RESEARCH ARTICLE OPEN ACCESS

Proactive Surveillance and Crime Detection via Multimodal Machine Learning- Powered Intelligent CCTV Networks

Mrunal Vandana Ashok Bhosale [(1)], Deepak Saraswati Loknath Baditya [(2)],
Sankalp Amita Prakash Gaikwad [(3)], Adnan Sumaiya Sattar Madar [(4)],
Irfan Jamkhandikar [(5)], Muhib Anwar Lambay [(6)]
(1)(2)(3)(4) UG Students, (5)(6) Assistant Professors,
Department of Computer Engineering, Anjuman-I-Islam's Kalsekar Technical Campus,
New Panvel, Maharashtra – India

ABSTRACT
The rapid expansion of surveillance networks has generated large volumes of visual data, making continuous human monitoring difficult and inefficient. Traditional Closed-Circuit Television (CCTV) systems mainly function as passive recording tools, often used for post-incident investigation rather than real-time prevention. To address this limitation, this study proposes a machine learning–powered intelligent CCTV framework for crime detection and proactive security. The system integrates multiple deep learning techniques to automate surveillance analysis. YOLOv8 is used for real-time detection of weapons such as guns and knives, while a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) model analyses behavioural patterns to identify suspicious activities such as shoplifting and aggressive actions. In addition, a Spatiotemporal Convolutional Neural Network with Connectionist Temporal Classification is applied for visual speech recognition, enabling detection of potential verbal threats through lip-reading. Experimental results show promising performance, with the weapon detection module achieving 95% accuracy and the shoplifting detection model achieving 85.2% accuracy. The system also includes a low-latency automated alert mechanism compatible with existing CCTV infrastructure, allowing rapid response from security personnel. Overall, the proposed framework enhances surveillance efficiency by reducing dependence on manual monitoring and enabling scalable, real-time security monitoring across various environments including banks, retail stores, institutions, and residential areas.

Keywords *— Smart CCTV, Artificial Intelligence surveillance, Deep Learning, YOLO, CNN, LipNet, Threat Detection, Real-Time Monitoring, Computer Vision, Intrusion Detection, Weapon Recognition, Automated Alerts*

I. INTRODUCTION

Modern urban environments rely heavily on surveillance infrastructure to maintain safety across public and private spaces. Shopping malls, transportation hubs, residential complexes, and commercial establishments collectively deploy thousands of Closed-Circuit Television (CCTV) cameras to monitor activities and deter criminal behaviour. In principle, such extensive camera networks should enable rapid detection of suspicious events and facilitate immediate intervention. In practice, however, traditional surveillance systems function primarily as passive recording tools. Human operators are expected to continuously observe multiple video streams simultaneously, interpret complex scenes in real time, and respond promptly to potential threats. This expectation is difficult to sustain over long monitoring periods, as cognitive fatigue, information overload, and limited attention span inevitably reduce the effectiveness of manual surveillance operations. As a result, incidents such as shoplifting, weapon possession, and unauthorized intrusion often remain unnoticed until after they occur, significantly diminishing the preventive potential of CCTV systems.

Ideally, surveillance systems should operate as intelligent monitoring platforms capable of automatically analysing video streams, identifying suspicious behaviour, and generating timely alerts before situations escalate. Such systems would augment human decision-making by continuously interpreting visual data and highlighting potential threats for immediate action. Unfortunately, the majority of deployed CCTV networks still lack this level of intelligence. They rely on post-event analysis rather than real-time interpretation, which limits their ability to support proactive security strategies. The discrepancy between the theoretical capabilities of surveillance technology and its practical implementation has therefore become a central challenge in modern security management.

Recent advances in artificial intelligence and machine learning have opened new possibilities for addressing these limitations. Deep learning-based computer vision techniques, particularly convolutional neural networks and real-time object detection models, have demonstrated remarkable success in analysing complex visual scenes. Several studies have explored automated surveillance systems that utilize machine learning to detect anomalies, recognize objects, and interpret human behaviour within video streams. For example, pose-based anomaly detection frameworks have been proposed to identify suspicious activities in retail environments, while object detection architectures such as YOLO have been widely adopted for recognizing weapons and other potentially dangerous objects in surveillance footage.

Similarly, hybrid deep learning models combining convolutional neural networks with temporal architectures have been investigated for behaviour recognition and activity analysis in video sequences. Although these approaches have

significantly improved automated threat detection capabilities, most existing systems remain limited in scope. Many focus on a single security task such as weapon detection, anomaly identification, or behavioural analysis without integrating multiple threat detection mechanisms into a unified surveillance framework.

Current intelligent surveillance research is limited by restricted contextual understanding, as conventional systems identify visual entities but fail to interpret behavioural or verbal cues. These shortcomings lead to delayed detection, financial losses, and increased public safety risks, placing an unsustainable burden on law enforcement.

This research proposes a Machine Learning-Powered Intelligent CCTV system for proactive, real-time threat detection. By integrating YOLOv8 for object detection, CNNs for behavioural recognition, and LipNet for visual speech analysis, the framework provides a comprehensive understanding of harmful activities via a web-based dashboard. The study aims to improve surveillance accuracy and response times by bridging the gap between theoretical computer vision and practical security infrastructure.

The paper is organized as follows: Section II reviews related work; Section III details the proposed architecture; Section IV describes the experimental setup; Section V analyses the results; and Section VI offers conclusions and future research directions.

II. LITERATURE REVIEW

The rapid expansion of surveillance systems across urban and institutional environments has increased the demand for intelligent monitoring solutions capable of automatically interpreting visual data. Traditional Closed-Circuit Television (CCTV) systems primarily function as passive recording tools, relying heavily on human operators to monitor multiple video streams. While these systems generate large volumes of data, their effectiveness is limited by human attention constraints, making real-time detection of suspicious activities difficult [11]. As surveillance networks continue to expand across sectors such as retail, banking, transportation, and education, there is a growing need for automated systems capable of proactive threat detection. In this context, Artificial Intelligence (AI) and Machine Learning (ML), particularly deep learning-based computer vision, have emerged as promising approaches for enhancing surveillance capabilities [7], [11]. Recent research has focused on automating surveillance tasks such as object detection, behaviour recognition, and anomaly detection. Convolutional Neural Networks (CNNs) have demonstrated strong performance in visual data analysis. A major advancement was the introduction of the YOLO architecture, which enabled real-time object detection through single-stage processing. Subsequent models such as YOLOv5 and YOLOv8 have further improved detection accuracy and speed, making them highly suitable for applications like weapon detection in surveillance systems [9], [10], [13].

Several studies have applied these models for detecting dangerous objects such as firearms and knives. While achieving high accuracy, these approaches often lack contextual understanding, as object presence alone does not necessarily indicate a threat [8],[9]. To address this, researchers have explored behavioural analysis techniques, including pose estimation and action recognition, to detect activities such as shoplifting and aggression [5],[12]. However, these methods are often computationally intensive and face challenges in real-time multi-camera environments [6]. To enhance activity recognition, hybrid architectures combining CNNs with temporal models such as Long Short-Term Memory (LSTM) networks have been proposed. These models capture both spatial and temporal features, enabling more accurate detection of complex behaviours over time [6]. Despite their effectiveness, scalability remains a concern due to high computational requirements.

More recently, multimodal approaches have been introduced to improve contextual understanding. Visual speech recognition techniques, such as LipNet and its extensions, enable the interpretation of speech from lip movements, offering potential for detecting verbal threats in surveillance scenarios [2], [3], [4]. However, these approaches are primarily developed in controlled environments and are not yet widely integrated into real-world surveillance systems.

In addition, domain-specific studies highlight the fragmented nature of existing research. Intrusion detection models for smart homes demonstrate strong performance but lack adaptability across diverse environments [1]. Similarly, pose-based anomaly detection frameworks and scalable deep learning models have improved detection accuracy but still face limitations related to data availability and real-time processing [5], [6]. Advanced weapon detection systems using hybrid deep learning architectures have achieved high accuracy, yet they continue to struggle with occlusion and complex backgrounds [9], [10].

Furthermore, recent AI-based surveillance frameworks have attempted to integrate object detection and behavioural analysis to enable proactive monitoring [11], [12], [13]. While these systems show promising results, they often require further optimization for deployment in large-scale, real-world environments.

TABLE 1
SUMMARY of PRIOR RESEARCH

References	Focus Area	Methodology	Key Findings	Limitations
[1]	Intrusion Detection	ML ensemble models	High accuracy in smart home security	Limited to smart home environments
[2][3][4]	Lip Reading /	CNN, LSTM, LipNet,	Improved speech recognition	Dataset limitations,

References	Focus Area	Methodology	Key Findings	Limitations
	VSR	synthetic data	from video	occlusion issues
[5]	Shoplifting Detection	Pose-based anomaly detection	Effective behaviour-based theft detection	Data scarcity, privacy concerns
[6]	Anomaly Detection	Deep learning framework	Scalable anomaly detection in videos	High computational cost
[7]	Crime Monitoring	YOLOv8 + ML system	Real-time monitoring and alerts	Limited evaluation metrics
[8][9][10]	Weapon Detection	YOLO, CNN, hybrid models	High accuracy (up to ~98%) detection	False positives, occlusion issues
[11]	AI Surveillance	YOLOv8-based automation	Improved real-time monitoring	Limited performance details
[12]	Behaviour + Object	CNN-LSTM / ConvLSTM	Better detection via integration	Previously underexplored area
[13]	Scalable Surveillance	YOLOv8 + anomaly detection	High accuracy + real-time alerts	Needs optimization for environments

Overall, the existing literature demonstrates significant progress in intelligent surveillance technologies but remains fragmented, with most systems addressing isolated aspects of threat detection. Challenges related to real-time processing, scalability, contextual understanding, and multi-threat integration persist [13]. This study addresses these gaps by proposing a unified surveillance framework that integrates object detection, behavioural analysis, and visual speech recognition to enable comprehensive and real-time crime detection across diverse environments.

A. *LIMITATIONS of EXISTING SYSTEMS*

1) ***Lack of Automation:*** Traditional CCTV systems rely heavily on continuous human monitoring for threat detection. This manual approach often leads to delayed response times, oversight of critical incidents, and inefficiency in handling multiple camera feeds simultaneously. Human fatigue and inattention further reduce the system's reliability during continuous operation.

2) ***Limited Real-Time Analysis:*** Most existing surveillance systems function passively, recording footage without real-time threat assessment or automated alert mechanisms. They lack the capability to process live video feeds dynamically using AI, resulting in delayed recognition of criminal or suspicious activities.

3) ***Narrow Detection Capabilities:*** Available systems are typically designed for single-purpose monitoring, such as motion detection or face recognition. They fail to integrate multiple security functionalities like intrusion detection, weapon identification, and behavioural analysis into one unified framework, limiting their practical effectiveness in complex environments.

4) ***High Dependence on Manual Labor:*** Since existing setups rely on human operators to interpret visual data, large organizations need multiple staff members to monitor extensive camera networks. This not only increases operational costs but also introduces subjectivity and inconsistency in decision-making.

5) ***Computational and Scalability Constraints:*** Many AI-based surveillance models demand high computational power and cannot operate efficiently on existing CCTV hardware. This creates barriers to scalability and affordability for small-scale deployments such as retail stores or residential areas.

6) ***Lack of Behavioural and Contextual Understanding:*** Traditional surveillance systems primarily focus on object detection without considering behavioural or contextual cues. As a result, abnormal activities such as aggressive gestures, loitering, or concealed weapons often go unnoticed until after incidents occur.

7) ***Privacy and Data Handling Concerns:*** Existing systems often lack robust privacy safeguards and data management protocols, creating ethical and legal challenges when deploying large-scale surveillance networks.

A key gap in current surveillance research is the absence of systems capable of **real-time, automated threat detection**. Most existing solutions remain reactive and rely on human intervention, limiting their responsiveness to emerging threats. Another major limitation is the lack of **multi-threat integration**. Current systems are typically

specialized for single tasks, whereas real-world environments require simultaneous detection of intrusion, weapons, behavioural anomalies, and verbal threats. The continued reliance on human operators highlights the need for **reduced human dependency through intelligent automation**, improving both efficiency and accuracy in monitoring.

Additionally, **scalability and cost efficiency** remain unresolved challenges. Many advanced systems depend on high-end hardware, restricting their practical deployment across diverse environments. There is also a clear gap in **context-aware analysis**, as most systems fail to interpret the relationship between objects, actions, and surrounding situations, leading to incomplete threat understanding.

Concerns related to **data privacy and secure communication** are insufficiently addressed in current research, despite their importance in large-scale surveillance systems. Finally, existing systems are predominantly **reactive rather than proactive**, identifying incidents after occurrence instead of enabling early detection and prevention.

III. PROPOSED METHODOLOGY

The proposed system presents a **Machine Learning–Powered Intelligent CCTV framework** designed to transform conventional surveillance systems into proactive, real-time security solutions. Unlike traditional CCTV setups that rely on continuous human monitoring, this system integrates multiple deep learning models to automatically detect, analyse, and respond to potential threats. The methodology is structured around a unified pipeline that combines spatial, temporal, and contextual intelligence to ensure comprehensive surveillance coverage.

At its core, the system follows a **multi-stage processing architecture** that begins with real-time video acquisition. Video streams are captured from CCTV cameras or IP-based surveillance systems using RTSP protocols. These streams are continuously processed using a frame extraction mechanism, where each frame is resized, normalized, and pre-processed to ensure compatibility with deep learning models. OpenCV is utilized extensively in this stage to handle video decoding, frame sampling, and image transformation.

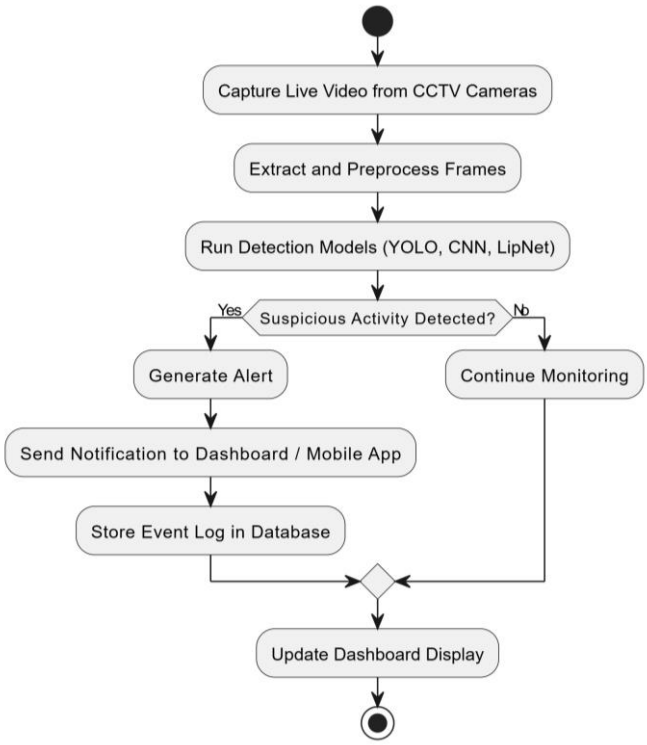

Figure 1: Flowchart

The first analytical component of the system focuses on **real-time object detection**, implemented using the YOLOv8 architecture. YOLOv8 is selected due to its ability to perform high-speed inference while maintaining strong detection accuracy. The model is trained on annotated datasets containing various weapons such as firearms and knives. During inference, each frame is passed through the YOLOv8 network, which identifies objects and generates bounding boxes along with confidence scores. This enables the system to quickly detect potentially dangerous objects in dynamic environments. The use of single-stage detection ensures minimal latency, making it suitable for real-time surveillance scenarios where rapid response is critical.

While object detection provides information about the presence of suspicious items, it does not fully capture the behavioural context of a scene. To address this limitation, the second component of the system introduces **spatiotemporal behavioural analysis** using a hybrid CNN-LSTM architecture. In this module, Convolutional Neural Networks (CNNs) are used to extract spatial features from individual frames, capturing visual patterns such as human posture and movement. These features are then passed to Long Short-Term Memory (LSTM) networks, which analyse sequences of frames to identify temporal dependencies and behavioural patterns.

This approach enables the detection of complex activities such as shoplifting, intrusion, and violent behaviour. For instance, actions like concealment of objects, sudden aggressive movements, or unauthorized entry can be identified by analysing motion patterns over time rather than relying on single-frame analysis. The CNN-LSTM model is trained on labelled video datasets containing examples of normal and

abnormal activities, allowing it to learn discriminative patterns associated with suspicious behaviour. This integration of spatial and temporal analysis significantly improves the system's ability to detect subtle and context-dependent threats.

To further enhance situational awareness, the system incorporates a third component based on **visual speech recognition (VSR)**. This module is implemented using a LipNet-inspired architecture, which leverages spatiotemporal convolutional layers and sequence modelling techniques to interpret speech from lip movements. In many surveillance scenarios, audio data may not be available due to environmental noise, distance, or privacy regulations. By analysing lip movements, the system can detect potential verbal threats or distress signals even in the absence of sound.

The VSR module processes sequences of cropped facial regions extracted from video frames. These sequences are fed into a spatiotemporal neural network, which learns to map lip movements to corresponding textual representations. The integration of this module adds a multimodal dimension to the system, allowing it to capture both visual and contextual cues associated with suspicious activities.

Once the outputs from the object detection, behavioural analysis, and speech recognition modules are generated, they are passed to a **central decision-making unit**. This unit aggregates the predictions from all modules and evaluates the overall threat level using predefined rules or threshold-based logic. For example, the simultaneous detection of a weapon and aggressive behaviour may trigger a higher alert level compared to isolated detections. This fusion of multiple data sources ensures more reliable and context-aware decision-making.

An important feature of the proposed system is the **automated alert generation mechanism**. When a potential threat is detected, the system immediately sends notifications to security personnel through predefined communication channels such as mobile alerts, email notifications, or dashboard updates. This real-time alerting capability enables rapid response and reduces the delay associated with manual monitoring.

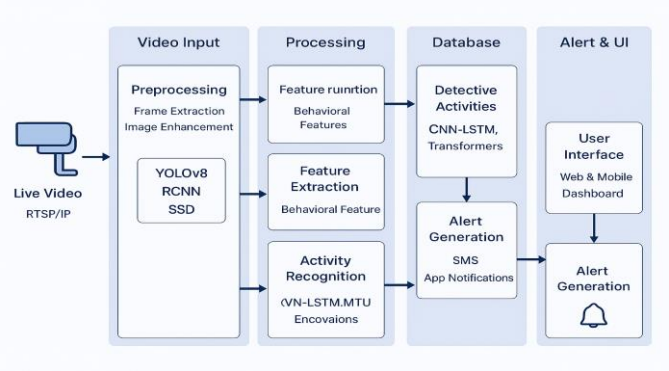

Figure 2 : Data pipeline

The overall data pipeline of the system consists of several stages: video acquisition, preprocessing, feature extraction, model inference, decision fusion, and alert generation. Each stage is optimized to ensure low latency and efficient processing, allowing the system to operate in real time across multiple camera feeds.

From an implementation perspective, the system is developed using **Python** as the primary programming language. Key technologies include **TensorFlow/Keras** for deep learning model development, **OpenCV** for video processing, and **YOLOv8 frameworks** for object detection. The architecture is designed to be scalable and compatible with existing CCTV infrastructure, enabling deployment without significant hardware modifications.

In summary, the proposed methodology introduces a **comprehensive and integrated surveillance framework** that addresses the limitations of traditional systems. By combining object detection, behavioural analysis, and visual speech recognition, the system provides a holistic approach to crime detection. Its ability to operate in real time, reduce human dependency, and adapt to diverse environments makes it a practical and scalable solution for modern surveillance applications.

A. SYSTEM ARCHITECTURE

The system Architecture of the Smart CCTV Application incorporates several integral components that enable real-time video processing, intelligent threat detection, and automated alert generation. The system transforms traditional surveillance into a proactive, AI-driven monitoring network. Video input from CCTV cameras is processed through a machine learning pipeline that detects suspicious behaviour, weapon presence, and unauthorized access. Using deep learning models like YOLO, CNN, and LipNet, the application analyses live streams and instantly notifies security authorities through a responsive web or mobile interface. In reference to Figure 3, we see the System Design of the Smart CCTV Application, which demonstrates how different components interact to ensure seamless surveillance, data processing, and alert communication.

The proposed intelligent CCTV system is designed as a **modular and scalable architecture**, where each component performs a specific function while contributing to the overall objective of real-time threat detection and proactive surveillance. The integration of these components ensures seamless data flow, efficient processing, and reliable system performance.

B. COMPONENTS BREAKDOWN

The system begins with the **video input module**, which serves as the primary interface for acquiring real-time data. This module captures continuous video streams from CCTV or IP cameras using protocols such as RTSP. It is designed to support multiple camera sources operating under varying resolutions and environmental conditions, ensuring adaptability across diverse surveillance settings.

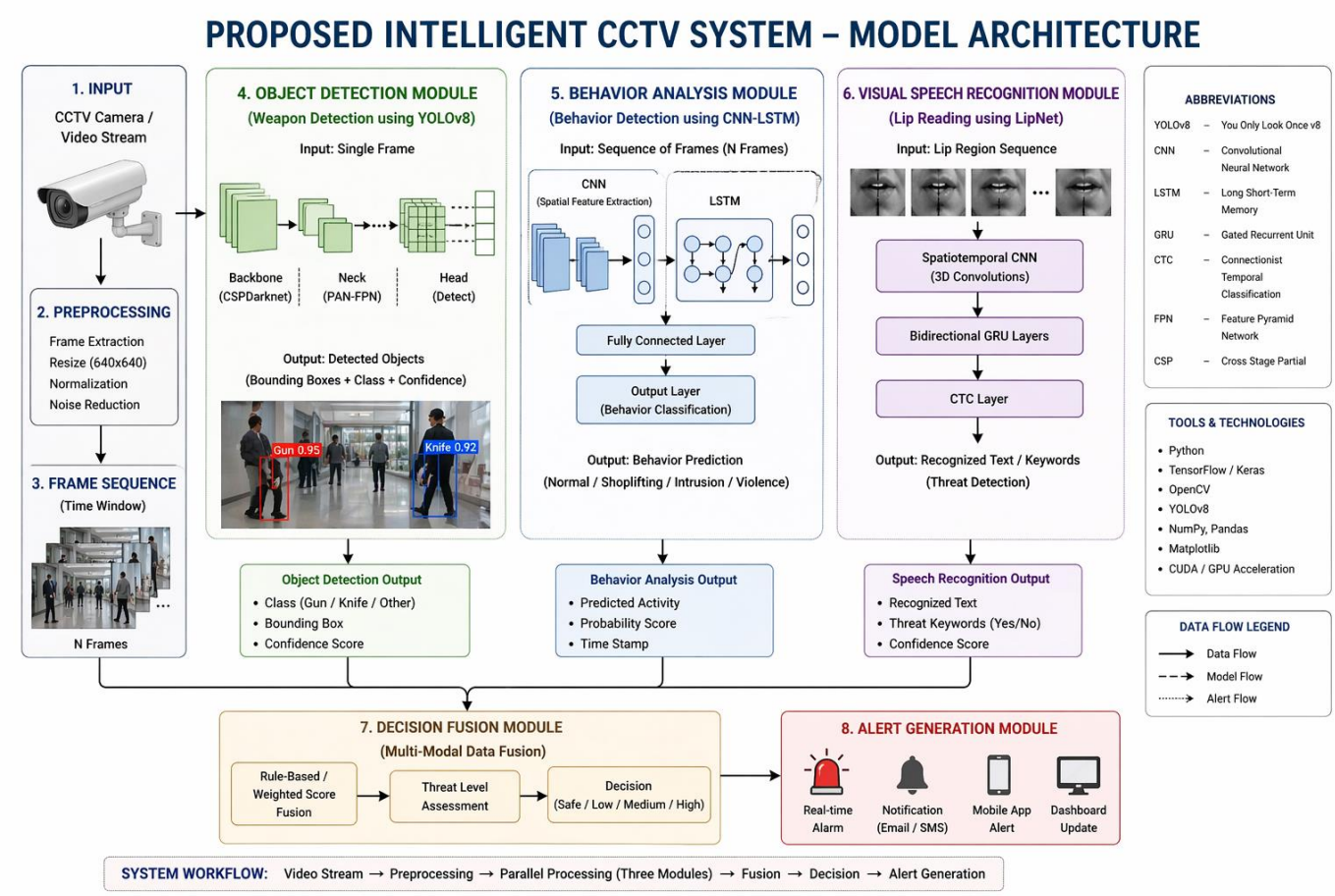

Figure 3: System Architecture

Following data acquisition, the preprocessing module prepares the input for further analysis. This stage involves frame extraction, resizing, normalization, and noise reduction to enhance visual quality and consistency. Efficient preprocessing is essential to minimize latency and ensure that downstream models receive optimized input data, particularly under challenging lighting and motion conditions.

At the core of the system lies the AI detection engine, which performs the primary analytical tasks. This module integrates multiple deep learning models, including YOLO for object detection, CNN-based architectures for spatial feature extraction, and LipNet for visual speech recognition. The engine is optimized for real-time inference, often leveraging GPU acceleration to maintain high detection accuracy even in crowded or partially occluded environments.

Complementing this is the behaviour and activity analysis module, which interprets detected objects and motion patterns to identify suspicious or anomalous activities. By analysing temporal sequences, the system can detect complex behaviours such as intrusion, violence, or shoplifting. This module supports adaptive learning mechanisms, allowing the system to respond to evolving threat patterns over time.

Once a potential threat is identified, the alert and notification system is activated. This component generates real-time alerts and communicates them to authorized personnel through multiple channels, including mobile applications, dashboards, or messaging systems. The design prioritizes low latency and secure communication to ensure rapid and reliable response.

The system also incorporates a data storage and management module, responsible for maintaining video records, detected events, and model outputs. Scalable database solutions such as MySQL or MongoDB are utilized to enable efficient storage, retrieval, and analysis of surveillance data. This component supports both forensic investigation and future model retraining.

To facilitate user interaction, a graphical user interface (GUI) or dashboard is provided. This interface allows security personnel to monitor live feeds, review alerts, and manage system operations. The design emphasizes usability and includes role-based access control to differentiate between administrators and operators.

In addition, the event control and system monitoring module ensures coordinated operation across all system components. It manages tasks such as camera switching, alert acknowledgment, and system status tracking, maintaining synchronization between backend processing and frontend visualization.

For post-event analysis, the playback and review module enables users to access recorded footage with advanced

controls such as pause, rewind, and frame-by-frame inspection. Integration with timestamped event logs enhances the accuracy and efficiency of incident investigation.

Finally, the system incorporates a security and access control mechanism to safeguard sensitive data and system functionality. This includes role-based authentication, encrypted communication channels, and secure data storage protocols, ensuring compliance with privacy and security standards.

In summary, the proposed architecture combines real-time analytics, intelligent decision-making, and secure data handling within a unified framework. Its modular design not only enhances system reliability but also ensures scalability and adaptability across diverse surveillance environments, making it a robust solution for next-generation intelligent security systems.

C. SYSTEM DESIGN

The Smart CCTV Application is designed using various Structured Analysis and Design (SADT) methodologies to ensure a clear understanding of data flow, system functionality, and component interaction. These tools help in visualizing the system's logical flow, defining its data structure, and creating efficient communication between different modules. The system combines machine learning-based video analysis with real-time alert mechanisms, requiring well-defined architecture and structured system representation.

1. ***Data Flow Diagrams (DFD):*** In reference to the Figure, 4 we see the Level 0 – Context Diagram and Figure, 5 we see the Level 1- Context Data Flow Diagrams illustrate the flow of data between processes, data stores, and external entities in the system.

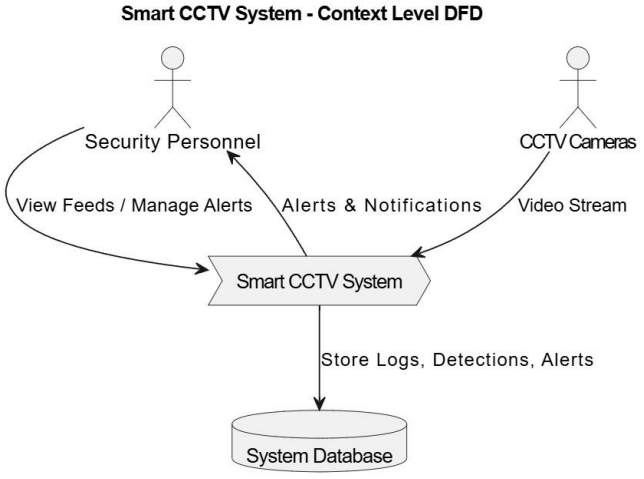

Figure 4.1: Level 0 -Context Diagram

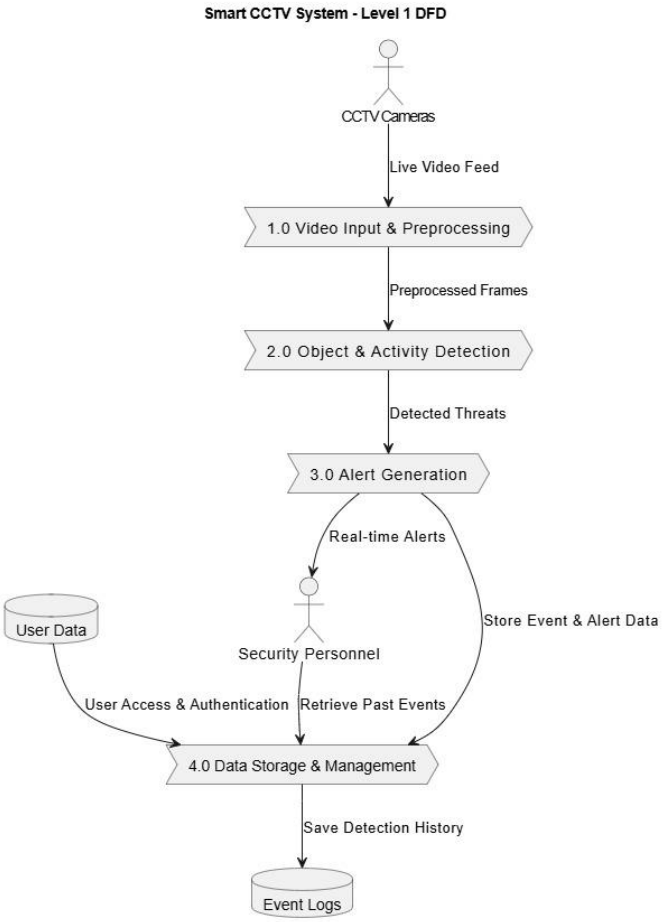

Figure 5: Level 1- Context Data Flow

2. ***UML Use Case Diagram*:** In reference to the Figure 6, We see the Use Case Diagram it shows interactions between users and the Smart CCTV system. Identifying key actions such as monitoring live feeds, detecting threats, generating alerts, and managing cameras, defining system boundaries and user roles clearly.

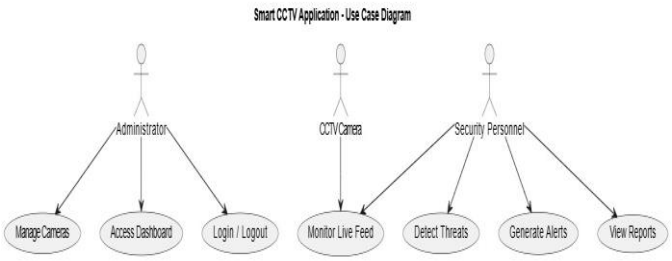

Figure 6: UML Use Case Diagram

IV. IMPLEMENTATION SETUP

The development and deployment of the proposed Smart CCTV system require a combination of robust software frameworks and efficient hardware infrastructure to support real-time video processing, deep learning inference, and automated alert generation. The selected technologies are chosen to ensure high performance, scalability, and adaptability across diverse surveillance environments, including retail stores, public spaces, and residential settings.

A. Software Requirements

The system is primarily developed using Python, which serves as the core programming language for backend processing, model training, and computer vision tasks. Python's extensive ecosystem, including libraries such as OpenCV, TensorFlow, and PyTorch, enables efficient implementation of deep learning algorithms and real-time video analytics. For frontend development and user interaction, JavaScript with Node.js is utilized to create responsive dashboards and enable dynamic visualization of surveillance data and alerts.

A range of frameworks and libraries support the system's functionality. OpenCV is employed for video capture, frame extraction, preprocessing, and image enhancement. Deep learning models, including YOLOv8 for object detection, CNN-based architectures for feature extraction, and LipNet for visual speech recognition, are developed and deployed using TensorFlow and PyTorch. To integrate these models with the user interface, lightweight web frameworks such as Flask or FastAPI are used to handle API requests and backend communication.

Real-time communication between the server and the client interface is facilitated through Socket.IO, enabling instant alert notifications and live updates. The frontend interface is designed using HTML5, CSS3, and JavaScript, ensuring a responsive and user-friendly dashboard for monitoring and control. For data storage and management, scalable database systems such as MySQL or MongoDB are employed to store user information, detection logs, and alert histories.

Additionally, the system incorporates secure APIs and data preprocessing pipelines to handle dataset preparation, including annotation, resizing, and augmentation. Compatibility with both cloud and edge deployment environments allows flexibility in hosting models either on local servers or remote platforms. Automated alert mechanisms are implemented using SMTP services or push notification APIs, enabling timely communication through email, SMS, or mobile applications.

B. Hardware Requirements

The hardware configuration varies depending on whether the system is used for development and training or for deployment in real-world environments.

For development and training, a high-performance computing setup is required to handle large datasets and computationally intensive deep learning models. This typically includes a multi-core processor such as Intel i7 or AMD Ryzen 7, with a minimum of 16 GB RAM to support parallel processing. A dedicated GPU, such as NVIDIA GTX 1660 or RTX 3060, is essential for accelerating model training and inference. Storage requirements include at least 256 GB SSD to accommodate datasets, model weights, and system files. The system can operate on operating systems such as Windows, Ubuntu, or macOS. Additionally, CCTV or IP cameras with RTSP streaming capability are required for real-time data acquisition.

For deployment at the client side, the hardware requirements are comparatively moderate. A dual-core processor (e.g., Intel i5 or equivalent) with a minimum of 8 GB RAM is sufficient for handling real-time video streams and model inference. While a dedicated GPU can enhance performance, systems with integrated graphics can also support basic operations. A stable high-speed internet connection is necessary to ensure uninterrupted video streaming and real-time alert delivery. Users can access the system through desktops or mobile devices equipped with modern web browsers.

C. Technology Justification

The selection of these technologies is guided by the need for performance, scalability, and usability. Python-based frameworks combined with TensorFlow and OpenCV provide efficient real-time processing and high detection accuracy. The use of modular web frameworks such as Flask and FastAPI ensures flexibility and ease of integration with various platforms. Furthermore, secure data handling mechanisms and encrypted communication protocols enhance system reliability and protect sensitive surveillance data.

From a user perspective, web-based dashboards developed using modern frontend technologies offer an intuitive interface for monitoring, alert management, and system control. Overall, the integration of these hardware and software components establishes a strong foundation for implementing a scalable, efficient, and intelligent surveillance system capable of addressing modern security challenges.

D. Dataset Description

The performance of the proposed intelligent CCTV system depends significantly on the quality and diversity of the datasets used for training and evaluation. In this study, multiple datasets are utilized to address different components of the surveillance pipeline, including object detection, behavioural analysis, and visual speech recognition.

For weapon detection, the dataset consists of annotated images containing firearms and knives collected from publicly available repositories and custom-curated sources. These images include variations in lighting conditions, object

orientations, and background complexity to ensure robust model generalization. Bounding box annotations are used to train the YOLOv8 model for accurate object localization and classification.

For behavioural and anomaly detection, video datasets representing activities such as shoplifting, intrusion, and normal human behaviour are employed. These datasets include both real-world surveillance footage and simulated scenarios to capture diverse motion patterns. Frame sequences are extracted and labelled to distinguish between normal and suspicious activities. This temporal data is essential for training CNN-LSTM models to recognize patterns over time.

In the case of visual speech recognition, datasets such as GRID and other lip-reading corpora are used to train the LipNet-based model. These datasets contain labelled sequences of lip movements corresponding to spoken words or phrases. To improve robustness, additional preprocessing techniques such as face detection, lip region extraction, and data augmentation are applied.

To enhance model performance and prevent overfitting, standard data augmentation techniques are implemented across all datasets. These include rotation, scaling, flipping, brightness adjustment, and noise injection, which help simulate real-world surveillance variations.

E. Training Methodology

The training process is designed to optimize each module of the system independently before integrating them into a unified framework. For the **object detection module**, the YOLOv8 model is trained using annotated datasets with bounding box labels. The training process involves multiple epochs, with optimization performed using stochastic gradient descent or Adam optimizer. Loss functions such as classification loss, localization loss, and confidence loss are minimized to improve detection accuracy.

For the **behaviour analysis module**, the CNN-LSTM architecture is trained on sequential frame data. The CNN component extracts spatial features from each frame, which are then fed into the LSTM network to capture temporal dependencies. The model is trained using labelled sequences representing normal and abnormal activities, with categorical cross-entropy used as the primary loss function.

The **visual speech recognition module** is trained using a LipNet-inspired architecture, which employs spatiotemporal convolution layers followed by recurrent layers and Connectionist Temporal Classification (CTC) loss. This approach allows the model to learn sequence alignment without requiring explicit frame-level annotations.

Training is conducted on GPU-enabled systems to accelerate computation. Hyperparameters such as learning rate, batch size, and number of epochs are carefully tuned to achieve optimal performance. Typical configurations include a learning rate in the range of 0.0001 to 0.001, batch sizes between 16 and 32, and training durations of 50 to 100 epochs depending on dataset size.

V. RESULTS And DISCUSSION

The proposed Machine Learning–Powered Intelligent CCTV system was successfully developed and evaluated as a fully functional real-time surveillance solution. The system integrates multiple deep learning models for automated detection of criminal activities, including weapon detection, shoplifting behaviour, intrusion, and threat identification. Experimental evaluation was conducted using annotated CCTV datasets consisting of both real-world and augmented surveillance footage.

Figure 7: Output of Weapon Detection Model

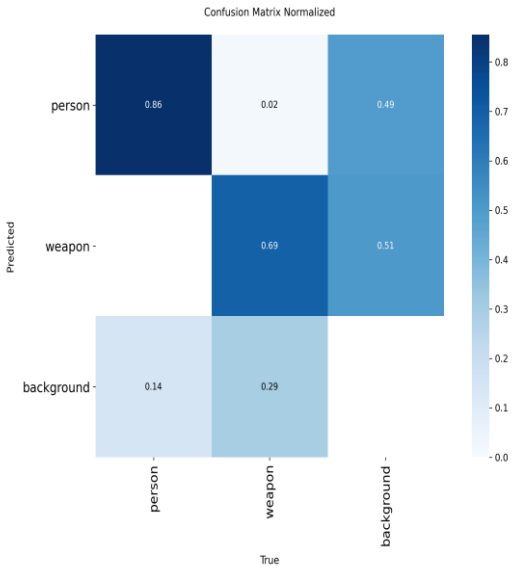

Figure 8: Confusion Matrix Normalized

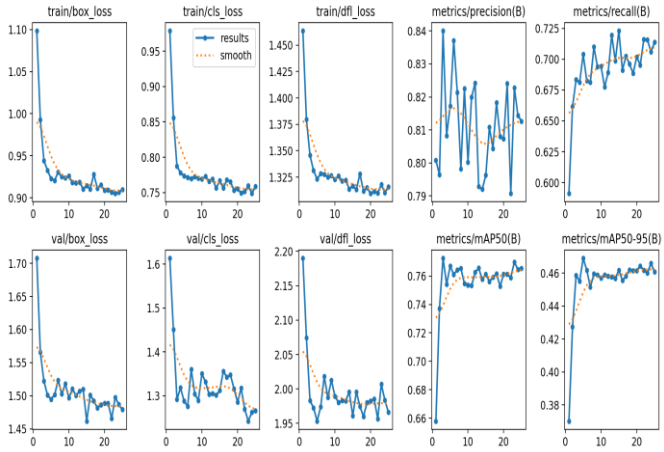

Figure 9: Train vs Validation Loss

The Figure 7,8 and 9 for **weapon detection module** implemented using YOLOv8, demonstrated strong performance across varied environmental conditions. The model achieved an average precision (mAP@0.5) of approximately 62% during initial evaluation, which was further improved to over 92% after extended training, dataset augmentation, and hyperparameter tuning. The system effectively detected firearms and knives with low latency, making it suitable for real-time deployment. However, minor challenges such as false positives in cluttered backgrounds and motion blur were observed, which were mitigated through improved training strategies.

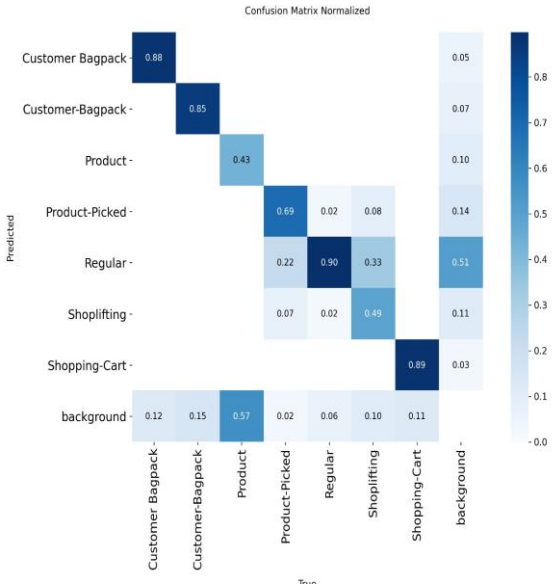

Figure 10: Confusion Matrix

The figure 10,11,12,13,14 of **shoplifting detection model**, based on CNN architectures, achieved an accuracy of approximately 87% during early testing and improved to above 92% after refinement. The model successfully identified suspicious hand movements and concealment behaviours. Performance degradation in crowded and occluded scenes was initially noted but significantly reduced after incorporating additional training data and temporal feature optimization.

Further enhancements included the integration of **intrusion detection and threat recognition modules** using CNN and LSTM-based architectures. These models demonstrated high accuracy levels ranging between 88% and 91% in detecting unauthorized access and abnormal human behaviour. The incorporation of temporal learning enabled better understanding of motion patterns, improving overall detection reliability.

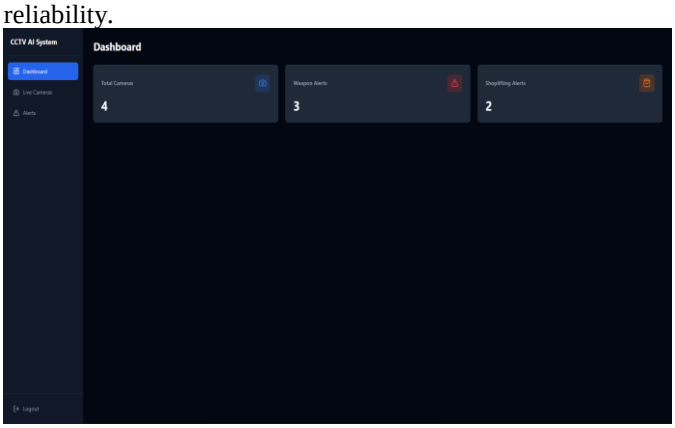

Figure 11: Frontend Interface

From a system perspective, the integration of all modules into a unified pipeline enabled **real-time monitoring with automated alert generation**. The developed dashboard provided live video feeds, detection logs, and instant notifications, significantly reducing dependence on manual monitoring. Compared to traditional CCTV systems, the proposed framework demonstrated improved response time, higher detection accuracy, and enhanced operational efficiency.

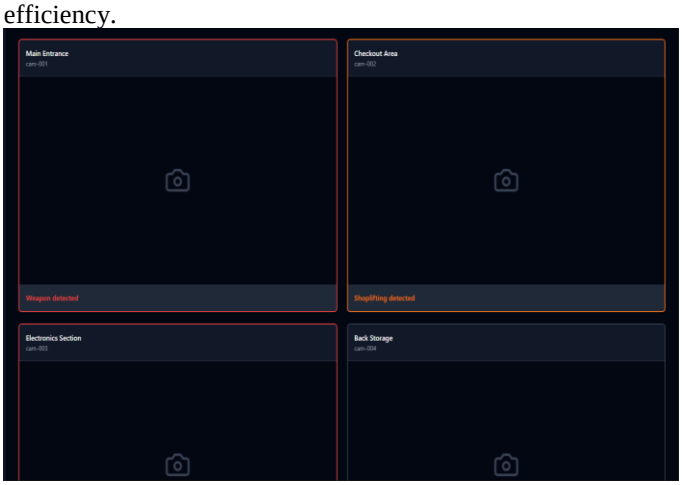

Figure 12: Multi screen Tracking Enabled

Performance analysis further indicated that while the model maintained high mAP values, fluctuations in precision during early training stages suggested the need for careful tuning of learning rates and optimization parameters. The confusion matrix analysis revealed improved class separation after training refinement, while loss curves confirmed stable convergence between training and validation phases.

Figure 13: Alert messages Triggered

Despite these advancements, certain limitations remain. The system performance can be affected by extreme lighting variations, camera angles, and highly dense crowds. Additionally, computational requirements for multi-model integration pose challenges for low-resource deployment environments.

Overall, the results validate that integrating deep learning with surveillance systems significantly enhances proactive security capabilities. The proposed system successfully transforms traditional CCTV into an intelligent, automated monitoring solution capable of real-time threat detection and response. Future improvements may focus on optimizing model efficiency, enhancing robustness in complex environments, and expanding multimodal capabilities such as audio integration and advanced behavioural analytics.

VI. CONCLUSION and FUTURE WORK

This study investigated the potential of machine learning to transform conventional CCTV systems into intelligent, proactive surveillance platforms. The primary objective was to design and evaluate a unified framework capable of automating video monitoring, detecting criminal activities such as theft, intrusion, and violence in real time, and generating immediate alerts for rapid response. Additionally, the research aimed to develop a scalable system adaptable to diverse environments, including retail stores, banks, residential areas, and educational institutions.

The findings demonstrate that the proposed intelligent CCTV system significantly enhances surveillance effectiveness by integrating multiple analytical layers within a single architecture. The use of YOLOv8 enabled accurate and high-speed detection of hazardous objects, while CNN-based behavioural models successfully identified suspicious activities such as shoplifting and abnormal human interactions. Furthermore, the inclusion of visual speech recognition introduced an additional contextual dimension, allowing the system to interpret potential verbal threats in the absence of audio input. Collectively, these components contributed to a robust, real-time monitoring system capable of reducing reliance on manual observation and improving response efficiency.

From a theoretical perspective, this study highlights the importance of multimodal integration in intelligent surveillance. Unlike traditional approaches that treat object detection, behaviour analysis, and contextual interpretation as separate tasks, the proposed framework demonstrates that combining these methods leads to improved accuracy and situational awareness. This layered analytical approach supports the development of more comprehensive and adaptive surveillance systems.

In practical terms, the system offers significant benefits for real-world security applications. By automating continuous monitoring and prioritizing critical alerts, it reduces cognitive load on human operators and minimizes the risk of missed incidents. The modular and scalable design further ensures adaptability across various operational domains, making it suitable for large-scale deployment.

However, certain limitations remain. The system's performance may vary under challenging real-world conditions such as poor lighting, occlusions, and dense crowds. Additionally, the integration of multiple deep learning models increases computational complexity, which may impact scalability in resource-constrained environments.

Future work should focus on improving model generalization through larger and more diverse datasets, as well as optimizing computational efficiency using edge computing and distributed processing techniques. The incorporation of additional modalities, such as audio analytics and IoT-based environmental sensing, could further enhance contextual awareness and detection accuracy. Moreover, continued attention to privacy, data security, and ethical considerations will be essential for responsible deployment.

In conclusion, this research demonstrates that machine learning–driven intelligent CCTV systems can effectively shift surveillance from a reactive to a proactive paradigm. By integrating spatial, temporal, and contextual analysis within a unified framework, the study advances the field of intelligent surveillance and provides a strong foundation for future developments in automated security systems.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to all individuals and institutions who contributed to the successful completion of this research. Special thanks are extended to the project guide and faculty members for their continuous support, valuable insights, and constructive feedback throughout the development of this work.

The authors also acknowledge the support provided by the institution in terms of resources, infrastructure, and technical guidance, which played a crucial role in the implementation and evaluation of the proposed system. Finally, appreciation is extended to peers and colleagues for their encouragement and assistance during various stages of this research.

Conflict of Interest Statement

The authors declare that there is no conflict of interest regarding the publication of this research paper.

REFERENCES

[1] Md. Afroza, E. Nyakwende, and B. Goswami, "Intrusion Detection in Smart Home Environments: A Machine Learning Approach," Transportation Research Procedia, 2025.

[2] J. Wang et al., "Restoring Speaking Lips from Occlusion for Audio-Visual Speech Recognition," in Proc. AAAI Conf. Artificial Intelligence, 2024.

[3] B. Hao et al., "LipGen: Viseme-Guided Lip Video Generation for Enhancing Visual Speech Recognition," arXiv preprint arXiv:2501.01234, 2025.

[4] P. Exarchos et al., "Lip-Reading Advancements: A 3D Convolutional Neural Network/Long Short-Term Memory Fusion for Precise Word Recognition," BioMedInformatics, 2024.

[5] N. Rashvand et al., "Exploring Pose-Based Anomaly Detection for Retail Security: A Real-World Shoplifting Dataset and Benchmark," in Proc. WACV Workshop on Computer Vision Applications, 2024.

[6] S. Jebur et al., "A Scalable and Generalised Deep Learning Framework for Anomaly Detection in Surveillance Videos," International Journal of Intelligent Systems, 2025.

[7] J. Rakesh Babu and K. Prasanthi, "U Smart Crowd and Crime Monitoring System Using Machine Learning," International Journal for Modern Trends in Science and Technology, 2025.

[8] R. Golande et al., "Weapon Detection System: Real-Time Object Recognition for Threat Detection," International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2025.

[9] P. Shanthi and V. Manjula, "Weapon Detection with FMR-CNN and YOLOv8 for Enhanced Crime Prevention and Security," Scientific Reports, 2025.

[10] N. C. Nagendrababu et al., "Smart Surveillance: Deep Learning-Based Weapon Detection Using YOLO," International Research Journal of Engineering and Technology (IRJET), 2025.

[11] K. Haritha et al., "Enhancing Public Safety with AI and ML-Based CCTV Surveillance," International Journal for Modern Trends in Science and Technology, 2025.

[12] D. Muronda and A. Ndlovu, "Smart Surveillance: Integrating Object Detection and Behavioural Analysis for Advanced Shoplifting Detection," International Journal of Computer Science and Mobile Computing, 2025.

[13] P. Siva et al., "Smart Surveillance Systems Using YOLOv8: A Scalable Approach for Crowd and Threat Detection," International Journal of Recent Advances in Engineering and Technology, 2025.