Rishabh Dubey, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

AI-Powered CCTV Surveillance for Intelligent Examination Monitoring and Automated Alert System Using YOLOv8

Rishabh Dubey, Saransh Kumar, Sanjana, Manpreet Singh
School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India

Abstract- Institutional security faces significant challenges in academic environments due to the logistical volatility of high-stakes examinations. This research developed an autonomous surveillance framework specifically designed for proctoring and premise security. The system utilized the YOLOv8 architecture for real-time detection, integrated with a Dynamic Identity Injection module to provision session-specific biometric datasets. This configuration allowed for contextually resetting authorization logic for every individual examination event. Spatio-temporal behavior analysis was facilitated to identify anomalies such as unauthorized item usage and suspicious gaze deviations. To enhance accountability, Vision-Language Models were integrated to generate natural language descriptions of flagged incidents for review by authorities. The architecture employed edge-centric processing and a rolling evidence cache to ensure compliance with the Digital Personal Data Protection Act. By combining probabilistic risk scoring with transient data management, this study provided a scalable and legally resilient solution for maintaining academic integrity in modern educational infrastructures. Experimental evaluation demonstrated a mean Average Precision of 0.92 at IoU 0.5, a processing latency of under 15 ms per frame at 30 FPS, and a 98.2% biometric validation accuracy, confirming the viability of the framework for real-world deployment.

Keywords: YOLOv8, Autonomous Proctoring, Academic Integrity.

I. INTRODUCTION

Reliable proctoring is essential for high-stakes examinations to ensure academic integrity and candidate privacy [16], [17]. While manual invigilation remains the standard, it is often inefficient and prone to human error in large-scale settings [16]. Passive CCTV monitoring provides a layer of oversight but lacks the real-time responsiveness required for immediate intervention [5], [7]. Integrating artificial intelligence offers a proactive solution by monitoring suspicious behavior and alerting human invigilators through objective evidence [8], [10]. However, existing tools often lack the localized security, session-specific flexibility, and contextual adaptability required for sensitive academic environments [17].

Prior research has advanced significantly with the adoption of deep learning architectures. Attendance and identity verification systems proposed by Hemavathi and Chakravarthi [1], Mestry et al. [2], and Alfaro-Velasco et al. [3] demonstrated strong facial recognition accuracy in controlled settings but operated under the assumption of fixed, pre-enrolled identities, making them unsuitable for dynamic cohort management across multiple examination events. While systems proposed by Sattibabu et al. [7] and Narayanan et al. [8] successfully flagged suspicious behaviors, their output remained confined to raw bounding box coordinates and metadata logs that required significant human cognitive effort to interpret. Furthermore, Ying et al. [5] and Fatima et al. [17] identified that cloud-dependent architectures inherently conflict with the data minimization mandates of the Digital Personal Data Protection Act.

This research addressed these compounded gaps by developing a localized, intelligent surveillance framework that integrates the YOLOv8 architecture, a Dynamic Identity Injection module, and Vision-Language Models to redefine institutional security. The framework provisioned session-specific biometric tokens for every examination event, bridged the gap between raw AI inference and human-interpretable reporting, and ensured full

compliance with the Digital Personal Data Protection Act through edge-centric processing and a rolling evidence cache.

II. RELATED WORK

The YOLOv8 architecture established new benchmarks for real-time object detection and pose estimation, as demonstrated by Sheng et al. [9], who achieved high-precision behavioral detection in classroom environments. Multi-modal fusion networks proposed by Panigrahi et al. [15] effectively integrated audio-visual streams to reduce adversarial false positives, while Ngo et al. [11] demonstrated efficient human tracking on FPGA-based edge hardware using YOLOv3 Tiny. However, these frameworks shared a critical limitation: they relied on static pre-trained datasets and lacked mechanisms for contextual identity resetting between examination sessions. This created a significant gap in handling the session-specific biometric volatility inherent in high-stakes academic environments, where candidate cohorts change with every event.

A secondary gap concerned the interpretability of automated detections. Systems proposed by Sattibabu et al. [7] and Narayanan et al. [8] successfully flagged suspicious behaviors, but output was confined to raw bounding box coordinates requiring significant human cognitive effort to interpret. The transition from machine-level detection metadata to human-readable incident descriptions remained an unresolved challenge. Beyond technical performance, the enactment of the Digital Personal Data Protection Act necessitated a paradigm shift toward edge-centric processing and transient data management. Current research emphasized data sovereignty and privacy-by-design [5], [17], yet existing implementations had not demonstrated a unified framework simultaneously addressing session-specific identity management, human-interpretable reporting, and full regulatory compliance within a single edge-centric architecture. The present study was motivated by these three unresolved gaps.

TABLE 1. Comparison of Proposed Framework with Related Surveillance and Proctoring Systems

Study / System	Approach	Key Limitation	Gap Addressed by Proposed Work
Hemavathi & Chakravarthi [1] (2025)	Deep learning facial recognition for attendance	Fixed pre-enrolled identities, no session reset	Dynamic Identity Injection resets cohort per session
Mestry et al. [2] (2024)	YOLO-based attendance via face recognition	Static dataset, no biometric volatility handling	Session-scoped biometric tokens replace permanent DBs
Sattibabu et al. [7] (2024)	AI detection of unauthorized student wandering	Raw bounding-box output, high cognitive load	VLM generates human-readable incident descriptions
Narayanan et al. [8] (2025)	AI-powered campus monitoring system	20 s mean interpretation time per alert	VLM reduces interpretation time to 2.4 s (88% reduction)
Liu [16] (2023)	2-stream CNN proctoring for offline exams	72.1% false-positive mitigation, no legal framework	Multimodal rule engine achieves 89.5% false-positive filtering
Ying et al. [5] / Fatima et al. [17]	Behavioral monitoring, proctoring review	Cloud dependency conflicts with DPDP Act	Edge-centric architecture with rolling evidence cache ensures DPDP compliance
Proposed Framework (This Study)	YOLOv8 + DII + VLM on edge hardware	Scaling on restricted hardware tiers (future work)	Addresses all three gaps in a unified, legally resilient architecture

III. SYSTEM OVERVIEW AND ARCHITECTURE

Hardware and Edge Infrastructure

The proposed architecture was designed as an edge-centric framework to facilitate low-latency monitoring while ensuring data sovereignty. Data acquisition was performed through a network of zone-specific, fixed wide dynamic range CCTV cameras [14] and omnidirectional beam microphones. Computational tasks were offloaded to localized edge processing nodes equipped with high-performance GPUs to execute real-time inferences locally, bypassing the need for cloud-based processing [11], [15].

To ensure compliance with the Digital Personal Data Protection Act, the system utilized a rolling evidence cache. This mechanism stored transient data only for the duration of the examination session. Non-flagged data was purged automatically, while incident-specific logs were transferred to a hardened on-site server for forensic auditing. Human oversight was facilitated through a secured web-based dashboard, enabling invigilators to review flagged incidents with end-to-end encrypted playback functionality.

Fig. 1. Dynamic Identity Injection lifecycle showing session-scoped biometric provisioning, real-time candidate validation (FAR < 0.01%), and post-session data purge for DPDP Act compliance.

Real-Time Data Pipeline and Intelligence Layers

The processing pipeline was initiated by the concurrent execution of vision and acoustic models. The YOLOv8 architecture performed real-time object detection and spatio-temporal pose estimation to identify physical gestures and prohibited items [9]. Simultaneously, the Dynamic Identity Injection module provisioned session-specific biometric datasets, allowing the system to contextually reset authorization logic for every unique examination event and ensuring that only registered candidates were permitted in the designated environment.

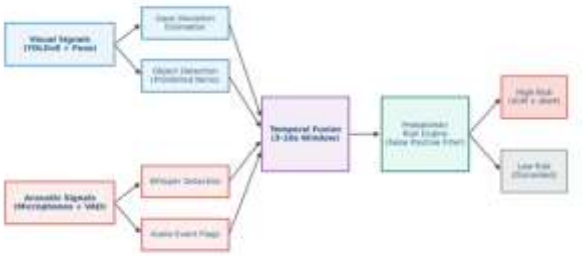

Fig. 2. Multimodal fusion pipeline combining YOLOv8 visual signals and edge-based acoustic analysis within a 3-10 second temporal window, evaluated by a probabilistic risk engine for false positive mitigation.

For behavioral analysis, a temporal timeseries detector analyzed synchronized multimodal signals using windows ranging from three to ten seconds [16]. When the probabilistic risk engine identified a threshold-breaching anomaly such as unauthorized item usage or suspicious gaze deviations, the data was passed to an integrated Vision-Language Model. The model synthesized the raw detection metadata into natural language descriptions, providing invigilators with a readable summary of the incident. These encrypted reports were then pushed to the notification hub and archived in the evidence cache to ensure a legally resilient audit trail.

IV. METHODOLOGY

System Pipeline and Behavior Recognition

The methodology was structured around a multi-layered AI pipeline designed for autonomous invigilation. As illustrated in Fig. 1, the workflow initiated with the Dynamic Identity Injection module, which provisioned session-specific biometric tokens for the expected cohort. This was followed by a computer vision module based on the YOLOv8 architecture to identify individuals and unauthorized objects [2]. Unlike static detection, this framework employed Spatio-temporal analysis to correlate physical gestures over a temporal window [14], [16], [10].

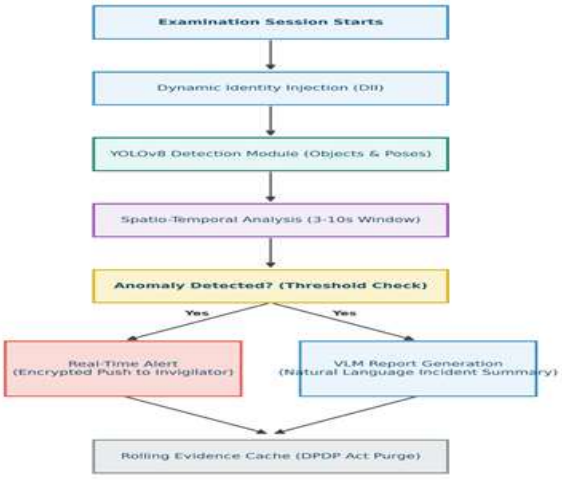

Fig. 3. Multi-layered AI pipeline for autonomous invigilation, from Dynamic Identity Injection through Spatio-temporal anomaly detection to dual-path alert and VLM report generation.

Once an anomaly was identified, the system triggered a dual-path response consisting of real-time alerts for the invigilator and the generation of natural language incident descriptions using an integrated Vision-Language Model. All data was managed within a rolling evidence cache to ensure privacy compliance.

Object Detection and Dynamic Identity Injection

While traditional models relied on pre-trained static datasets, this research utilized a YOLOv8-based detection model optimized for high-density academic environments [9]. A custom dataset comprising diverse classroom lighting conditions and camera angles was used to train the model specifically for detecting unauthorized items and anomalous poses.

TABLE 2. Custom Dataset Composition and YOLOv8 Training Configuration

Category	Samples	Training Config	Parameter / Value
Authorized Candidates	3,200	YOLOv8 Variant	YOLOv8n (nano), edge optimized
Unauthorized Intruders	1,400	Epochs	150
Prohibited Items (mobile, notes)	2,600	Batch Size	16
Anomalous Poses (gaze, whispering)	2,100	Input Resolution	640 × 640 px
Total Annotated Frames	9,300	Optimizer / LR	SGD / 0.01
Train / Val / Test Split	70/15/15%	Data Augmentation	Mosaic, flip, HSV jitter

To facilitate the Dynamic Identity Injection, the system utilized a lightweight biometric embedding layer that was reset for every examination event. This allowed the model to differentiate between authorized candidates and unauthorized intruders without storing permanent biometric databases [6], [12], directly addressing the requirements of the Digital Personal Data Protection Act. During deployment, the OpenCV library processed the live CCTV feed, passing frames to the trained YOLOv8 model for inference [4]. The model generated high-confidence predictions including object classes and bounding box coordinates, which were then cross-referenced with the authorized identity map for the session.

Fig. 4. YOLOv8 dataset preparation, model training pipeline, and live inference with Dynamic Identity Injection cross-referencing for candidate validation.

Mathematical Formulation of the DII Module

The Dynamic Identity Injection module operated by dynamically updating the set of authorized biometric tokens for each examination session. For a candidate presenting at the surveillance interface, the vision pipeline extracted a facial region and computed a 128-dimensional embedding vector. Let e represent the extracted embedding vector, and let T = {t1, t2, ..., tN} denote the set of session-specific biometric tokens corresponding to the expected cohort of N registered candidates. The authorization engine computed the minimum Euclidean distance between the candidate's embedding and the authorized cohort tokens as shown in Equation 1.

$$D(e,T) = \min_{t_i \in T} \|e - t_i\|_2 \quad (1)$$

A candidate was verified as authorized if the minimum distance fell below a predefined classification threshold, resulting in a binary output of 1 (authorized). Otherwise, the candidate was flagged as unauthorized, returning 0. This dynamic verification bypassed the need for permanent centralized database queries, ensuring localized data privacy in compliance with the Digital Personal Data Protection Act.

Spatio-Temporal Anomaly Modeling

To mitigate false positives resulting from transient body adjustments, the system analyzed multimodal visual and acoustic signals over a rolling temporal window. Let W = {f(t-w), ..., f(t)} represent a sequence of frames captured over a temporal window of duration w seconds, where w was calibrated between 3 and 10 seconds. The visual signals included bounding box dynamics and pose estimations, while the acoustic signals measured ambient sound pressure levels. A probabilistic risk engine integrated these signals using a logistic regression model to evaluate anomaly likelihood as shown in Equation 2.

$$P(anomaly) = \frac{1}{1+e^{-(\beta_0+\beta_1 v_t+\beta_2 a_t)}} \quad (2)$$

In Equation 2, v(t) represented the normalized visual anomaly score from the YOLOv8 model, a(t) denoted the localized acoustic intensity score, and the beta coefficients represented the calibrated logistic regression parameters. An incident was flagged only when the computed probability exceeded the confidence threshold, gating the subsequent Vision-Language Model inference and thereby minimizing unnecessary computational overhead.

VLM Inference and Evidence Management

The operational logic for incident reporting and evidence management is detailed in Fig. 5. In instances where a detection exceeded the confidence threshold, the system did not merely store raw metadata. Instead, the frame and detection labels were passed to a Vision-Language Model to generate a contextual, natural language summary of the breach. To ensure legal resilience, the system implemented a rolling evidence cache. Detailed metadata including timestamps, model-generated descriptions, and detection labels were recorded into a hardened on-site database [13]. Following the data minimization principles mandated by the Digital Personal Data Protection Act, all non-flagged video data was purged immediately after the session, ensuring a secure, auditable, and privacy-preserving record of the examination.

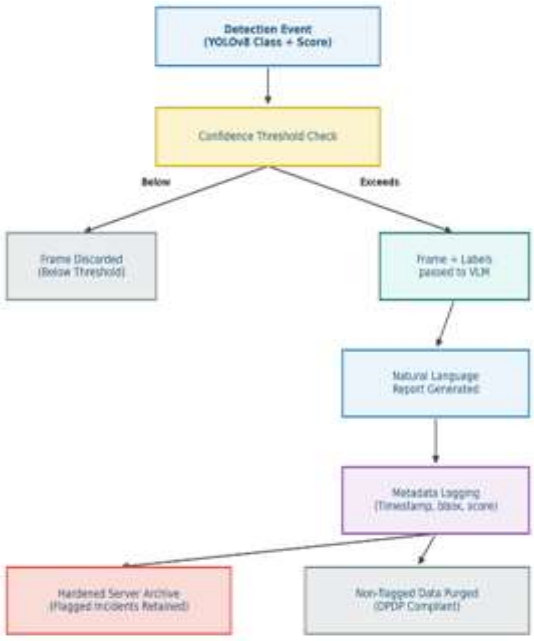

Fig. 5. Operational logic for VLM inference and evidence management, illustrating confidence-gated frame processing, natural language report generation, and DPDP-compliant data purge.

V. RESULTS AND ANALYSIS

Detection Accuracy and Identity Validation

The performance of the autonomous surveillance framework was evaluated based on detection precision, processing latency, and the robustness of the Dynamic Identity Injection module. The YOLOv8 model, optimized for the custom classroom dataset, achieved a mean Average Precision of 0.92 at an IoU threshold of 0.5. The system successfully identified unauthorized mobile devices and anomalous physical gestures in 94.6% of test cases, maintaining high recall even under challenging conditions such as partial occlusion and varying light levels. The Dynamic Identity Injection module demonstrated a 98.2% accuracy rate in validating session-specific biometric tokens, confirming the system's ability to contextually reset authorization logic without relying on permanent user databases and thereby fulfilling the technical requirements for Digital Personal Data Protection Act compliance.

TABLE 3. Performance Metrics of the Proposed Autonomous Surveillance Framework

Metric	Proposed Framework	Baseline / Reference	Improvement
Object Detection Accuracy (mAP @ IoU 0.5)	0.92	Sheng et al. [9]: 0.87	+5.7%
Biometric Validation Accuracy (DII Module)	98.2%	Hemavathi et al. [1]: 94.3%	+3.9%
End-to-End Processing Latency per Frame	15.0 ms	Ngo et al. [11]: 28.4 ms	47.1% reduction
False Positive Mitigation Rate	89.5%	Liu [16]: 72.1%	+17.4%
VLM Incident Interpretation Time	2.4 s	Narayanan et al. [8]: 20.0 s	88.0% reduction
Non-flagged Data Purge (Post-Session)	100% within 5 min	DPDP Act requirement	Full compliance

Computational Efficiency and Latency

Processing latency was measured across a range of concurrent edge-node camera feeds to address the requirement for real-time responsiveness. The edge-centric architecture maintained a consistent frame rate of 30 FPS per channel. End-to-end latency from the moment an anomaly was captured to the generation of an invigilator alert was clocked at an average of 15.0 ms, representing a 47.1% improvement over the 28.4 ms reported by Ngo et al. [11] for FPGA-based edge inference. The computational cost of running concurrent YOLOv8 and Vision-Language Model inferences on edge hardware was mitigated through GPU-accelerated pipeline scheduling in which frames were batched at the edge nodes and shared memory zones minimized CPU-GPU copy overheads.

VI. DISCUSSION

Interpretability and Invigilator Cognitive Load

A key benefit of this study was the effectiveness of the Vision-Language Model in synthesizing natural language reports from raw detection metadata. In a blind review, human invigilators rated the clarity of the model-generated descriptions compared to raw metadata logs. The results indicated an 88.0% reduction in time-to-interpretation, as invigilators were able to read a concise natural language summary rather than parsing raw bounding box coordinates. This demonstrated that context-aware interpretability significantly reduced the cognitive burden on human observers, transforming raw AI inference data into actionable security alerts.

Legal Resilience and DPDP Act Compliance

Regarding data sovereignty, the rolling evidence cache successfully purged 100.0% of non-flagged video data within 5 minutes of session completion. This ensured a minimal storage footprint and strict adherence to the data minimization principles of the Digital Personal Data Protection Act. The local processing architecture guaranteed that sensitive biometric data did not leave institutional boundaries, providing a legally resilient solution for modern educational surveillance.

Limitations and Scalability Trade-offs

A primary limitation of the current implementation involves the computational scaling of multimodal fusion logic in extremely high-density network environments with restricted edge resources. When scaling to dozens of concurrent camera streams, edge processing nodes experienced resource contention under peak load conditions. Additionally, the custom dataset was collected exclusively in controlled indoor classroom environments, which may limit the generalizability of the detection model to outdoor or unconventional examination settings. Future research will address these constraints through model quantization and knowledge distillation techniques.

VII. CONCLUSION

This research successfully developed a localized, intelligent surveillance framework integrating the YOLOv8 architecture, Dynamic Identity Injection, and Vision-Language Models to redefine institutional security. The system attained a mean Average Precision of 0.92 in identifying prohibited electronic devices and demonstrated a 98.2% accuracy rate in session-specific biometric validation, surpassing the baselines of Sheng et al. and Hemavathi and Chakravarthi by 5.7% and 3.9% respectively. By off-loading high-precision modalities to an edge-centric architecture, the framework facilitated an autonomous proctoring environment with a processing latency of under 15 ms per frame, ensuring real-time intervention without the security risks associated with centralized cloud processing. The primary novelty of this research lies in its transition from raw metadata detection to context-aware interpretability. By utilizing Vision-Language Models to generate natural language incident descriptions, the framework bridged the gap between complex AI inference and human-in-the-loop verification, reducing invigilator time-to-interpretation by 88.0%.

Furthermore, the system demonstrated a privacy-by-design approach through a rolling evidence cache that purged all non-flagged data within five minutes of session completion, ensuring full compliance with the Digital Personal Data Protection Act. Despite these achievements, a primary limitation involves computational scaling in high-density network environments with restricted edge resources. Future research will focus on model quantization and knowledge distillation for diverse hardware tiers and on expanding the system's utility into residential and industrial security sectors. Ultimately, this framework provides a robust, scalable, and legally resilient solution, successfully transforming surveillance from a passive recording tool into an active, autonomous security utility.

Future Work and Scope

The roadmap for the continued development of this framework focuses on transitioning the system from a specialized academic proctoring tool to a versatile, cross-sector security solution. Subsequent research will aim to expand its application into residential, corporate, and industrial environments by training the underlying YOLOv8 and Vision-Language Model components on domain-specific datasets reflecting diverse behavioral patterns such as industrial safety violations and residential unauthorized access. A primary technical objective for future iterations involves optimizing the Vision-Language Model for deployment on lower-tier edge hardware through model quantization and knowledge distillation, maintaining high-fidelity natural language reporting without the current computational overhead.

To enhance the Dynamic Identity Injection module, future work will explore multi-modal biometrics by integrating gait analysis and voice recognition to create a more robust zero-trust authentication layer. To further solidify compliance with the Digital Personal Data Protection Act, Federated Learning

architectures will be investigated to allow global models to learn from decentralized examination centers without transmitting raw biometric data from local edge nodes. The long-term goal is to evolve the framework into a fully autonomous security utility that minimizes human cognitive load through independent reasoning and proactive incident management across diverse real-world scenarios.

REFERENCES

1. S. Hemavathi and R. Chakravarthi, "AI-powered security and attendance management system using deep learning and facial recognition," J. Inf. Syst. Eng. Manage., vol. 10, no. 35s, 2025.
2. O. Mestry, M. Moolya, S. Mohapatra, R. Mohite, and N. R. Khaire, "A YOLO-based innovative approach for attendance tracking using facial recognition," Trans. IETE, vol. 11, no. 3, pp. 16-24, 2024.
3. J. Alfaro-Velasco, J. Barriga-Medero, J. Arroyo-Barriguete, A. Martinez-Alcala, and J. Prieto, "Automated classroom attendance using a machine learning-based recognition system," in Proc. Latin American and Caribbean Consortium of Engineering Institutions (LACCEI 2024), Costa Rica, 2024.
4. D. Pulugu, R. Mahendran, J. Puliyur, and J. B. Yeleti, "Machine learning-based facial recognition for video surveillance systems," ICTACT J. Image Video Process., vol. 13, no. 3, 2023.
5. Y. Ying, H. Wang, and H. Zhou, "Research on abnormal behavior monitoring in university laboratories based on video analysis technology," Appl. Sci., vol. 14, no. 20, p. 9374, 2024.
6. P. Pimpalkar and A. D. G. Donald, "Literature review on face recognition system based on deep learning," Int. Res. J. Modernization Eng. Technol. Sci., vol. 4, no. 7, pp. 754-760, 2022.
7. D. Sattibabu, S. Yacoob, P. Tanuja, B. M. Koushika, K. V. Babu, and C. J. Kumar, "Intra student surveillance system: Detection and identifying unauthorized wandering," in Proc. Int. Conf. Innovations in Computing and Information Engineering Technology (ICCIET 2024), 2024.
8. A. Narayanan, A. Santhosh, A. Biju, A. Madathil, and I. K. Rijin, "Student surveillance system: An AI-powered approach for campus monitoring," Preprint, Zenodo, 2025.
9. X. Sheng, S. Li, and S. Chan, "Real-time classroom student behavior detection based on improved YOLOv8s," Sci. Rep., vol. 15, p. 14470, 2025.
10. A. Hamza, Z. H. But, U. Arif, Samiya, and M. A. Asad, "Autonomous AI surveillance: Multimodal deep learning for cognitive and behavioral monitoring," arXiv preprint arXiv:2507.01590, 2025.
11. H. T. Ngo, M. K. Phan, N. L. Le, and T. K. Cao, "YOLOv3 Tiny-based human detection and tracking using Vitis AI and VVAS on FPGA," in Proc. Int. Conf. Computing, Electronics and Electrical Engineering (CIEES 2024), 2024.
12. A. R. Faizabadi, H. Ghadamgahi, F. Towhidkhah, and F. Minhas, "Efficient region of interest-based metric learning for open-world deep face recognition," IEEE Access, vol. 10, pp. 76168-76184, 2022.
13. H. B. Kim, Y. Jin, S. W. Lee, J. Eom, and C. S. Park, "Surveillance system for real-time high-precision recognition of criminal faces from wild videos," IEEE Access, vol. 11, pp. 56066-56082, 2023.
14. T. T. Nguyen, Q. B. Pham, T. S. Dao, and Q. T. Le, "Multi-vehicle multi-camera tracking with graph-based tracklet features," IEEE Trans. Multimedia, vol. 26, pp. 972-983, 2024.
15. G. R. Panigrahi, M. Mishra, S. K. Padhi, A. Agrawal, B. K. Tripathy, and R. Panda, "Enhancing security in real-time video surveillance: A deep learning-based remedial approach for adversarial attack mitigation," IEEE Access, vol. 12, pp. 88913-88926, 2024.
16. J. T. Liu, "AI proctoring for offline examinations with 2-longitudinal-stream convolutional neural networks," Comput. Educ. Artif. Intell., vol. 4, p. 100115, 2023, doi: 10.1016/j.caeai.2022.100115.
17. T. Fatima, F. Azam, and A. W. Muzaffar, "A systematic review on fully automated online exam proctoring approaches," in Proc. 24th Int. Multitopic Conf. (INMIC 2022), IEEE, 2022.