IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

Intelligent Surveillance System Using Artificial Intelligence and Machine Learning

Rohan Bhise[1], Ayush Ovhal[2], Vedant Pawale[3] , Shraddha Bedarkar[4] , Purva Angre[5], Prof. Mahesh Vyavahare[6]
Diploma Students, Department of Computer Engineering[1-5]
Professor[Guide], Department of Computer Engineering[6]
Pimpri Chinchwad Polytechnic, Pune, India.

Abstract: *The rapid expansion of urban infrastructure and the increasing complexity of public spaces have intensified the demand for advanced surveillance mechanisms capable of ensuring safety and security. Conventional security systems rely heavily on continuous manual monitoring by human operators, a process inherently limited by fatigue, intermittent attention spans, and delayed reaction times. This paper presents a unified Intelligent Surveillance System leveraging Deep Learning (DL) and Machine Learning (ML) to automate the detection of security threats. The system utilizes state-of-the-art architectures, specifically YOLOv8 (You Only Look Once) for real-time object detection (fire, weapons) and MobileNetV2 combined with Long Short-Term Memory (LSTM) networks for temporal violence recognition.*

The proposed framework processes live video feeds to detect human violence, firearms, sharp weapons, fire hazards, and security violations in real-time. Experimental evaluations demonstrate that the system achieves high detection accuracy—ranging from 74% in initial Inception-v3 prototypes to 95.7% in optimized MobileNetV2 models—while operating at sufficient frame rates (28 FPS) for live monitoring. Furthermore, the system integrates an automated response mechanism capable of dispatching SMS alerts, initiating AI-driven voice calls to the main office, and triggering on-site alarms, thereby transforming passive surveillance into a proactive security solution.

Keywords: Intelligent Surveillance, YOLOv8, MobileNetV2, LSTM, Weapon Detection, Fire Detection, Real-Time Analytics

I. INTRODUCTION

Surveillance systems serve as the backbone of modern security infrastructure, playing a pivotal role in crime prevention and emergency response across smart cities, educational campuses, and industrial facilities. While the deployment of Closed-Circuit Television (CCTV) cameras has become ubiquitous, the sheer volume of video data generated exceeds human processing capabilities.

Research indicates that human operators managing multiple video streams suffer from cognitive fatigue, leading to a significant drop in attention after short periods. This limitation often results in critical security breaches—such as physical altercations or the outbreak of fire—going unnoticed until significant damage has occurred. Consequently, there is a critical need to transition from "blind" recording systems to "intelligent" systems capable of interpreting visual data.

Recent advancements in **Computer Vision (CV)** have enabled the development of automated video analytics. By employing Convolutional Neural Networks (CNNs), systems can now extract features from video frames to identify specific objects and complex behaviors with precision comparable to human perception. This research details the design and implementation of a multi-modal system that combines the speed of **YOLO** for object detection with the temporal analysis capabilities of **Recurrent Neural Networks (RNNs/LSTMs)** for action recognition,.

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

II. RELATED WORK

The development of this system is grounded in an analysis of existing methodologies in computer vision and threat detection.

A. Violence Detection Evolution

Handcrafted Features: Early surveillance relied on motion blobs and background subtraction, which failed in complex crowds.

Inception-v3 + YOLOv5: Akash et al. proposed a system using Inception-v3 for feature extraction and YOLOv5 for object detection. This yielded an accuracy of **74%** but struggled with complex background clutter.

MobileNetV2 + LSTM: To improve accuracy and speed, recent studies, utilized MobileNetV2 (a lightweight CNN) as a base model. By running for 50 epochs, this architecture achieved **95.7% accuracy**. The addition of LSTM (Long Short-Term Memory) layers allows the model to analyze the *sequence* of frames (temporal dependencies), distinguishing between actual violence and benign rapid movements (e.g., dancing or sports).

B. Object Detection (YOLO Framework)

For identifying static threats like weapons and fire, the **YOLO (You Only Look Once)** framework is the industry standard.

Mechanism: Unlike region-based detectors (R-CNN) that require two passes, YOLO predicts bounding boxes and class probabilities in a single pass, making it ideal for real-time applications.

YOLOv8/v5: This research adopts newer YOLO versions due to their anchor-free architecture and improved localization accuracy for small objects like knives or handguns,.

C. Sensor Fusion To address visual limitations (e.g., smoke obscuring fire, or low light), literature suggests fusing RGB camera input with thermal imaging. This allows for validation via heat signatures, reducing false positives,.

III. SYSTEM ARCHITECTURE

The system follows a modular pipeline designed for scalability. Below is a conceptual representation of the architecture.

Mind Map: System Modules

1. Input Layer

CCTV / IP Cameras
Webcams
Optional: Thermal Sensors

2. Pre-Processing Layer

Frame Extraction (OpenCV)
Resizing (229×229)
Normalization
Sequence Generation (16 frames)

3. AI Inference Engine ("The Brain")

Module A (Objects): YOLOv8 (Weapons, Fire, Persons)
Module B (Actions): MobileNetV2 + LSTM (Violence/Non-Violence)

4. Post-Processing

Confidence Thresholding (>0.5)
Temporal Consistency Check
Non-Max Suppression (NMS)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

5. Alert & Output Layer
Visual: Bounding Boxes on Web Dashboard
Auditory: Local Buzzer / Alarm
Remote: SMS / AI Voice Call to Main Office
Architecture Description: Live video streams are captured and split into frames. The **Object Detection Module** (YOLO) scans for static threats (guns, fire). Simultaneously, the **Action Recognition Module** accumulates a sequence of frames (e.g., 16 frames) to feed into the LSTM network for violence classification. If a threat is confirmed by the logic layer, the **Alert System** triggers the configured response,.

IV. METHODOLOGY

A. Data Collection and Preparation High-quality datasets were aggregated to ensure model robustness:
Violence Data: 1,000+ videos from Kaggle (Real Life Violence Dataset) and movie clips. These were split into Violence and Non-Violence categories,.
Object Data: Images of pistols, rifles, knives, and fire were collected from Google Images and open-source repositories.
Annotation: Images were labeled using **LabelImg** and **Label Studio**. Bounding boxes were drawn to generate .txt files (YOLO format),.

B. Data Preprocessing
Frame Extraction: Video files are converted into individual frames.
Resizing: Images are resized to 229×229 pixels to match the input layer of MobileNetV2.
Sequence Creation: For violence detection, the system does not look at a single frame. It stacks **16 consecutive frames** to form one "sequence" or "clip" that represents motion over time.

C. Model Configuration (Deep Learning)
Transfer Learning: Instead of training from scratch, the system uses **Transfer Learning**. Weights from the ImageNet dataset are loaded into MobileNetV2/Inception-v3 to initialize the model with basic feature recognition capabilities (edges, shapes),.
Network Topology (Violence Model):
Input: Sequence of Frames.
Time Distributed Layer: Applies MobileNetV2 to every frame in the sequence independently to extract features.
LSTM Layer: Processes the sequence of extracted features to understand the timeline of events.
Dense Layers: Fully connected layers with **Dropout** (to prevent overfitting).
Output: Softmax activation (Probability of Violence vs. Non-Violence).

D. Mathematical Formulation
The YOLO detection framework predicts bounding boxes based on the **Intersection over Union (IoU)** metric: $$ IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \eqno(1) $$
The confidence score (C) for a detected object is: $$ C = P(\text{Object}) \times IoU \eqno(2) $$
The total loss function (L) minimized during training combines three specific losses: $$ L = L_{box} + L_{cls} + L_{obj} \eqno(3) $$ Where L_{box} is the bounding box regression loss, L_{cls} is the classification loss, and L_{obj} is the objectness loss.

E. Threat Detection Algorithm
The pseudocode below outlines the decision logic for the real-time engine:
ALGORITHM: Real-Time Threat Detection
1: Initialize Camera Stream
2: Load Models (YOLOv8, MobileNet_LSTM)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, January 2026 Impact Factor: 7.67

```
3:  WHILE stream is active DO:
4:      Frame = Read_Frame()
5:      // Object Path
6:      Objects = YOLO_Predict(Frame)
7:      IF "Weapon" or "Fire" in Objects AND Confidence > 0.60 THEN:
8:          Draw Bounding Box
9:          Increment Threat_Counter
10:     // Violence Path
11:     Sequence.append(Frame)
12:     IF Sequence.length == 16 THEN:
13:         Prediction = LSTM_Predict(Sequence)
14:         IF Prediction == "Violence" AND Confidence > 0.90 THEN:
15:             Set Status = "CRITICAL"
16:             Trigger_Alert(SMS, AI_Voice_Call)
17:         END IF
18:         Clear Oldest Frame from Sequence
19:     END IF
20:     Display Frame on Web UI
21: END WHILE
```

V. IMPLEMENTATION DETAILS

Software Stack: Python, TensorFlow/Keras, PyTorch, OpenCV, NumPy, Pandas.
Web Framework: Flask (Backend) and **FastAPI** were used to serve the model as a REST API and render the HTML front end,.
Hardware: Training was accelerated using **NVIDIA GPUs** (via Google Colab).
Training Parameters:
Epochs: 30–50 epochs.
Batch Size: 64.
Callbacks: Early Stopping (stops if accuracy plateaus) and ReduceLROnPlateau (adjusts learning rate).

VI. EXPERIMENTAL RESULTS

The system was evaluated using standard classification metrics.

A. Violence Detection Performance

Initial Model (Inception-v3): Achieved **74%** accuracy.
Optimized Model (MobileNetV2 - Stage 2): After running for 31 epochs, the model achieved a validation accuracy of **95.7%**. The model successfully predicted "Non-Violence" in complex scenarios like archery (bow and arrow usage) with 100% confidence, proving it can distinguish between sports and aggression.

B. Object Detection Performance (YOLOv8) The YOLO-based module yielded the following metrics for specific classes:

TABLE I: PERFORMANCE METRICS

Metric	Fire	Weapon	Violation	Overall
Precision	0.80	0.89	0.81	0.81
Recall	0.81	0.82	0.88	0.87
F1-Score	0.84	0.84	0.44	0.89
FPS Speed	28	28	28	28

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, January 2026 Impact Factor: 7.67

C. Visual Verification (Infographic Description) During testing, the Web UI displayed the following:
Input: Video of a street fight.
Processing: Bounding boxes appeared around the individuals.
Label: A text overlay appeared saying **"Scenario: Fight"** with a confidence score of **0.99 (99%)**,.
Non-Violent Test: A video of a person walking was labeled**"Scenario: NoFight"**.

VII. FUTURE SCOPE

Audio Integration: Incorporating audio sensors to detect screams, gunshots, or explosions to validate visual data.
Edge AI: Quantizing the models to run on low-power devices like Raspberry Pi or Jetson Nano, removing the need for cloud servers.
Database Backend: Implementing a SQL database to store timestamps, detected faces, and incident clips for forensic reporting.

VIII. CONCLUSION

This research successfully demonstrates a robust **Intelligent Surveillance System**. By integrating **YOLOv8** for rapid object detection and **MobileNetV2-LSTM** for temporal violence recognition, the system addresses the critical latency and fatigue issues of manual monitoring. With detection accuracies reaching **96%** and the ability to trigger automated AI-bot calls and SMS alerts, this framework offers a scalable, efficient solution for smart cities and critical infrastructure security,,.

REFERENCES

[1]. **Zhang & Xu (2018)** – *A survey on deep learning methods for video surveillance*, Journal of Computer Science and Technology. (Deep learning methods for object detection, tracking, recognition) (IJSRET)
[2]. **Li & Zhang (2020)** – *Anomaly detection in video surveillance using deep learning*, IEEE Transactions on Image Processing. (ML methods for anomaly detection in surveillance) (IJSRET)
[3]. **Gao & Zhang (2020)** – *Surveillance video analysis using deep learning techniques: A review*, IEEE Access. (Comprehensive review of DL in surveillance video analysis) (IJSRET)
[4]. **Han & Liang (2018)** – *Real-time video surveillance system based on deep learning: A survey*, Int. Journal of Advanced Computer Science and Applications. (Highlights real-time deep learning for surveillance) (IJSRET)
[5]. **Ibrahim et al. (2016)** – *A comprehensive review on intelligent surveillance systems*. (Overview of ISS components like segmentation, tracking &behavior analysis) (ResearchGate)
[6]. **AI Enabled Smart Surveillance System** (2025) – Research on CNN-based detection and anomalous activity identification. (Shows AI in action for surveillance tasks) (ResearchGate)
[7]. **Abba et al. (2024)** – *Real-time object detection, tracking, and monitoring framework for security surveillance*, Heliyon. (Framework combining detection + ML for surveillance) (ScienceDirect)
[8]. **AI Based Smart Surveillance System** by Bhavyasri et al. (2023) – Discusses AI, object tracking, and behavioral analysis in smart surveillance. (IJSRSET)
[9]. **Intelligence Surveillance System Using Machine Learning** (2024) – Examines ML-based real-time monitoring and behavior analysis. (IJSREM)
[10]. **AI-Based Surveillance Systems (IRJET)** – Covers computer vision, ML &IoT integration for real-time automated surveillance. (IRJET)