iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80717

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Implementation of an Intelligent Multi-Scenario Surveillance System for Violence and Cheating Detection in Educational Institutions

Prof. Rushikesh S. Bhalerao[1], Mr. Rahul D. Kakad[2], Ms. Chetana P. Suryavanshi[3], Mr. Devidas K. Tambe[4], Mr. Dinesh R. Thorat[5]
[1,2,3,4,5]*Department of Information Technology, Sir Visvesvaraya Institute of Technology, Savitribai Phule Pune University*

Abstract: ***Maintaining safety within public areas like schools and ensuring fairness in examination halls are of paramount importance in the modern educational landscape. Traditional surveillance systems depend heavily on manual monitoring, which is inherently prone to human error, fatigue, and causes critical delays in identifying sporadic events. This paper presents the implementation of a comprehensive, multi-scenario automated surveillance system that leverages advanced deep learning and computer vision techniques to actively analyze closed-circuit television (CCTV) feeds. We categorize our approach into two distinct routing pipelines: a campus safety mode utilizing YOLOv7 for rapid human detection combined with a MobileNet-BiLSTM classifier for spatiotemporal violence recognition , and an academic integrity mode utilizing an improved SE-YOLOv8 model coupled with a ResNet 3D Convolutional Neural Network (CNN) for subtle cheating detection. By employing Squeeze Aggregated Excitation (SaE) modules and Deep Keyframe Detection, the proposed framework minimizes computational overhead. We analyze benchmark datasets, evaluation metrics, and overall performance, demonstrating that this hybrid framework achieves a 91% accuracy in violence detection and a 96.0% mean Average Precision (mAP) in cheating recognition. The deep learning backend is further integrated with a full-stack administrative dashboard and automated alerting mechanism, bridging the gap between theoretical computer vision models and deployable institutional needs.***

Keywords: Multi-Scenario Surveillance, Deep Learning, Violence Detection, Smart Proctoring, YOLOv7, SE-YOLOv8, MobileNet-BiLSTM, ResNet 3D CNN, Action Recognition.

I. INTRODUCTION

Digital video surveillance serves as the primary mechanism for evidence collection and real-time security in educational institutions. Public violence cases and academic misconduct are emerging threats that require immediate intervention. However, the vast majority of security systems installed at these locations rely entirely on closed-circuit video surveillance systems that depend exclusively on human observation.

This traditional mode of observation leads to severe operational bottlenecks. Human monitors are susceptible to attention fatigue, subjective judgment variations, and the inability to simultaneously process multiple high-definition feeds. This results in a situation of "monitoring without detailed inspection," where rapid, sporadic acts of violence or subtle instances of academic dishonesty go completely unnoticed until post-incident investigations are launched. Furthermore, when incidents of public violence occur, a delayed response can lead to significant harm to individuals and property.

To overcome these critical limitations, this study proposes a dual-model-based artificial intelligence framework designed to navigate the intricacies of different examination and campus environments. The opportunity to address these challenges emerges with the fast growth of machine learning, allowing for the transition from passive video recording to active, intelligent threat detection. This paper details the complete implementation of a multi-scenario system that dynamically routes video frames based on their contextual environment, utilizing optimized YOLO (You Only Look Once) architectures and temporal classifiers to ensure both high-speed frame processing and highly accurate action recognition.

II. TAXONOMY OF SURVEILLANCE DETECTION TECHNIQUES

Research in the field of automated surveillance has evolved significantly over the past two decades, transitioning from traditional computer vision methods to advanced deep learning architectures.

A. *Traditional (Hand-Crafted Feature) Methods*

Early automated systems relied on basic motion detection, background subtraction, and optical flow algorithms. While these methods are computationally inexpensive and simple to implement, they struggle heavily with dynamic lighting, occlusions, and varying crowd densities. They depend on handcrafted features which fail to capture the complex contextual nuances of human interactions, leading to unacceptably high false alarm rates in busy campus environments.

TABLE I
SUMMARY OF TRADITIONAL VS. DEEP LEARNING METHODS

Method Type	Feature Used	Strength	Limitation
Traditional Vision	Background Subtraction, Optical Flow	Low computational requirement	High false alarm rate in dynamic environments
CNN-based Spatial	Xception, ResNet	High spatial accuracy	Lacks temporal motion understanding
Temporal Models	LSTM, 3D CNN	Excellent motion/action analysis	High computational cost
Hybrid (YOLO + RNN)	YOLOv7 + BiLSTM	Balance of speed and temporal accuracy	Requires careful hyperparameter tuning

B. *Deep Learning-Based Spatial and Temporal Approaches*

The introduction of Convolutional Neural Networks (CNNs) revolutionized object detection by automatically learning hierarchical spatial features. However, standard 2D CNNs evaluate single frames in isolation, discarding the crucial temporal data required to understand a sequence of movements. To address this, researchers integrated Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which analyze frame-to-frame coherence to detect manipulations or specific actions. While effective, pure temporal models are computationally expensive. Therefore, a hybrid approach that utilizes a rapid spatial detector followed by a lightweight temporal classifier represents the state-of-the-art optimal balance for edge device deployment

III.RELATED WORK

A. *Violence Detection Approaches*

Recent studies on violence detection systems have utilized various deep learning models like YOLOv7, SSD, and Faster R-CNN. While Faster R-CNN achieves high accuracy, it does so at the expense of speed, making it computationally heavy and unsuitable for real-time edge devices. In contrast, YOLOv7 has demonstrated exceptional capabilities in addressing obstacles in real-time object detection, such as recognizing small objects and optimizing gradient paths through its extended layer aggregation networks (E-ELAN). However, YOLOv7 alone cannot classify complex temporal actions. Consequently, integrating YOLOv7 with a lightweight classifier like MobileNet ensures rapid human detection alongside efficient violent action recognition.

B. Cheating Detection Approaches

Existing smart proctoring technologies often rely on singular object detection algorithms, which are inadequate in addressing the complex conditions of examination environments where examinees are densely seated. Furthermore, traditional 3D convolutions offer higher accuracy than 2D methods but present large parameter sizes and substantial computational overhead. To ensure a highly responsive framework, recent methodologies emphasize incorporating Multilayer Perceptrons (MLP) into the detection network to enhance feature representation, followed by a hierarchical approach combining 2D and 3D convolutions to ensure both detection speed and accuracy.

IV. SYSTEM ARCHITECTURE AND METHODOLOGY

The implemented architecture acts as an intelligent router, categorizing video feeds and applying the appropriate deep learning pipeline to minimize computational waste. The system is divided into three primary modules.

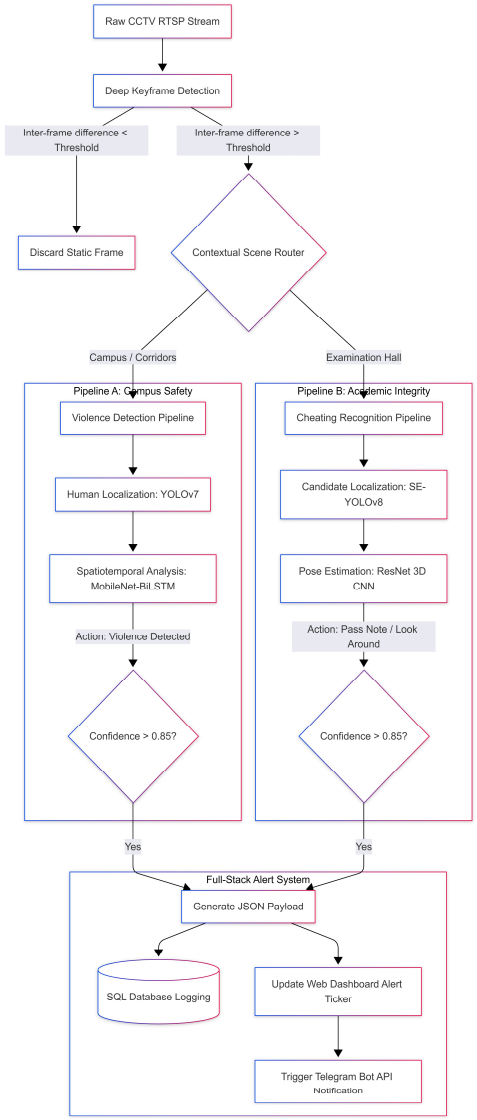


Fig 1. Proposed multi-scenario system architecture illustrating the data flow from video input to the final full-stack alert mechanism.

A. Deep Keyframe Detection and Preprocessing

Continuous CCTV footage contains massive amounts of redundant data. For a standardized examination, candidates typically exhibit minimal movement for the majority of the time, resulting in small inter-frame differences. Analyzing every frame would lead to redundant information and processing delays. However, cheating behaviors and violent altercations often involve rapid, discreet, or exaggerated movements, leading to significant differences between consecutive frames.

The system employs frame differencing to extract the contours of moving objects. By subtracting one frame from another and calculating the absolute difference of pixel values at corresponding positions, the algorithm determines if the difference exceeds a predefined threshold. A scatter plot based on the time series is utilized, where frames with local maximum values in inter-frame difference intensity are selected as keyframes, while static frames are discarded.

B. Violence Detection Pipeline (Campus Mode)

When deployed in corridors or public spaces, the system activates the violence detection sub-system.

1) *Human Detection (YOLOv7):* YOLOv7 is employed due to its ability to achieve a good trade-off between speed and accuracy. The algorithm first scans the keyframes, discarding those where no human presence is identified. The E-ELAN architecture preserves computational efficiency while providing robust detection in dynamic, cluttered environments.
2) *Pose Estimation (ResNet 3D CNN):* The localized spatial data is subsequently analyzed using a ResNet architecture enhanced with 3D convolutions. This hierarchical approach compensates for the lack of temporal information in 2D convolutions. By processing the spatiotemporal sequence of the candidate's posture, the system precisely categorizes actions such as leaning forward, looking around, or passing notes.

C. Cheating Recognition Pipeline (Exam Mode)

When monitoring examination halls, the system requires granular analysis to detect subtle anomalies.

1) *Enhanced Localization (SE-YOLOv8):* The baseline YOLOv8 algorithm was structurally improved using Multilayer Perceptron principles. Specifically, the C2f module in the neck layer of YOLOv8 was replaced with a new module combining the characteristics of C2f and SENetV2, named C2f_SENetV2. This incorporates a Squeeze Aggregated Excitation (SaE) module, which applies global average pooling to "squeeze" features, computes channel weights via fully connected layers, and adaptively scales the convolutional features . This SE-YOLOv8 variant significantly boosts the algorithm's capacity for meticulous candidate localization.
2) *Pose Estimation (ResNet 3D CNN):* The localized spatial data is subsequently analyzed using a ResNet architecture enhanced with 3D convolutions. This hierarchical approach compensates for the lack of temporal information in 2D convolutions. By processing the spatiotemporal sequence of the candidate's posture, the system precisely categorizes actions such as leaning forward, looking around, or passing notes.

V. BENCHMARK DATASETS AND EXPERIMENTAL SETUP

Evaluating these models requires diverse datasets simulating real-world anomalies. The model training and evaluation were executed utilizing an Intel Core i7-7700K CPU (4.20GHz) paired with an NVIDIA GTX 1080Ti GPU. The software stack was built on Ubuntu 16.04, utilizing Python 3.8, OpenCV 4.8.1, and PyTorch 2.1.1 running on CUDA 7.5.

A. Violence Dataset

For the public safety pipeline, a balanced dataset comprising 1000 non-violent and 1000 violent videos was utilized. Data augmentation procedures were implemented to generate variations in terms of illumination, viewpoint, and crowd density, thereby improving model generalization across complex environments.

B. Cheating and Normal (CAN) Dataset

Due to privacy concerns and the classified nature of examination footage, a custom simulated dataset was constructed. High-fidelity replication of exam settings was achieved using 4K cameras under varied lighting and angles. A rigorous sampling technique extracted every twelfth frame, amassing a dataset of 17,000 images. The dataset represented standard behavior alongside four specific annotated cheating behaviors: looking forward, looking back, looking around, and passing notes. Additional simulations introduced visual distortions like blur, low-light, and occlusions to mimic real-world challenges. The dataset was split into training, validation, and test sets in a ratio of 7:2:1, and the model was trained for 20 epochs with a batch size of 256 and a learning rate of 0.01.

TABLE II
DATASETS UTILIZED FOR MODEL TRAINING

Dataset	Type	Samples	Strength	Limitation
Violence Dataset	Public Altercations	1,000 non-violent, 1,000 violent videos	Balanced classes, real-world scenarios	Limited extreme low-light samples
CAN Dataset	Exam Behaviors	17,000 images	Contains normal actions and 4 common cheating behaviors	Requires specific lighting configurations

VI. EVALUATION METRICS

To accurately assess the multi-scenario system, standard classification metrics were employed. The evaluation relies on the calculation of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).
Precision measures the ratio of correctly predicted positive observations to the total predicted positives, which is crucial for controlling false alarms in automated alerting systems:

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

Recall calculates the ratio of correctly predicted positive observations to all actual positive observations, defining the detection completeness:

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

The F1-Score provides the harmonic mean of Precision and Recall, ensuring a balanced performance evaluation even if class distributions are uneven:

$$F1 = 2 \times \text{Precision} \times \text{Recall} / \text{Precision} + \text{Recall}$$

Mean Average Precision (mAP) is also utilized, specifically mAP_{50}, which calculates the average precision across all classes at an Intersection over Union (IoU) threshold of 0.50

VII. RESULTS AND PERFORMANCE ANALYSIS

The system was evaluated on its ability to balance processing speed with classification accuracy.

A. Violence Detection Performance

The integration of YOLOv7 significantly improved real-time processing capabilities compared to baseline models. The YOLOv7 algorithm demonstrated an average processing time of 0.0308 seconds per frame. This was substantially faster than the SSD MobileNet model (0.1015 seconds) and the TensorFlow Faster RCNN model (1.4498 seconds). In terms of total detections in the benchmark video, YOLOv7 successfully identified 6,262 instances, drastically outperforming SSD MobileNet (1,092) and TensorFlow (3,204).
The combined YOLOv7 and MobileNet-BiLSTM architecture achieved an overall Accuracy Score of 0.91. The confusion matrix exhibited a balanced detection of both behavior types, recording 92 true positives for non-violence and 90 true positives for violence, while maintaining low false positive and false negative rates. Specifically, the model achieved a precision of 0.93 for violent instances, indicating that 93% of the instances predicted as violence were correctly assessed.

TABLE III
HUMAN DETECTION SPEED COMPARISON

Model Used	**Time Taken to Detect Object**	**Total Objects Detected**
YOLOv7 (Proposed)	0.0308 seconds	6,262
SSD MobileNet	0.1015 seconds	1,092
Faster RCNN	1.4498 seconds	3,204

B. Heating Behavior Recognition Accuracy

For the academic integrity scenario, the SE-YOLOv8 combined with the ResNet 3D CNN yielded a high mean Average Precision (mAP_{50}) of 96.0% across all 2000 test examples. The system showed exceptional accuracy in identifying distinct physical interactions. The behavior "pass note" achieved the highest precision at 98.7% with a recall of 95.8%. Normal behavior was accurately classified with a 96.7% precision. Detecting highly ambiguous movements like "look around" presented the greatest challenge, yielding a slightly lower precision of 78.8%, though it maintained a strong recall of 92.2%. The overall model average confirmed an accuracy of 91.6%.

TABLE IV
CHEATING BEHAVIOR RECOGNITION ACCURACY

Behavior Class	Precision (%)	Recall (%)	mAP@50 (%)
Normal Behavior	96.7	85.0	97.7
Look Forward	96.2	95.5	97.9
Look Back	87.7	90.1	94.9
Look Around	78.8	92.2	90.1
Pass Note	98.7	95.8	99.4
Overall Model	91.6	91.7	96.0

VIII. FULL-STACK SOFTWARE INTEGRATION

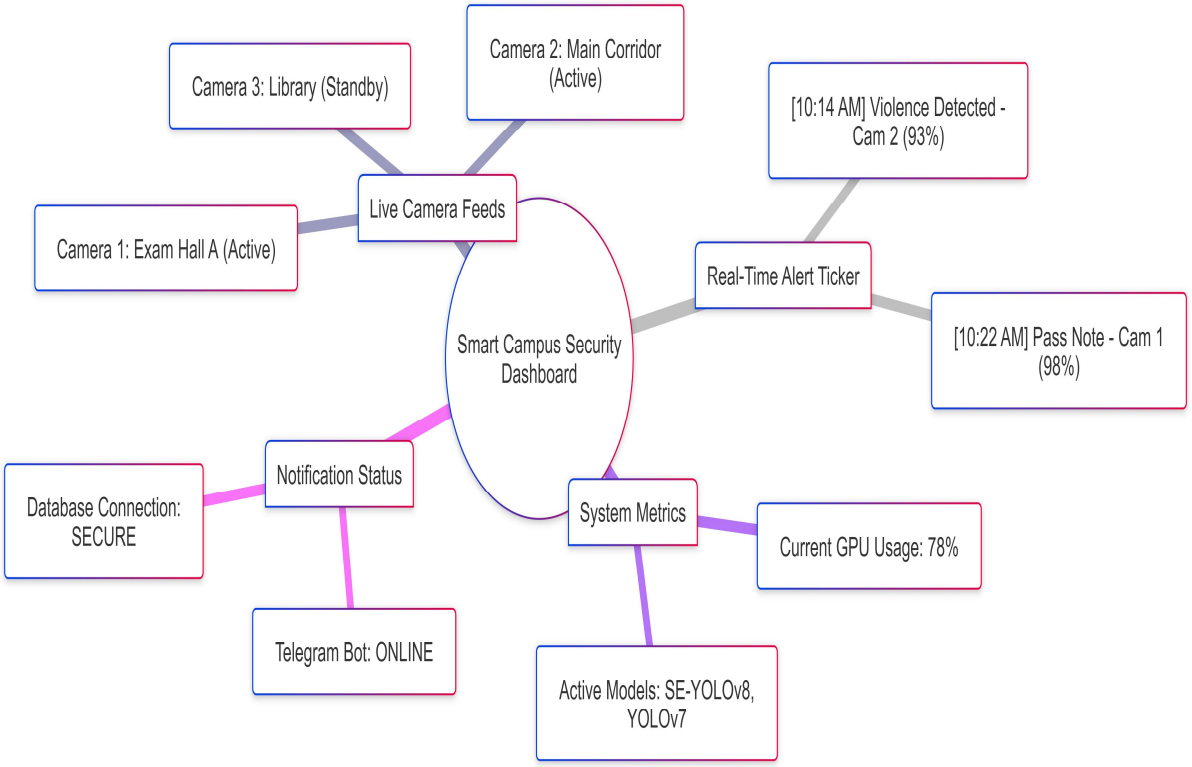


Fig 2. Conceptual layout of the administrative web dashboard for real-time institutional monitoring.

To transition this research from a theoretical deep learning framework to a functional institutional product, the analytical Python backend was seamlessly integrated with a centralized administrative web dashboard.

The inference engines run continuously, processing the RTSP video streams from the institutional CCTV network. When the confidence score of a detected anomaly (violence or cheating) exceeds the designated threshold, the deep learning pipeline generates a JSON payload. This payload contains critical metadata: the unique camera ID, a highly precise timestamp, the classified action label, and the bounding box coordinates.

This data is transmitted via RESTful APIs to a Spring Boot backend, which logs the incidents into a relational SQL database for historical auditing and evidence preservation. Simultaneously, a responsive frontend—built using modern JavaScript frameworks (e.g., React or Angular)—updates dynamically. The dashboard provides security personnel with live video feeds alongside a real-time event ticker. To ensure rapid response times when administrators are away from the dashboard, the system incorporates an automated alert mechanism that utilizes third-party messaging integrations, such as a Telegram bot, to send immediate push notifications directly to the mobile devices of relevant authorities.

IX. MAJOR CHALLENGES AND RESEARCH GAPS

While the proposed system achieves high accuracy, deploying AI surveillance in the real world presents several ongoing challenges:

1) *Computational Overhead vs. Accuracy:* While 3D convolutions offer superior accuracy, they present large parameter sizes. Balancing this on resource-constrained edge devices remains a significant engineering hurdle.
2) *Dataset Bias and Diversity:* The current systems rely on laboratory-curated datasets. There are limitations in dataset scale and behavior diversity due to practical constraints and privacy considerations.
3) *Visual Distortions:* The existing system suffers from various limitations, including false alarms in highly complex environments and the inability to handle low light or extreme occlusion states.

TABLE V
RESEARCH GAP ANALYSIS AND PROPOSED SOLUTIONS

Gap	Explanation	Research Opportunity (Our Approach)
Single-Domain Focus	Most models only detect one anomaly (e.g., *only* cheating)	Multi-Scenario Routing: Dynamic switching between Exam and Campus pipelines.
Real-time Latency	High computational cost delays alerts	Hybridization: Using YOLOv7 (0.03s/frame) + lightweight MobileNet.
Action Confusion	2D CNNs fail to understand temporal movements	3D CNN Integration: Utilizing ResNet 3D CNNs to capture time-series poses.

X. EMERGING RESEARCH DIRECTIONS

Future iterations of this framework will focus on optimizing the existing architecture and exploring new modalities.

1) *Lightweight Architectures:* Further research into ultra-lightweight classifiers like EfficientNet-B0 is necessary for the optimization of computational efficiency, leading to higher usability on ultra-low-power and edge devices.
2) *Graph Neural Networks (GNNs):* Advanced techniques such as GNNs can be explored to model complex spatial relationships between individuals in densely populated, crowded spaces.
3) *Multimodal Learning:* Future improvements can be achieved by incorporating multi-modal learning methods where video data will be combined with audio input for better detection performance in cases where relying on only visual cues may not work.

XI. CONCLUSIONS

Surveillance systems in educational institutions must transition from passive recording to active, intelligent monitoring to ensure the safety of students and the integrity of academic processes. This paper presented the comprehensive implementation of a dual-pipeline, multi-scenario framework designed to efficiently categorize and analyze video feeds. By utilizing Deep Keyframe Detection to minimize computational waste, followed by a YOLOv7 and MobileNet-BiLSTM combination, the system achieved a 91% accuracy in violence detection with rapid frame processing speeds. Concurrently, the MLP-enhanced SE-YOLOv8 combined with a ResNet 3D CNN successfully mitigated temporal confusion, achieving a 96.0% mAP_{50} for subtle cheating recognition.

The integration of these deep learning models with a robust full-stack administrative dashboard and automated Telegram alerting mechanism elevates this research into a practical, deployable product. This unified approach successfully balances computational efficiency with high detection accuracy, providing a modern, automated solution that significantly reduces the reliance on manual human monitoring in educational environments.

XII. ACKNOWLEDGMENT

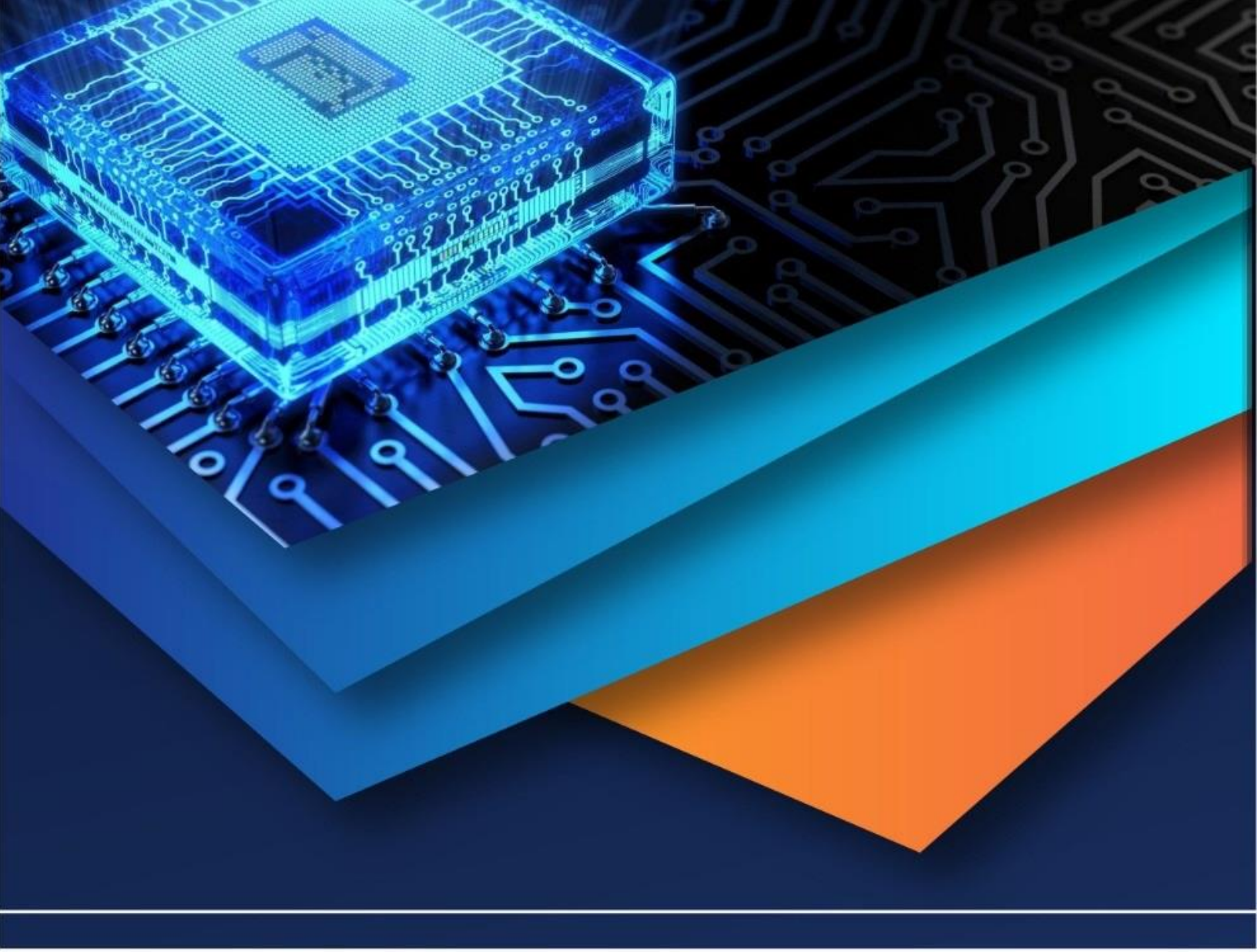
We would like to express our profound gratitude to our Project Guide, **Mr. Rushikesh S. Bhalerao**, for their valuable guidance, continuous support, and constructive feedback throughout the conceptualization and implementation of this research. Their expertise in deep learning and computer vision was instrumental in shaping our multi-scenario surveillance architecture.

We also extend our sincere thanks to Dr. Pratibha V. Kashid, Head of the Information Technology Department, and Dr. Sarang Pande, Principal of Sir Visvesvaraya Institute of Technology, for providing the necessary facilities, computational resources, and a conducive environment to carry out this work.

Finally, we are thankful to Savitribai Phule Pune University for structuring a curriculum that fosters practical research, technological innovation, and academic growth. We also acknowledge the broader open-source AI community for providing the foundational frameworks that made this project possible.

REFERENCES

[1] S. Senthilkumar, S. Kolte, G. Agarwal, and A. Shirish, "Real Time Violence Detection System using YOLOv7 and Deep Learning Techniques," 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), IEEE, 2025, pp. 1447-1454.
[2] J. Lu, J. Wang, N. Song, Z. Luo, W. Zhang, and Y. Wang, "Cheating Recognition in Examination Halls Based on Improved YOLOv8," 2024 International Conference on Artificial Intelligence of Things and Systems (AIoTSys), IEEE, 2024.
[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.
[5] M. Narayanan, "SENetV2: Aggregated dense layer for channelwise and global representations," arXiv preprint arXiv:2311.10807, 2023.
[6] X. Yan, S. Z. Gilani, H. Qin, M. Feng, L. Zhang, and A. Mian, "Deep keyframe detection in human action videos," arXiv preprint arXiv:1804.10021, 2018.
[7] R. Kumar, A. Gupta, and D. Rajeswari, "Violence Detection System using MobileNetV2," 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), IEEE, 2024, pp. 1555-1560.
[8] L. A. Siddique, R. Junhai, T. Reza, S. S. Khan, and T. Rahman, "Analysis of Real-Time Hostile Activity Detection from Spatiotemporal Features Using Time Distributed Deep CNNs, RNNs and Attention-Based Mechanisms," arXiv preprint arXiv:2302.11027, 2023.
[9] L. Rahmawati, S. Rustad, A. Marjuni, M. A. Soeleman, and P. N. Andono, "Foggy-Based Object Detection In Video Using Faster R-CNN, YOLOv3, and SSD," 2023 International Seminar on Application for Technology of Information and Communication (iSemantic), IEEE, 2023, pp. 412-416.
[10] M. Malhotra and I. Chhabra, "Automatic invigilation using computer vision," International Conference on Integrated Intelligent Computing Communication & Security, Atlantis Press, 2021, pp. 130-136.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY