahmed, Taskeed Jabid, Md Sawkat Ali, Mohammad Rifat Ahmmad Rashid, Mohammad Manzurul Islam, and Maheen Islam. "Design of a real-time crime monitoring system using deep learning techniques." *Intelligent Systems with Applications*, Vol. 21,2024.
- [12] Panda, Mrutyunjaya, A. Mousa Abd Allah, and Aboul Ella Hassanien. "Developing an efficient feature engineering and machine learning model for detecting IoT-botnet cyber-attacks." *IEEE Access*, Vol. 9, 2021.
- [13] Vuyyuru, Lakshma Reddy, NagaMalleswara Rao Purimetla, Kancharakunt Yakub Reddy, Sai Srinivas Vellela, Sk Khader Basha, and Ramesh Vatambeti. "Advancing automated street crime detection: a drone-based system integrating CNN models and enhanced feature selection techniques." *International Journal of Machine Learning and Cybernetics* 16, no. 2, 2025.
- [14] Akhtar, Muhammad Shoaib, and Tao Feng. "Deep Learning-Based Framework for the Detection of Cyberattacks Using Feature Engineering." *Security and Communication Networks*, no. 1,2021.
- [15] Nazir, Amril, Rohan Mitra, Hana Sulieman, and Firuz Kamalov. "Suspicious behavior detection with temporal feature extraction and time-series classification for shoplifting crime prevention." *Sensors*, Vol. 23, no. 13,2023.
- [16] Sahay, Kishan Bhushan, Bhuvaneswari Balachander, B. Jagadeesh, G. Anand Kumar, Ravi Kumar, and L. Rama Parvathy. "A real-time crime scene intelligent video surveillance system in a violence detection framework using deep learning techniques." *Computers and Electrical Engineering*, Vol. 103,2022.
- [17] Ullah, Waseem, Amin Ullah, Ijaz Ul Haq, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks." *Multimedia tools and applications*, Vol. 80,2021.

- [18] Raza, Ali, Hafeez Ur Rehman Siddiqui, Kashif Munir, Mubarak Almutairi, Furqan Rustam, and Imran Ashraf. "Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction." *Plos one*, Vol. 17, no. 11, 2022.
- [19] Verma, Kamal Kant, Brij Mohan Singh, and Amit Dixit. "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance systems." *International Journal of Information Technology*, Vol. 14, No. 1, 2022
- [20] Palanivinayagam, Ashokkumar, Siva Shankar Gopal, Sweta Bhattacharya, Noble Anumbe, Ebuka Ibeke, and Cresantus Biamba. "An optimized machine learning and big data approach to crime detection." *Wireless Communications and Mobile Computing* 2021, no. 1, 2021
- [21] Alzahrani, Mohammed Y., and Alwi M. Bamhdi. "Hybrid deep-learning model to detect bot-net attacks over internet of things environments." *Soft Computing*, Vol. 26, No. 16, 2022.
- [22] Zahrawi, Mohammad, and Khaled Shaalan. "Improving video surveillance systems in banks using deep learning techniques." *Scientific Reports*, Vol. 13, No. 1, 2023.
- [23] Zhang, Xinwei, Yaoci Han, Wei Xu, and Qili Wang. "HOBAs: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture." *Information Sciences*, Vol. 557, 2021.
- [24] Akhtar, Muhammad Shoaib, and Tao Feng. "Detection of malware by deep learning as CNN-LSTM machine learning techniques in real time." *Symmetry* Vol. 14, no. 11, 2022.
- [25] Chinedu, Paschal Uchenna, Wilson Nwankwo, Florence U. Masajuwa, and Simon Imoisi. "Cybercrime detection and prevention efforts in the last decade: an overview of the possibilities of machine learning models." *Rigeo*, Vol. 11, No. 7, 2021.
- [26] Awotunde, Joseph Bamidele, and Sanjay Misra. "Feature extraction and artificial intelligence-based intrusion detection model for a secure internet of things networks." In *Illumination of artificial intelligence in cybersecurity and forensics*, Cham: Springer International Publishing, 2022.
- [27] Wei, Yuanyuan, Julian Jang-Jaccard, Fariza Sabrina, Amardeep Singh, Wen Xu, and Seyit Camtepe. "AMLP: A hybrid deep learning approach for DDoS detection and classification." *IEEE Access* 9, 2021.
- [28] Tutar, Hayati, Ali Güneş, Metin Zontul, and Zafer Aslan. "A hybrid approach to improve the video anomaly detection performance of pixel-and frame-based techniques using machine learning algorithms." *Computation* 12, no. 2, 2024.
- [29] Ahmim, Ahmed, Faiz Maazouzi, Marwa Ahmim, Sarra Namane, and Imed Ben Dhaou. "Distributed denial of service attack detection for the Internet of Things using a hybrid deep learning model." *IEEE Access*, Vol. 11, 2023.
- [30] Hayder, Israa M., Taief Alaa Al-Amiedy, Wad Ghaban, Faisal Saeed, Maged Nasser, Ghazwan Abdulnabi Al-Ali, and Hussain A. Younis. "An intelligent early flood forecasting and prediction leveraging machine and deep learning algorithms with an advanced alert system." *Processes*, Vol. 11, no. 2, 2023.
- [31] Kuklin, V. Zh., Ivanov, N. Z., Muranov, A. N., Alexandrov, I. A., & Linskaya, E. Yu. (2024). Improving the reliability of biometric authentication processes using a model for reducing data drift. *Emerging Science Journal*, 8(6), 2449–2459. Available at: <http://www.ijournalse.org>
- [32] Kurdthongmee, W., & Kurdthongmee, P. (2024). Fast and accurate pupil estimation through semantic segmentation fine-tuning on a shallow convolutional backbone. *HighTech and Innovation Journal*, 5(2), Article 016. DOI: 10.28991/HIJ-2024-05-02-016
- [33] Abumalloh, R. A., Alrahhal, M., El-Haggar, N., Alsumayt, A., Alfawaer, Z. M., & Aljameel, S. S. (2024). Exploring individuals' experiences with security attacks: A text mining and qualitative study. *Emerging Science Journal*, 8(1). DOI: 10.28991/ESJ-2024-08-01-010