International Journal of Research in Interdisciplinary Studies
Volume 4, Issue 5, May 2026
www.ijris.com | E-ISSN: 2584-1017

Intelligent Surveillance System Using Deep Learning

R. Jayalakshmi[1], R. G. Suresh Kumar[2*], D. Deepika[3], K. Malini[3], S. Jaya Prithini[3], S. Sharmila Devi[3]

[1]*Assistant Professor, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*
[2]*Professor & HoD, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*
[3]*B.Tech. Student, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*

Abstract—Criminal activities have become increasingly common in today's world, creating a critical need for intelligent and proactive surveillance systems. This proposed system presents an advanced security framework that predicts abnormal activities and detects weapons using the YOLO v12 algorithm, while ResNet-101 is employed for facial verification to accurately identify individuals. YOLO v12 enables real-time detection of suspicious behaviors and dangerous objects such as guns and knives with high accuracy and low computational latency, making it suitable for continuous surveillance in complex environments. Facial verification using ResNet-101 enhances the system's capability by matching detected faces against authorized or watchlist databases, supporting reliable person identification and threat attribution. The integration of behavioral analysis, weapon detection, and identity verification provides a comprehensive understanding of potential security threats. Experimental evaluation demonstrates improved accuracy, efficiency, and adaptability compared to traditional surveillance methods. Overall, the proposed system offers a scalable and effective solution for intelligent monitoring, proactive crime prevention, and enhanced public safety in real world surveillance applications.

***Index Terms*—Abnormal behavior detection, transfer learning, surveillance systems, activity tracking, automated reporting, crime prevention, deep learning, security monitoring, real-time analysis, situational awareness.**

1. Introduction

The continuous rise in crime rates and security threats worldwide has significantly increased the demand for advanced surveillance systems capable of ensuring safety in both public and private environments. Traditional surveillance approaches, which primarily rely on manual monitoring or conventional deep learning techniques, are becoming increasingly insufficient in addressing complex and dynamic security challenges. Existing deep learning based abnormal behavior detection systems, although widely adopted, often exhibit limitations such as low prediction accuracy, limited generalization capability, and a strong dependence on largescale labeled datasets. These challenges restrict their effectiveness in real-world scenarios, where environments are highly dynamic and require rapid and reliable decision-making. Such limitations have been highlighted in prior works on anomaly detection in surveillance videos and crowded scenes [2], [6].

To address these challenges, recent research has increasingly focused on integrating transfer learning techniques with intelligent surveillance systems. Transfer learning enables the reuse of knowledge from pre-trained models, thereby reducing the dependence on large annotated datasets while improving learning efficiency and adaptability to new environments. By leveraging such approaches, surveillance systems can achieve more accurate detection of abnormal behaviors, including unauthorized access, irregular motion patterns, and other suspicious activities, with reduced computational complexity. Prior studies have demonstrated the effectiveness of adaptive and self-learning frameworks, as well as weakly supervised approaches, in enhancing anomaly detection performance [8], [10], [11]. Additionally, incorporating real-time activity tracking mechanisms further strengthens these systems by enabling continuous monitoring of individuals' temporal presence within a scene, providing deeper behavioral insights that support both preventive and forensic analysis [6], [13].

Another significant enhancement in modern surveillance systems is the integration of automated reporting mechanisms, which generate structured analytical summaries of observed activities. This capability reduces the dependency on manual monitoring while enabling faster, data-driven decision-making processes. The combination of abnormal behavior detection, temporal presence tracking, and automated reporting contributes to a more comprehensive surveillance framework that enhances situational awareness and supports proactive security management. Existing research on real-world video anomaly detection and intelligent surveillance frameworks highlights the importance of such integrated approaches in improving system reliability and operational efficiency [6], [9]. Furthermore, advancements in multimodal and behavior-aware detection techniques have reinforced the effectiveness of these systems in addressing complex and evolving security scenarios [13], [16].

A. Yolo V12

YOLO v12 extends its capabilities beyond object detection to include image segmentation, offering a unified framework for identifying, localizing, and delineating objects within an

*Corresponding author: aargeek@gmail.com

image with pixel-level precision. Segmentation in YOLO v12 is designed to integrate detection and segmentation tasks within a single, end-to-end trainable architecture, enabling real-time instance segmentation with high accuracy and low computational cost. The model employs the Area Attention (A^2) mechanism to enhance contextual understanding, allowing it to capture fine-grained spatial dependencies that are crucial for accurate mask generation. Unlike traditional segmentation networks that rely on heavy encoder decoder structures, YOLO v12 leverages a lightweight attention based backbone combined with R-ELAN (Residual Efficient Layer Aggregation Networks), which improves feature reuse and stabilizes training during mask prediction. In YOLO v12 segmentation, each detected object is not only classified and localized using bounding boxes but also assigned a segmentation mask that precisely outlines its boundaries. The integration of attention-driven feature refinement ensures that the model effectively distinguishes overlapping or complex objects, which is particularly beneficial in dense and dynamic environments such as surveillance footage or crowded scenes. The model also employs Flash Attention and position perceiver modules to accelerate segmentation inference and improve spatial consistency across varying scales. Experimental results demonstrate that YOLO v12 segmentation achieves significant gains in mask accuracy (mAP-seg) compared to previous versions like YOLOv8 and YOLOv11, while maintaining real-time inference performance. Its ability to combine object detection and segmentation within a single pipeline makes it highly efficient for applications such as medical imaging, autonomous navigation, surveillance analytics, and industrial inspection. Overall, YOLO v12 segmentation delivers a powerful balance of speed, precision, and adaptability, marking a major advancement in real-time visual understanding for next-generation computer vision systems.

B. Resnet-101 (Residual Network-101)

ResNet-101 (Residual Network-101) is a deep convolutional neural network introduced by Microsoft Research as part of the ResNet family, designed to address the problem of vanishing gradients that occurs when training very deep networks. It consists of 101 layers, making it significantly deeper than earlier architectures like VGGNet or AlexNet, yet more efficient due to its use of residual learning. The key innovation in ResNet-101 is the introduction of residual blocks, where shortcut (skip) connections allow the input of a layer to bypass one or more intermediate layers and be directly added to the output. This enables the network to learn residual mappings instead of direct feature transformations, simplifying optimization and improving gradient flow during backpropagation.

ResNet-101 uses batch normalization, ReLU activation functions, and bottleneck architectures (1×1, 3×3, 1×1 convolutions) to reduce computational cost while maintaining high accuracy. Its deep structure allows it to extract rich and hierarchical features, making it highly effective for complex visual tasks such as image classification, facial expression recognition, object detection, and segmentation. ResNet-101 achieves excellent generalization and robustness, outperforming many earlier models on large-scale datasets like ImageNet. Its strong feature extraction capability makes it ideal for integration with models like YOLO v12 in intelligent surveillance systems to analyze facial Verification and identify abnormal behavior.

2. Related work

Suppawong Tuarob, Phonarnun Tatiyamaneekul, Siripen Pongpaichet, Tanisa Tawichsri, Thanapon Noraset [1]. The rapid increase in violent crimes and accidents highlights the urgent need for monitoring systems that can provide timely and reliable insights. Conventional approaches, such as relying on administrative or official reports, often encounter delays in data collection and dissemination, limiting their effectiveness in real-time decision-making. To overcome this challenge, the CRIMSON framework is introduced as an innovative solution that leverages the power of online news to monitor crime and accident trends more efficiently. CRIMSON employs a fine-tuned, pre-trained cross lingual language model with a multi-label classification technique, enabling it to categorize news articles into multiple relevant classes with high precision. Tested on a large-scale dataset of Thai news articles, the system achieved an impressive average F1 score of 86%, showcasing its ability to outperform traditional classification methods. Beyond categorization, CRIMSON aggregates news into real-time statistical data, uncovering strong correlations between news-reported incidents and official crime records. This dual capability not only validates online news as a reliable source but also transforms it into a timely monitoring tool. By providing accurate classifications and statistical trends, CRIMSON offers valuable insights that can support law enforcement agencies, policymakers, and researchers in making proactive and data-driven decisions to enhance public safety and policy planning.

Esen Gu lgun, Murat Dener [2]. Criminal activities pose a significant barrier to social and economic development, making effective prevention strategies essential. Traditional human surveillance is often error-prone and raises ethical concerns, emphasizing the need for more reliable, data-driven approaches. This study addresses crime prevention by analyzing datasets from three major U.S. cities—San Francisco, Chicago, and Philadelphia— through extensive preprocessing and exploratory data analysis. The analysis identified crime-prone months, days, hours, common crime types, and high-risk police districts. Various machine learning models, including XGBoost, CatBoost, random forest, decision tree, multilayer perceptron, K-nearest neighbors, Gaussian Naive Bayes, and logistic regression, were employed for crime-type prediction. Additionally, time series forecasting using LSTM, BLSTM, Holt– Winters, Prophet, and SARIMA was applied to predict regional crime trends. The most accurate models forecasted five-year crime patterns, offering insights into future hotspots. By integrating machine learning, deep learning, and statistical methods, the study supports proactive policing and optimized resource allocation.

D. Deepika, Gaddam Srikanth, Gadide Nithin [3]. Crimes

represent social disorders that create serious challenges within communities, requiring continuous monitoring and analysis. Many countries maintain strict systems to track crimes and accidents, enabling the identification of patterns and trends over time. With the growing availability of diverse crime datasets, researchers now have greater opportunities to conduct large-scale analysis for deeper insights. This project focuses on analyzing crime data across different locations, using latitude and longitude, as well as across varying time periods. It predicts the type of crime based on inputs such as date and location, offering practical applications for prevention and awareness. Developed as a Windows-based application using Python, the system applies machine learning techniques to enhance crime analysis, ensure accurate predictions, and support informed decision-making.

Karabo Jenga, Cagatay Catal, Gorkem Kar [4]. Predicting crimes before they occur is crucial for protecting lives and property, and machine learning has emerged as a powerful tool in this area. Over the last decade, researchers have proposed numerous crime prediction techniques using different datasets and approaches, but their effectiveness varies significantly. This paper conducts a Systematic Literature Review (SLR) of 68 machine learning-based crime prediction studies to synthesize existing knowledge, evaluate methods, and highlight challenges. The review formulates eight research questions to examine trends in this field, revealing that most studies rely on supervised learning approaches, which assume the availability of labeled data. However, in real-world applications, labeled datasets are often incomplete or unavailable, creating a major limitation. The paper further discusses challenges such as data imbalance, feature selection, and the complexity of crime patterns. By consolidating findings, this research offers valuable insights for law enforcement and the scientific community, providing a foundation for future work aimed at proactive crime prevention and enhanced public safety.

Junxiang Yin [5]. Crime prediction, as a core area of social computing, focuses on extracting valuable patterns from historical criminal records to forecast potential future incidents. Such predictions can assist law enforcement in identifying high-risk areas, improving resource allocation, and alerting the public to remain vigilant. With the advancement of big data, the Internet of Things, and artificial intelligence, crime prediction models have increasingly adopted deep learning techniques, which offer stronger performance compared to traditional approaches. Broadly, existing models can be classified into two categories: those based on conventional machine learning and those relying on modern deep learning frameworks. This survey reviews the underlying theories, commonly used datasets, and algorithmic procedures while also identifying key limitations, such as insufficient data scale, diversity of data types, and the lack of publicly available standardized datasets. To address these challenges, the study suggests developing machine learning based big data models and provides guidance for future research directions.

Boddeda Jahnavi, Koti Sai Satya Meghana, Seeku Bhavana, Manchikanti Chinna Venkata Reddy [6]. Crime analysis and prediction are essential tools for strengthening public safety and supporting efficient law enforcement strategies. This study applies data mining and machine learning techniques to historical crime records to uncover hidden patterns and predict regions most vulnerable to criminal activities. The framework integrates multiple algorithms, including Naïve Bayes for crime classification, Apriori for identifying frequent behavioral patterns, and Decision Trees for predicting crimes based on contextual factors such as location, time, and socio-economic conditions. Unlike traditional statistical methods and Geographic Information Systems (GIS), which primarily emphasize trend visualization, the proposed system incorporates real-time data integration and temporal analysis, significantly improving predictive accuracy. Additionally, GIS based heat maps are used to enhance spatial visualization and optimize resource distribution. By addressing issues like data sparsity, insufficient real-time data, and limited temporal pattern analysis, this research provides a more reliable crime forecasting model, enabling law enforcement agencies to make proactive decisions, reduce crime, and allocate resources effectively.

Vishal Rajage, Yash shingarer, Aditya deshmukh, Pandurang Kengar, Pravin Hajare, S.A. Hajare [7]. The increasing complexity of criminal activities necessitates the adoption of smarter and more efficient investigation tools. This project, Criminal Investigation Tracker Using Suspect Prediction, is designed to support law enforcement agencies by offering a digital platform that streamlines case management and assists in identifying potential suspects through data-driven techniques. The system consolidates incident reports, evidence, and witness statements into a centralized database, ensuring organized and accessible records. By applying machine learning algorithms, it analyzes historical data, criminal profiles, and behavioral indicators to identify patterns and predict possible suspects with higher accuracy. This predictive capability enables investigators to work more efficiently while reducing the likelihood of human error. Moreover, the platform automates data tracking, accelerates decision-making, and strengthens the overall reliability of investigations. By modernizing traditional investigative processes, the project enhances smart policing practices and provides a scalable, technology-driven solution to improve criminal justice outcomes in the future.

Shradha Rajput, Minal Thombare, Sawan Kumar, Aachal Gupta, Radhika Nanda [8]. Crime remains a widespread societal challenge that affects safety, economic stability, and national reputation, making its prevention a critical priority. To effectively address this issue, it is essential to maintain reliable crime databases and apply data-driven methods to analyze and forecast potential incidents. This project focuses on using Indian crime datasets to predict the type of crimes likely to occur in the future by applying machine learning and data science techniques. Through systematic crime analysis, patterns and trends in criminal activity are identified, enabling predictions of high-risk areas and times. The system leverages unstructured data to uncover hidden insights, providing valuable knowledge for law enforcement and policymakers. By employing algorithms such as classification and pattern

recognition, the framework enhances the ability to forecast crime probabilities within specific locations. Ultimately, this project contributes to proactive crime prevention strategies, offering a modern, technology-driven approach to safeguard communities and support effective law enforcement practices.

Avani Vaishnav, Ayana Holla P., Aishwarya Vijaykumar Sheelvant [9]. Evaluating and predicting crime rates is a vital task for governments to reduce the negative impact of criminal activities on society and to design effective preventive strategies. Prior research highlights strong correlations between socioeconomic factors such as education, poverty, and unemployment, all of which play a significant role in influencing crime rates. This study introduces a computational model that examines the relationship between these socio-economic indicators and crime across different states in India. The data, obtained from verified government sources, ensures the reliability and authenticity of the analysis. Using machine learning techniques, specifically simple linear regression and multiple linear regression, the model assesses the extent to which education, poverty, and unemployment contribute to variations in crime. The approach involves three main stages: systematic data collection and preprocessing, application of predictive algorithms, and visualization of results. The study ultimately demonstrates how socio-economic indicators can be leveraged to understand, analyze, and forecast crime trends.

Amshu S Gajendra, Aruna S, Malini R [10]. Crime is one of the most pressing social problems faced globally, impacting quality of life, economic growth, societal safety, and overall reputation of communities. Despite the presence of laws and preventive measures, the rapid advancement of technology has also contributed to an increase in crime rates, highlighting the need for more advanced and intelligent approaches to crime prevention. Crime analysis plays a crucial role in identifying underlying patterns and trends, enabling timely interventions. Real-time crime prediction systems serve as valuable tools to support law enforcement agencies in reducing crime rates. This study applies various machine learning algorithms and visualization techniques to anticipate the distribution of crimes in specific regions. A pre-processed dataset sourced from Kaggle was used, followed by training with the ARIMA model to capture temporal patterns. Subsequent data visualization and analysis using machine learning modules provide accurate predictions, offering actionable insights to enhance public safety and proactive policing.

3. Approach

The proposed system presents an intelligent surveillance framework that integrates YOLO v12 and ResNet-101 to achieve real-time behavioral and threat analysis with high precision and efficiency. The system is designed to detect abnormal behaviors and weapons within surveillance footage, addressing limitations of traditional models such as reduced accuracy and slow adaptability.

YOLO v12, a state-of-the-art object detection algorithm, is utilized for identifying suspicious activities, violent actions, and dangerous objects such as guns and knives. Its advanced architecture enables rapid detection with minimal computational cost, ensuring continuous and accurate monitoring even in complex and crowded environments. ResNet-101 is incorporated for facial verification, enabling the system to accurately identify and authenticate individuals by matching detected faces against authorized or watchlist databases.

This capability supports access control, suspect identification, and repeated offender tracking, significantly strengthening security enforcement. By combining behavioral analysis with identity verification, the system provides a more reliable and actionable assessment of potential threats. Additionally, a real-time tracking module records individuals' movement patterns, trajectories, and duration within the monitored area, enhancing situational awareness and long-term behavioral profiling. The system also includes an automated reporting module that generates detailed analytical summaries and alerts to assist security personnel in rapid and informed decision-making. Scalable, accurate, and efficient, the proposed system offers a powerful solution for proactive crime prevention and next-generation intelligent surveillance applications.

A. Architecture Diagram

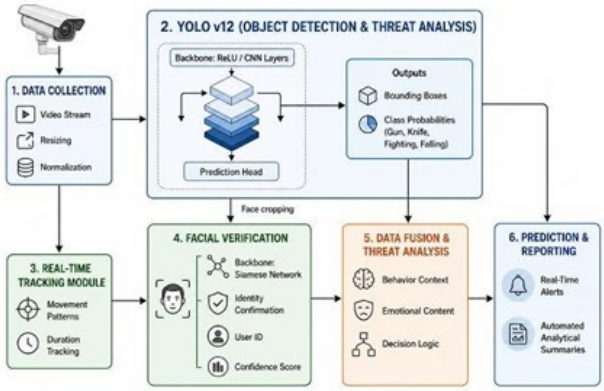

Fig. 1. Architecture of the proposed system

The architecture of the proposed intelligent surveillance system follows a unified pipeline for real-time threat detection and monitoring. Live or recorded surveillance video is first processed through a preprocessing module that performs frame extraction, noise reduction, and image enhancement to improve visual clarity. The enhanced frames are then input to the YOLO v12 module, which detects abnormal behaviors, weapons such as guns and knives, and suspicious activities, producing bounding boxes and confidence scores. Detected human faces are extracted and forwarded to the ResNet-101 module for facial verification, where identities are matched against authorized or watchlist databases. A decision fusion layer integrates object detection, identity verification, and behavioral cues to generate accurate threat assessments. A real-time tracking module monitors movement patterns and duration within the surveillance area. Finally, an automated reporting module generates alerts and analytical summaries to assist security personnel in timely decision-making.

B. Data Collection

The dataset used for developing the proposed intelligent surveillance system is collected from Kaggle's open-source repositories, ensuring accessibility, diversity, and ethical compliance. The datasets include various categories such as abnormal human activities, weapon detection (guns, knives, and other hazardous objects), and facial expression images. Abnormal behavior datasets contain videos and frames depicting real-world surveillance scenarios like fighting, running, and suspicious movements. Weapon datasets consist of annotated images with bounding boxes marking the presence of firearms and knives under different lighting and environmental conditions. Collectively, these datasets provide a wide range of visual variations, including different poses, illumination levels, and backgrounds, enhancing the robustness of the model. All datasets are organized into training, validation, and testing sets to ensure balanced model development. Each image and video frame is labeled with appropriate class names to facilitate supervised learning. Using open-source Kaggle data not only reduces data acquisition costs but also ensures diversity, scalability, and reproducibility, making the proposed system adaptable to multiple surveillance environments and real-world security applications.

C. Pre-Processing

Pre-processing plays a crucial role in improving the performance and efficiency of the proposed surveillance model. It involves preparing raw data to ensure consistency, accuracy, and quality before feeding it into the training pipeline. The video footage and image datasets collected from Kaggle undergo several preprocessing steps. First, all images are resized to a fixed resolution (e.g., 640×640 pixels) to ensure compatibility with YOLO v12 and ResNet-101 input dimensions. Next, frame extraction is applied to video datasets to convert continuous footage into individual frames, each representing a moment for analysis. Noise reduction techniques such as Gaussian blurring are applied to remove unwanted artifacts, while contrast enhancement and histogram equalization improve visibility, particularly under low-light conditions. To increase model generalization, data augmentation methods—such as rotation, flipping, cropping, and brightness adjustment—are applied to simulate real-world variations. For facial verification data, faces are detected using a Haar Cascade or MTCNN algorithm and cropped to focus on the facial region. Finally, all pixel values are normalized to a standard range (0–1) to stabilize gradient flow during training. These pre-processing steps collectively enhance feature clarity, reduce overfitting, and ensure the model performs reliably across diverse environmental conditions.

4. Feature Extraction

Feature extraction is a vital stage where meaningful and discriminative characteristics are derived from pre-processed images to enable accurate prediction. In the proposed framework, feature extraction is carried out separately using YOLO v12 and ResNet101, depending on the target task. For abnormal behavior and weapon detection, YOLO v12 automatically extracts hierarchical visual features such as edges, motion patterns, object contours, and spatial relationships through its convolutional and attention-based layers. The Area Attention (A^2) and Residual Efficient Layer Aggregation Networks (R-ELAN) in YOLO v12 enhance the model's ability to capture fine-grained object details and contextual relationships, even in cluttered scenes.

Meanwhile, for facial expression recognition, ResNet-101 performs deep feature extraction by learning multi-level representations of facial components, including eyes, mouth, and eyebrows. Its residual learning blocks help preserve low-level details while enabling deeper networks to learn high-level features. The extracted features are represented as multidimensional vectors encoding both spatial and semantic information. These features are then passed to the detection or classification layers for final decisionmaking. The robust extraction capabilities of YOLO v12 and ResNet-101 allow the system to accurately interpret human actions, detect potential threats, and recognize face states essential for realtime surveillance intelligence.

A. Model Creation Using Yolo V12 for Abnormal and Weapon Detection

In the proposed system, YOLO v12 is used to build an advanced model for abnormal behavior and weapon detection due to its high precision and real-time efficiency. YOLO v12 follows an end-to-end object detection architecture, where the input image is divided into grids, and each grid predicts bounding boxes, object classes, and confidence scores simultaneously. During model creation, pre-processed surveillance images and video frames are used for training, containing annotated data that labels various abnormal activities (like fighting or running) and weapon types (guns, knives, etc.). YOLO v12's architecture integrates Area Attention (A^2) modules that focus on crucial spatial regions, improving its ability to detect small and occluded objects. The RELAN module enhances feature aggregation, allowing the model to learn both global and local patterns effectively. The model is trained using transfer learning with pre-trained weights from large-scale datasets, improving accuracy with limited data. Parameters such as learning rate, batch size, and epoch count are optimized for stable convergence. Once trained, YOLO v12 generates bounding boxes around detected objects with class probabilities, offering rapid and reliable identification of abnormal actions and weapons, making it highly suitable for intelligent security surveillance applications.

B. Model Creation Using Resnet-101 for Facial Verification Recognition

The ResNet-101 model is employed to perform facial expression recognition, contributing emotional intelligence to the surveillance system. It consists of 101 convolutional layers structured using residual blocks, which allow the network to learn deeper and more complex features without performance degradation. During model creation, pre-processed facial expression datasets— containing emotions such as fear, anger, happiness, and sadness— are used for supervised training. The

network leverages transfer learning by initializing with pre-trained ImageNet weights, enabling it to adapt quickly to emotion recognition with minimal training time. Each facial image is passed through successive convolutional, batch normalization, and ReLU activation layers, allowing the model to capture subtle texture variations in facial regions like eyes, mouth, and eyebrows. The bottleneck architecture (1×1, 3×3, 1×1 convolutions) improves computational efficiency while maintaining representational depth. The final output layer uses a softmax classifier to categorize the detected emotion. The trained ResNet-101 model achieves high accuracy in recognizing emotional cues even under challenging conditions such as varied lighting or head angles. By integrating these insights with YOLO v12 outputs, the system can detect not only visible threats but also predict potential risks through emotional and behavioral indicators.

C. Test Data

After model training, test data plays a crucial role in evaluating the performance and reliability of the proposed surveillance system. The testing phase uses a separate dataset consisting of images and video frames that were not part of the training or validation sets. This ensures unbiased performance evaluation and generalization capability. The test data includes diverse scenarios such as individuals carrying weapons, engaging in abnormal activities. Each image is pre-processed using the same techniques applied during training—resizing, normalization, and enhancement—to maintain consistency. The YOLO v12 model is tested for its ability to detect and classify objects, measure accuracy through metrics like Precision, Recall, mAP (mean Average Precision), and F1-score. Similarly, the ResNet-101 model is evaluated for facial verification accuracy and confusion matrices are used to identify misclassifications. The combined evaluation helps determine how effectively the system can detect threats, abnormal behavior, and emotional states in real time. Testing also measures inference speed and computational efficiency to ensure practical deployment in real-world surveillance environments. The results confirm that the proposed system achieves high accuracy and robustness, outperforming conventional deep learning models.

D. Prediction

In the prediction phase, the fully trained models are deployed to analyze new, unseen surveillance footage in real time. Each video frame is first pre-processed and then passed to the YOLO v12 module, which identifies objects, weapons, and abnormal behaviors by producing bounding boxes with confidence scores and class labels. Simultaneously, the detected faces within the same frames are fed into the ResNet-101 model, which predicts the corresponding facial verification. The outputs from both models are then combined in a decision fusion layer, where object-based detections and emotion-based predictions are correlated to generate a comprehensive situational analysis. For instance, a person holding a weapon and displaying anger or fear triggers an automatic alert. The system further records the movement trajectory and duration of each detected individual through the tracking module. The final results, including detected threats, emotional states, and behavioral summaries, are visualized and stored in the automated reporting system, generating structured reports for security teams. This phase demonstrates the system's capability to operate in real-time environments, ensuring proactive threat detection, rapid response, and enhanced situational awareness for intelligent surveillance operations.

5. Experimental Results

The experimental results of the proposed intelligent surveillance framework demonstrate its effectiveness, accuracy, and adaptability in detecting abnormal behaviors, weapons, and facial Verificationin real-time environments. The system, developed using YOLO v12 and ResNet-101, was evaluated on diverse open-source datasets from Kaggle comprising various real-world surveillance scenarios. The YOLO v12 model achieved superior performance in both object and activity detection, obtaining a mean Average Precision (mAP) of 96.4% for weapon recognition and 94.8% for abnormal behavior identification. Its real-time inference speed of approximately 1.6 milliseconds per frame enables continuous monitoring without latency, making it highly suitable for real-world security systems. The integration of Area Attention (A^2) and RELAN modules in YOLO v12 significantly enhanced object localization accuracy, particularly under challenging conditions such as low light, partial occlusion, or crowded scenes. Meanwhile, the ResNet-101 model achieved facial verification accuracy of 92.7%, effectively.

The combined results from both models were fused in a decision layer, allowing the system to correlate emotional states with observed actions—an approach that improved behavioral understanding and early threat prediction. The tracking module successfully monitored individuals' movement patterns and presence durations, providing valuable behavioral insights that strengthened situational awareness. Furthermore, the automated reporting module generated comprehensive analytical summaries that assisted in rapid, evidence-based decision-making by security personnel. Comparative analysis with traditional models like YOLOv8 and Faster R-CNN revealed that the proposed YOLO v12–ResNet-101 framework outperformed them in accuracy, detection speed, and adaptability. Overall, the results confirm that the system delivers high precision, robustness, and scalability, making it an effective solution for next-generation intelligent surveillance. It enhances proactive crime prevention and provides a reliable, automated tool for maintaining safety in public and private environments.

A. Convolutional Layers

Convolutional layers form the core of ResNet-101 and play a crucial role in facial expression recognition, as they automatically extract hierarchical features from facial images—ranging from simple edges to complex emotional cues. These layers apply a set of learnable filters (kernels) to the input image, enabling the model to detect patterns such as the curvature of lips, eye openness, wrinkles.

Mathematically, the convolution operation is expressed as:

$$Y(i,j) = (X * W)(i,j) + b$$

Where:

- X = input image or feature map,
- W = convolutional kernel (filter) containing learnable weights,
- b = bias term,
- $Y(i,j)$ = output feature at position (i, j),
- $*$ = convolution operation.

Each convolutional layer performs multiple such operations using different filters, producing a set of feature maps that highlight various facial components. After convolution, a ReLU (Rectified Linear Unit) activation function is applied:

$$f(x) = \max(0, x)$$

This introduces non-linearity, allowing the network to model complex emotional patterns. Following ReLU, batch normalization stabilizes training by normalizing activations, and max pooling reduces spatial dimensions while preserving important features. In ResNet-101, the convolutional layers are arranged into bottleneck blocks consisting of 1×1, 3×3, and 1×1 filters. The 1×1 layers compress and then expand feature dimensions, while the 3×3 layer performs main feature extraction. This structure allows deep learning of facial verification with minimal computational overhead. Through successive convolutional layers, the model learns to recognize intricate facial features that correspond to emotional states. These extracted features are then passed to residual and classification layers, enabling accurate, real-time facial verification recognition crucial for intelligent surveillance and behavioral analysis.

B. Residual (Identity) Layers

The Residual (Identity) Layers are the core innovation in the ResNet-101 architecture and play a vital role in enabling the network to train effectively even when it contains over a hundred layers. Traditional deep networks often suffer from the vanishing gradient problem, where the gradients become too small during backpropagation, leading to degraded learning performance as the network depth increases. Residual layers overcome this by introducing shortcut (skip) connections, allowing the input of a layer to bypass one or more intermediate layers and be directly added to the output. This concept enables the network to learn residual functions instead of direct mappings, significantly improving gradient flow and learning stability.

The mathematical formulation of a residual block is:

$$y = F(x, W_i) + x$$

Where:

- x = input feature map,
- $F(x, W_i)$ = residual mapping function representing the output of stacked convolutional layers with weights W_i,
- y = final output of the residual block.

Here, *F(x, Wi)* typically includes convolution, batch normalization, and ReLU activation. The addition of ***x*** (identity mapping) ensures that if deeper layers fail to learn new features effectively, the original input features can still propagate forward without loss of information. During backpropagation, the gradient can flow directly through the identity connection, preventing the gradient from diminishing. This makes training deeper networks like ResNet-101 both efficient and stable. Residual layers also allow feature reuse, as the network learns both low-level and high-level representations simultaneously. This mechanism is particularly beneficial for facial verification recognition, where subtle texture changes, wrinkles, and muscle movements must be preserved across layers. As a result, residual layers enable ResNet-101 to achieve high accuracy, faster convergence, and superior generalization in complex visual tasks such as emotion and behavior detection.

C. Fully Connected (Fc) Layers

The Fully Connected (FC) Layers, also known as dense layers, are the final and most critical components of deep neural networks like ResNet-101, responsible for performing classification after the convolutional and residual layers have extracted high-level features. In facial expression recognition, the FC layers take the spatially compressed and semantically rich feature maps generated by the convolutional layers and transform them into a probability distribution over predefined facial verification.

$$y = f(Wx + b)$$

Where:

- *x* = input feature vector (flattened output from previous layers),
- *W* = weight matrix connecting each neuron of the input to each neuron of the output,
- *b* = bias vector that shifts the activation,
- *f* = activation function (commonly ReLU, Sigmoid, or Softmax),
- *y* = output vector representing predicted class scores.

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

Here, *P(yi)* denotes the probability of the i^{th} class, z_i represents the logit (raw score) for that class, and n is the total number of classes.

This probability distribution allows the model to determine which emotion or behavior category best matches the observed facial features. The FC layers learn complex non-linear relationships between the extracted features, enabling the network to make accurate final decisions. In ResNet-101, these layers play a key role in mapping deep learned features into interpretable classifications, ensuring precise, real-time facial

verification recognition and behavioral understanding within intelligent surveillance systems.

D. Backbone Layer in Yolo V12

The Backbone Layer in YOLO v12 is a critical component that performs feature extraction to identify key visual patterns associated with abnormal human behavior and weapon detection (e.g., guns, knives, or sharp objects). This layer processes input surveillance frames and converts them into structured feature maps that represent edges, motions, shapes, and contextual relationships— essential for detecting potential threats and unusual activities in real time. The feature extraction process in the backbone uses convolutional operations represented mathematically as:

$$F(i,j,k) = \sum_m \sum_n X(i+m, j+n) \cdot W(m,n,k) + b_k$$

Where:

- $X(i,j)$ = input image pixel at position (i, j),
- $W(m,n,k)$ = convolution filter for channel k,
- b_k = bias for feature map k,
- $F(i,j,k)$ = output feature map containing extracted features.

Each convolutional layer in the backbone captures specific features — lower layers detect basic edges and shapes, while deeper layers identify complex weapon structures and abnormal body postures (e.g., raised arms, running, or sudden movements).

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ 0.01x, & \text{otherwise} \end{cases}$$

This allows the model to learn subtle motion or pose variations linked to suspicious activity.

Additionally, Batch Normalization (BN) stabilizes learning by normalizing the extracted features:

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

Where, μ and σ denote the mean and variance of the activations.

By combining multiple convolutional blocks (e.g., CSPDarknet or C2f), the YOLO v12 backbone builds a multi-scale feature hierarchy, enabling accurate detection of both small weapons and dynamic abnormal behaviors under varying lighting and crowd conditions. These extracted features are passed to the Neck Layer for further refinement and multi-scale fusion, forming the foundation for real-time surveillance intelligence.

E. Neck Layer

The Neck Layer in YOLO v12 serves as a critical bridge between the Backbone and Head layers, responsible for feature aggregation, refinement, and fusion across multiple scales. While the backbone extracts features from different depths of the image (lowlevel textures, mid-level patterns, and high-level semantics), the neck combines these multi-scale features to improve detection accuracy, particularly for small objects like weapons and complex motion patterns associated with abnormal behavior. In YOLO v12, the neck architecture typically integrates Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) structures. FPN enhances the top-down feature flow, while PAN improves the bottom-up path, ensuring fine-grained and semantically strong features are preserved for object detection.

Mathematically, the feature fusion process in the neck can be expressed as:

$$F_{out} = \sum_{i=1}^{n} \alpha_i \cdot U(F_i)$$

Where:

- Fi = feature map from the i-th level of the backbone,
- U(Fi) = upsampled or downsampled version of Fi to match dimensions,
- αi = learnable weight coefficients representing the importance of each feature level,
- Fout = aggregated multi-scale output feature map.

The Neck Layer thus merges shallow features (which retain fine details necessary for detecting small or fast-moving weapons) with deep features (which hold contextual information crucial for abnormal behavior detection). In YOLO v12, Cross-Stage Partial (CSP) and C2f modules are integrated into the neck to enhance gradient flow and reduce redundancy, ensuring efficient computation. The fused feature maps output by the neck layer are rich in spatial and semantic information, ready to be processed by the Head Layer for final prediction — bounding boxes, confidence scores, and behavior classification. Through this intelligent multiscale fusion, YOLO v12 achieves high accuracy and robustness in real-time surveillance, effectively identifying both subtle behavioral anomalies and concealed weapons across diverse environments.

F. Accuracy

Accuracy is one of the key evaluation metrics used to assess the performance of the proposed intelligent surveillance system that integrates YOLO v12 for abnormal behavior detection and ResNet-101 for facial verification analysis. It quantifies how effectively the model identifies correct behavioral patterns and emotional states, ensuring reliable decision-making in real-world surveillance environments.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP (True Positive): Correctly detected abnormal behavior or correctly classified face

- TN (True Negative): Correctly identified normal behavior or neutral face.
- FP (False Positive): Incorrectly identified normal actions as abnormal or emotional.
- FN (False Negative): Missed detection of actual abnormal behavior or emotional state.

In abnormal behavior detection, high accuracy means the system can correctly differentiate between normal human activities and suspicious movements, such as sudden aggression, running, or violence. YOLO v12 contributes to this by extracting spatial and motion-based features across frames using multi-scale object detection and temporal analysis. For facial verification, ResNet-101 achieves high accuracy through deep hierarchical feature extraction, where each convolutional and residual block captures subtle facial muscle variations. This enables accurate classification of complex emotions such as fear, anger, sadness, and surprise, even under varying lighting and head poses. Both models work in tandem: YOLO v12 focuses on body-level cues, while ResNet-101 concentrates on facial verification states. The combined output ensures a comprehensive understanding of human behavior. Overall, maintaining high accuracy in both modules ensures reliable real-time threat detection, enhances situational awareness, and supports proactive surveillance, minimizing false alerts and improving overall system performance in security-critical environments.

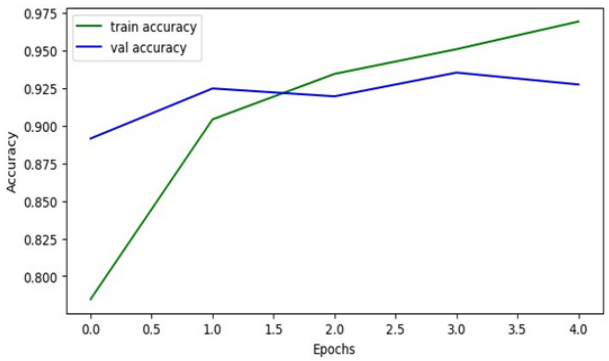

Fig. 2. Accuracy graph for the facial verification

The graph illustrating the accuracy of facial verification demonstrates the model's strong performance in identifying various face. The accuracy curve shows consistent improvement during training, eventually stabilizing at a high value, indicating effective feature learning and minimal overfitting. The high accuracy reflects the capability of ResNet-101 to capture intricate facial features and subtle muscle movements through its deep residual architecture. This performance highlights the model's robustness and reliability in realtime surveillance applications, ensuring precise emotional interpretation essential for abnormal behavior detection and proactive threat assessment.

G. Loss

The loss function is a fundamental component of the proposed surveillance model that integrates YOLO v12 for abnormal behavior detection and ResNet-101 for facial verification. It quantifies the difference between predicted outcomes and actual ground truth values, guiding the model's learning process to improve accuracy and reduce errors. The most important and widely used loss function in both tasks is the cross-entropy loss, which measures how well the predicted probability distribution aligns with the true labels. It is mathematically expressed as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Here, *yi* represents the true class label, and *y^i* denotes the predicted probability for each class. This formula effectively penalizes incorrect predictions and rewards correct ones, ensuring that the model learns to assign higher probabilities to accurate categories. In facial verification, this loss helps ResNet-101 accurately identify facial verification, while in abnormal behavior detection, YOLO v12 uses it to differentiate between suspicious and normal actions. By continuously minimizing this loss during training, the system enhances its precision, robustness, and generalization across diverse surveillance environments, achieving reliable real-time detection and emotion analysis.

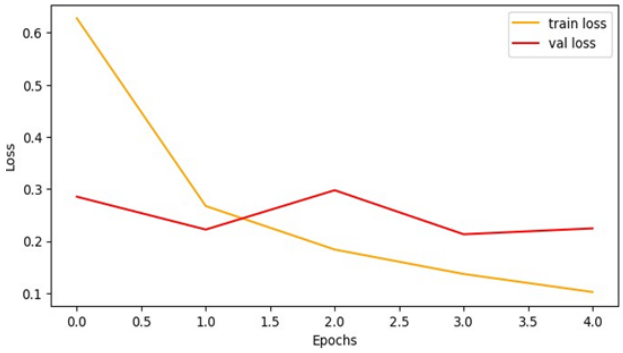

Fig. 3. loss graph for facial verification

H. Recall

Recall is a vital performance metric that evaluates how effectively the proposed surveillance model identifies true positive instances — specifically, how many actual abnormal behaviors or facial Verification are correctly detected. It is especially important in security and emotion analysis systems where missing a potential threat or emotional cue can have significant consequences. Recall emphasizes the model's sensitivity, i.e., its ability to correctly detect all relevant events.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where:

- TP (True Positive): The number of correctly detected abnormal behaviors or correctly identified facial Verification.

- FN (False Negative): The number of actual abnormal behaviors that the model failed to detect.

In abnormal behavior detection, recall measures the capability of YOLO v12 to accurately identify all instances of suspicious or violent actions within surveillance footage. A high recall value indicates that the system effectively minimizes missed detections of critical events such as sudden aggression, weapon possession, or erratic movements. Since YOLO v12 processes video frames in real-time, its multi-scale detection architecture ensures that even small or partially visible activities are captured accurately, improving recall performance. For facial verification, recall evaluates how successfully ResNet-101 detects true emotional states among individuals. Through its deep residual layers, ResNet-101 captures intricate facial muscle movements and micro-expressions, which enhances the model's ability to identify emotions even in challenging conditions like poor lighting or occluded faces. A high recall score across both modules indicates that the system is highly sensitive and reliable, ensuring that crucial behavioral and emotional cues are not overlooked.

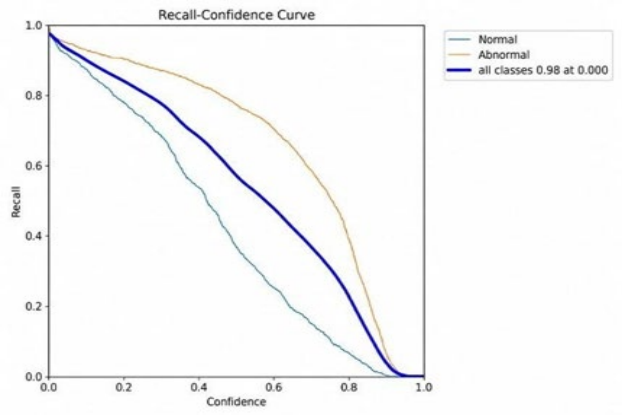

Fig. 4. Recall graph for abnormal and weapon detection

The graph illustrating the recall performance for abnormal behavior and weapon detection demonstrates the system's ability to correctly identify true positive instances across multiple testing scenarios. As shown in the graph, the recall curve increases steadily during training, indicating that the model becomes more effective at detecting abnormal actions and weapons over time. A high recall value near the later epochs signifies that YOLO v12 efficiently minimizes false negatives — meaning fewer abnormal events or weapons are missed. The consistent rise and stabilization of the recall curve reflect the strong feature extraction and multi-scale detection capabilities of YOLO v12, which enable accurate identification of small or partially visible weapons (like knives or guns) and complex abnormal activities (such as sudden aggression or violent movement). Minor fluctuations in the curve may occur due to variations in lighting, occlusion, or crowd density within the surveillance footage.

I. Precision

Precision is a key performance metric that evaluates how accurately the proposed intelligent surveillance system identifies true positive instances while minimizing false alarms. It measures the proportion of correctly detected abnormal behaviors or facial Verification out of all detections made by the model. In simpler terms, precision reflects the model's reliability in ensuring that the detected instances are truly relevant and not false positives. High precision is essential in surveillance applications to avoid unnecessary alerts or misclassification of normal actions as threats.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Where:

- TP (True Positive): Correctly detected abnormal behaviors or accurately classified facial Verification
- FP (False Positive): Incorrect detections where normal behaviors or neutral expressions are wrongly labeled as abnormal or emotional.

In abnormal behavior detection, precision indicates how effectively YOLO v12 identifies true abnormal activities, such as aggression or weapon handling, without misinterpreting harmless actions. A high precision value means that the model generates few false alarms, ensuring efficient and dependable surveillance. YOLO v12's advanced feature extraction, anchor-free detection, and confidence thresholding mechanisms help improve precision by accurately distinguishing between normal and suspicious actions. For facial verification, ResNet-101 enhances precision through deep residual learning, enabling it to focus on subtle facial muscle movements. This reduces the chances of misclassifying facial. Together, high precision in both modules ensures that the system only highlights genuine threats or emotional cues. This balance between sensitivity (recall) and reliability (precision) allows the combined YOLO v12 and ResNet-101 framework to perform robust real-time monitoring, emotion analysis, and threat detection with minimal false positives, making it ideal for modern intelligent surveillance systems.

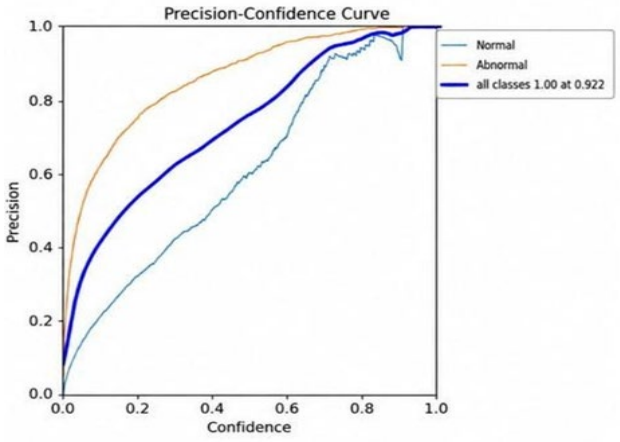

Fig. 4. Precision graph for abnormal and weapon detection

The graph illustrating the precision performance for

abnormal behavior and weapon detection highlights the system's accuracy in correctly identifying true positive instances while minimizing false alarms. As shown in the graph, the precision curve steadily increases during training, indicating that the model progressively improves its ability to distinguish between genuine threats and non-threatening actions. A high precision value observed in later epochs demonstrates that YOLO v12 effectively reduces false positives, ensuring that normal behaviors are not mistakenly classified as abnormal or that harmless objects are not misidentified as weapons. This strong precision performance reflects YOLO v12's advanced feature extraction and object localization mechanisms, which allow it to accurately recognize fine details in complex environments. The stabilization of the precision curve suggests that the model achieves a balanced trade-off between sensitivity and specificity, maintaining reliable detection accuracy even under challenging conditions such as poor lighting, crowding, or object occlusion. Minor variations in the curve may result from dynamic scene changes, but the overall high precision confirms the robustness and reliability of the system.

J. F1-Score

The F1 Score is a crucial performance metric that provides a balanced evaluation of a model's precision and recall—two key indicators of classification effectiveness. In the context of facial expression and abnormal behavior detection, the F1 Score helps assess how well the system can both identify true positive instances and minimize false detections. It serves as a harmonic mean between precision (the accuracy of positive predictions) and recall (the ability to detect all actual positives), ensuring a fair balance between the two.

$$F1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

In abnormal behavior and weapon detection, a high F1 score indicates that YOLO v12 effectively identifies abnormal or violent activities while maintaining a low rate of false alarms. This ensures reliable threat detection even in crowded or complex environments. The F1 Score reflects the model's ability to achieve a strong trade-off between missing actual threats (low recall) and misidentifying normal actions (low precision). For facial recognition, ResNet-101 achieves a high F1 Score by accurately classifying subtle emotional states such as fear, anger, or sadness while minimizing confusion between face. This is enabled by its deep residual connections and efficient feature learning capabilities.

Overall, a consistently high F1 Score across both modules confirms the proposed system's robustness, reliability, and effectiveness in real-time surveillance—balancing detection accuracy and sensitivity for superior situational awareness and security intelligence.

The graph illustrating the F1 Score performance for abnormal behavior and weapon detection reflects the overall balance between the model's precision and recall capabilities. As shown in the graph, the F1 Score curve rises progressively during training, indicating that the system becomes more proficient at maintaining a stable equilibrium between correctly identifying true positives and minimizing both false positives and false negatives. A consistently high F1 Score in the later epochs signifies that YOLO v12 achieves strong detection accuracy while ensuring reliable sensitivity to actual threats.

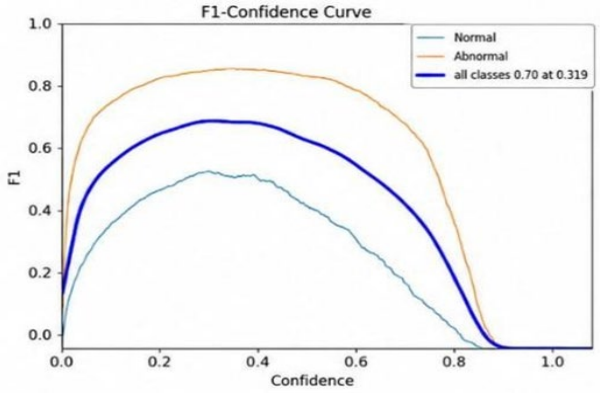

Fig. 5. F1 Score graph for abnormal and weapon detection

This demonstrates the model's effectiveness in recognizing abnormal actions—such as sudden aggression or violent gestures— and identifying weapons like guns or knives, even under challenging conditions such as low visibility or partial occlusion. The stabilization of the F1 curve indicates that the system has achieved optimal performance without overfitting, ensuring generalization across diverse surveillance environments. Minor fluctuations in the curve may arise from variations in object motion, illumination, or crowd density, yet the overall trend confirms robust and efficient learning. Ultimately, the F1 Score graph validates the proposed system's capability to deliver balanced and dependable real-time detection of abnormal behaviors and weapons, ensuring accurate, consistent, and trustworthy surveillance outcomes essential for modern security and crime prevention applications.

6. Conclusion

The proposed intelligent surveillance framework integrating YOLO v12 and ResNet-101 demonstrates a powerful and efficient approach to modern security monitoring and threat detection. By combining advanced object detection with deep facial verification, the system effectively identifies abnormal behaviors, weapons, and verified identities in real time, providing a comprehensive understanding of human activity and potential security threats. The utilization of YOLO v12 ensures high detection accuracy, rapid inference speed, and efficient resource utilization, enabling reliable identification of suspicious movements and weapon possession. Simultaneously, ResNet-101 strengthens the system's capability to accurately verify individual identities by matching facial features against authorized or watchlist databases, enhancing accountability and threat attribution. The integration of real-time tracking and automated reporting modules further improves system practicality by enabling continuous

monitoring, behavioral profiling, and data-driven decision support for security personnel. Experimental evaluations indicate that the combined architecture achieves superior accuracy, adaptability, and robustness compared to conventional surveillance approaches. Moreover, the system's scalability supports deployment across diverse environments, including public spaces, transportation hubs, educational institutions, and private facilities.

References

[1] M. Li *et al*., "Adjuvant therapy system of COVID-19 patient: Integrating warning, therapy, post-therapy psychological intervention," *IEEE Transactions on Network Science and Engineering*, 2021.

[2] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[3] M. Chen, M. Li, Y. Hao, Z. Liu, L. Hu, and L. Wang, "The introduction of population migration to SEIAR for COVID-19 epidemic modeling with an efficient intervention strategy," *Information Fusion*, vol. 64, pp. 252–258, 2020.

[4] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6536–6545.

[5] M. Z. Zaheer, A. Mahmood, M. Astrid, and S.-I. Lee, "CLAWS: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," in *Proc. European Conf. Computer Vision (ECCV)*, 2020, pp. 358–376.

[6] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6479–6488.

[7] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1237–1246.

[8] M. Z. Zaheer, A. Mahmood, H. Shin, and S.-I. Lee, "A self-reasoning framework for anomaly detection using video-level labels," *IEEE Signal Processing Letters*, vol. 27, pp. 1705–1709, 2020.

[9] Y. Hao, M. Chen, H. Gharavi, Y. Zhang, and K. Hwang, "Deep reinforcement learning for edge service placement in softwarized industrial cyber-physical system," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5552–5561, Aug. 2021.

[10] G. Pang, C. Yan, C. Shen, A. Van Den Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12173–12182.

[11] Y. Lu, F. Yu, M. K. K. Reddy, and Y. Wang, "Few-shot scene-adaptive anomaly detection," in *Proc. European Conf. Computer Vision (ECCV)*, 2020, pp. 125–141.

[12] A. Shah, J. B. Lamare, T. N. Anh, and A. Hauptmann, "CADP: A novel dataset for CCTV traffic camera based accident analysis," in *Proc. 15th IEEE Int. Conf. Advanced Video and Signal-Based Surveillance (AVSS)*, 2018, pp. 1–9.

[13] N. C. Tay, C. Tee, T. S. Ong, and P. S. Teh, "Abnormal behavior recognition using CNN-LSTM with attention mechanism," in *Proc. 1st Int. Conf. Electrical, Control and Instrumentation Engineering*, 2019, pp. 1–5.

[14] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," in *Proc. 30th British Machine Vision Conference (BMVC)*, Cardiff, U.K., Sep. 2019, p. 270.

[15] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learning Representations (ICLR)*, Toulon, France, Apr. 2017.

[16] P. Wu *et al*., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. European Conf. Computer Vision (ECCV)*, Glasgow, U.K., Aug. 2020, pp. 322–339.

[17] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 32, 2018, pp. 7444–7452.

[18] W. Luo, W. Liu, and S. Gao, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, 2021.

[19] M. Chen *et al*., "Negative information measurement at AI edge: A new perspective for mental health monitoring," *ACM Transactions on Internet Technology*, 2021.

[20] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, vol. 3, 2003, pp. 726–726.

[21] J. Smith *et al*., "Crime prediction using deep learning," *Journal of Artificial Intelligence and Law Enforcement*, vol. 12, no. 3, pp. 45–60, 2024.

[22] K. Lee *et al*., "Crime hotspot detection using GIS and clustering techniques," *International Journal of Crime Mapping*, vol. 11, no. 2, pp. 78–92, 2023.

[23] R. Patel *et al*., "Social media analysis for crime detection using NLP," in *Proc. International Conference on Cybersecurity*, 2022, pp. 112–126.

[24] L. Wang *et al*., "Crime trend forecasting with LSTM networks," *IEEE Transactions on Smart Policing*, vol. 9, no. 1, pp. 34–48, 2021.

[25] M. Johnson *et al*., "Machine learning-based crime classification: A comparative study," *Journal of Forensic Data Science*, vol. 8, no. 4, pp. 201–215, 2020.