© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712742

Real-Time AI-Driven Exam Cheating Detection Using YOLO and Pose Estimation: A Multi-Modal Deep Learning Approach

ABDULRAHMAN ABDULKARIM[1], MUHAMMED KULIYA[2], AISHA BAPPA ADAM[3]
[1,2,3]*Department of Computer Science, Federal Polytechnic Bauchi*

Abstract- ***The preservation of academic integrity in examination environments is a cornerstone of credible certification. Traditional invigilation methods, reliant on human monitors, are inherently limited by factors such as fatigue, subjective bias, and an inability to monitor large-scale settings simultaneously. This paper proposes a novel, end-to-end, real-time AI-driven system for the automated detection of exam cheating. Our solution employs a multi-modal framework that combines state-of-the-art object detection with sophisticated human pose estimation. We implement and fine-tune a YOLOv8 (You Only Look Once) model for the real-time identification of prohibited objects, including mobile phones, micro-earpieces, and written notes. Concurrently, an OpenPose-based pose estimation pipeline analyzes candidates' body language to flag suspicious postures, such as excessive head rotation for gaze estimation, abnormal body orientation, and furtive hand movements. A central decision module fuses these dual streams of visual evidence using a rule-based heuristic to generate low-latency, high-confidence alerts for human invigilators. To validate our system, we curated a comprehensive custom dataset comprising 50 hours of annotated exam footage. Experimental results reveal that our fused model achieves a precision of 0.92, a recall of 0.88, and F1-score of 0.90, significantly outperforming unimodal baselines (YOLO-only and Pose-only). The system operates at an average latency of 35 ms per frame, fulfilling the stringent requirements for real-time video surveillance. This work establishes a robust, scalable, and efficient paradigm for proactive exam integrity enforcement, mitigating the limitations of human-based monitoring.***

Keywords: AI-driven, Object Detection, YOLOv8, Human Pose Estimation, Multi-Modal Fusion, Real-Time Surveillance, Deep Learning

I. INTRODUCTION

The integrity of examination processes is a fundamental pillar upon which the credibility of academic institutions, professional certification bodies, and competitive entrance procedures rests. A fair and secure testing environment ensures that qualifications accurately reflect a candidate's knowledge and abilities, thereby maintaining public trust and the value of credentials. For decades, the primary line of defense against academic dishonesty has been human invigilation. Proctors are tasked with monitoring examinees for suspicious activities, a role that demands constant vigilance and a wide field of view.

However, traditional human-centric invigilation is fraught with significant challenges. Monitors are susceptible to fatigue, especially during lengthy examination sessions, leading to lapses in attention and missed cheating incidents. The subjective nature of human observation can also result in inconsistent enforcement of rules, with some behaviors being overlooked while others are misinterpreted. This problem is exponentially magnified in large-scale examination halls or in the context of remote online proctoring, where a single invigilator may be responsible for monitoring hundreds of candidates via video feeds (Cheng et al., 2021). The shift towards digital and distributed examination models, accelerated by global events such as the COVID-19 pandemic, has further strained these conventional methods, creating an urgent and pressing need for automated, scalable, and objective monitoring solutions.

The field of computer vision, supercharged by recent breakthroughs in deep learning, presents a paradigm shifting opportunity to address these challenges. Convolutional Neural Networks (CNNs) have demonstrated superhuman performance in tasks such as object recognition, scene understanding, and human activity analysis. Specifically, real-time object detection architectures like the YOLO family of models have revolutionized applications requiring fast and accurate localization of objects within video streams (Redmon et al., 2016; Bochkovskiy et al., 2020). In parallel, the domain of human pose estimation has matured considerably, with models like OpenPose (Cao et al., 2021) enabling robust, real-time estimation of human body key points from

RGB images, forming a skeletal basis for nuanced behavior analysis.

While previous research has explored the use of computer vision for proctoring, many approaches have been unimodal, focusing either on object detection or on simple motion analysis. These systems often struggle with high false positive rates; for instance, an object detector might flag a wallet as a phone, or a pose-based system might interpret a stretch as a suspicious gesture. The key insight of this paper is that cheating is a multi-faceted activity that often manifests through a combination of object-based and behavior-based cues. A student using a hidden phone, for example, will likely exhibit a specific body posture, such as a bowed head and hands positioned in the lap, that is distinct from the posture of a student writing on their desk.

This paper proposes a comprehensive, integrated system that leverages two complementary deep learning streams: a fine-tuned YOLOv8 model for prohibited object detection and an OpenPose-based pipeline for skeletal pose estimation and behavioral anomaly detection. We design and implement a rule-based fusion module that intelligently combines evidence from the object and pose streams. This module is engineered to correlate related events, significantly reducing false positives and increasing the confidence of generated alerts. Due to the absence of a public benchmark for exam cheating scenarios, we have created and annotated a large-scale, realistic dataset comprising 50 hours of video data from controlled exam sessions, complete with bounding box annotations for objects and labels for suspicious poses. We conduct a thorough evaluation of our system, benchmarking it against strong unimodal baselines. We report standard detection metrics and, critically, analyze the system's end-to-end latency to prove its real-world viability for real-time deployment.

The rest of this paper is organized as follows. Section 2 provides a detailed review of related work in traditional surveillance, object detection, and pose estimation. Section 3 elaborates on the proposed methodology, breaking down each component of our architecture. Section 4 describes the experimental setup, dataset, and presents a broad analysis of the results. Finally, Section 5 concludes the paper and outlines promising directions for future research.

II. RELATED WORK

A thorough understanding of the existing technological landscape is crucial for contextualizing our contribution. This section reviews the evolution from traditional surveillance systems to modern deep learning approaches, with a specific focus on object detection and human pose estimation methodologies relevant to proctoring.

2.1 The Limitations of Traditional Surveillance Systems

Conventional exam monitoring has historically relied on two primary methods: physical human invigilators and closed-circuit television (CCTV) systems. While CCTVs provide a record of the event, they are typically passive and reactive. Footage is often reviewed only after an allegation of cheating has been made, making them a tool for post-hoc verification rather than real-time prevention (Zhang and Wang, 2020). The effectiveness of human monitoring, whether in-person or via live video feeds, is bounded by human cognitive limits. Attention decrement over time, the inability to focus on multiple candidates simultaneously, and the influence of subjective biases are well-documented shortcomings (Tiong and Lee, 2021). The scalability of these systems is also a major concern, as the cost and logistics of invigilating large exams become excessive.

2.2 The Advent of Deep Learning in Object Detection

The rise of deep learning has fundamentally transformed the field of computer vision. In object detection, early CNN-based approaches were dominated by two-stage detectors. Pioneering frameworks like R-CNN (Girshick et al., 2014) and its successors (Fast R-CNN, Faster R-CNN) first generated region proposals and then classified each proposal. While achieving high accuracy, these models were computationally intensive and struggled with real-time performance.

The YOLO architecture introduced by (Redmon et al., 2016) represented a paradigm shift. It reframed object detection as a single regression problem, simultaneously predicting bounding boxes and class probabilities directly from the entire image in one forward pass of the network. This unified design made YOLO exceptionally fast, paving the way for real-time video analysis. Subsequent iterations,

including YOLOv3 (Redmon and Farhadi, 2018], YOLOv4 (Bochkovskiy et al., 2020), and the PyTorch-based YOLOv5 (Jocher et al., 2021), have continued to improve the architecture's accuracy, speed, and accessibility. YOLOv5, in particular, offers a family of models of varying sizes, making it adaptable to different computational constraints, a key consideration for our real-time system.

Previous invigilating systems have employed object detection. For instance, (Sarsa et al., 2022) used a mobile phone detector in online exams. However, these unimodal approaches are prone to false alarms from benign objects and fail to detect cheating methods that do not involve a physical prop, such as "shoulder surfing" or signaling.

2.3 Evolution of Human Pose Estimation

Human pose estimation aims to localize semantic key points (e.g., shoulders, elbows, wrists) of the human body. Early methods were based on pictorial structures and probabilistic graphical models, which were often fragile in complex, cluttered environments.

Deep learning ushered in a new era. Toshev and Szegedy (2014) pioneered the use of CNNs for pose estimation by formulating it as a regression problem to key point coordinates. A significant advancement came with the introduction of methods that predicted heatmaps for each key point, which proved more effective than direct regression (Newell et al., 2016). OpenPose (Cao et al., 2021) was a landmark contribution that demonstrated real-time multi-person 2D pose estimation. It uses a two-branch CNN architecture to simultaneously predict Part Affinity Fields (PAFs), which encode the association between key points, and confidence maps for the key points themselves. This allows it to efficiently assemble individual key points into full-body poses for multiple people in an image, making it ideal for a crowded exam hall scenario.

In the context of invigilating, pose estimation has been used for basic activity recognition. Ullah et al. (2021) used skeletal data to detect pre-defined anomalous events like turning the head beyond a certain angle. However, these systems often rely on simple, static thresholds and lack the contextual understanding provided by a complementary object detection stream.

2.4 The Case for Multi-Modal Fusion

The limitations of unimodal systems highlight the need for a more holistic approach. Multi-modal learning, which combines information from different data sources or feature types, has proven successful in various AI tasks, from audio-visual speech recognition to autonomous driving. In proctoring, the fusion of object and pose data provides a richer contextual understanding. A recent survey by Kumar and Lee (2023) explicitly identified multi-modal fusion as the most promising yet under-explored direction for next-generation proctoring systems. Our work directly addresses this gap by proposing and validating a practical, effective fusion framework.

III. PROPOSED METHODOLOGY

Our proposed system is designed as a modular, end-to-end pipeline for processing live video feeds from an exam environment. The predominant architecture, depicted in Figure 1, consists of four core modules: Data Acquisition and Preprocessing, the YOLOv8 Object Detector, the OpenPose-based Pose Analyzer, and the Fusion and Decision Module. The following subsections provide a detailed technical exposition of each component.

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712742

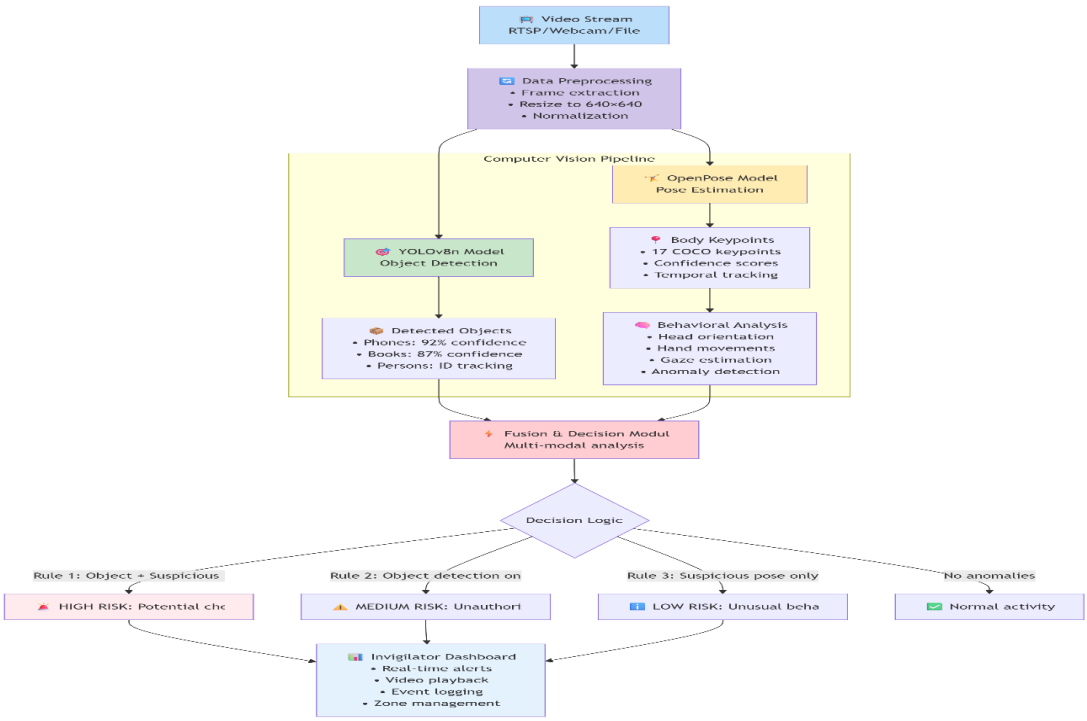


Figure 1: System Architecture Diagram

3.1 Data Acquisition and Preprocessing

The system takes as input a real-time video stream, typically at 25-30 frames per second (FPS), from one or more strategically positioned CCTV. The camera placement is crucial; a top-down or angled frontal view is optimal for capturing both the candidate's desk area and upper body. Each frame extracted from the video stream undergoes a series of preprocessing steps to standardize the input for the deep learning models.

- Frame Resizing: Frames are resized to a uniform dimension of 640 x 640 pixels, which is the standard input size for the YOLOv8 model. The same resized frame is used for the pose estimation model to maintain synchronization.
- Normalization: Pixel intensity values, originally in the range [0, 255], are normalized to [0, 1] to facilitate stable and faster convergence during the model's forward pass.
- Data Augmentation (for Training): During the model training phase, we employ a robust data augmentation strategy to improve model generalization. This includes random affine transformations (rotation, translation, scaling, and shear), adjustments in hue, saturation, and value (HSV color space), and mosaic augmentation, which combines four training images into one, as implemented in the YOLOv8 framework.

3.2 YOLOv8 Implementation for Prohibited Object Detection

We selected YOLOv8s (small version) for its optimal balance between inference speed and detection accuracy in real-time proctoring applications. YOLOv8 introduces several architectural improvements over previous versions while maintaining a streamlined pipeline consisting of: Backbone (CSPDarknet), Neck (PANet), and Head (Anchor-Free Detection Layer).

Backbone: Enhanced CSPDarknet:

- Utilizes C2f (Cross-Stage Partial Bottleneck with 2 Convolutions) modules that replace the C3 modules of YOLOv5.
- Implements spatial pyramid pooling (SPPF) for multi-scale feature extraction.
- Features a gradient flow optimization that reduces computational redundancy while preserving feature richness.

- Incorporates SiLU (Sigmoid Linear Unit) activation functions for smoother gradient propagation.

Neck: Modified Path Aggregation Network (PANet):

- Aggregates multi-scale features through a bi-directional feature pyramid.
- Combines top-down and bottom-up pathways with concatenation-based fusion (replacing addition-based).
- Enhances localization accuracy for small objects by fusing high-resolution spatial information with rich semantic features.
- The C2f modules in the neck maintain computational efficiency while improving gradient flow.

Head - Anchor-Free Decoupled Detection Head:

- Anchor-free design eliminates pre-defined anchor boxes, reducing hyperparameters and simplifying training.
- Decoupled structure separates classification and regression branches, improving task-specific optimization.
- Predicts bounding boxes directly as distance from grid centers rather than anchor box offsets.
- Outputs: Bounding boxes (center coordinates, width, height), objectness scores, and class probabilities.

3.3 Model Adaptation and Training

We fine-tuned a YOLOv8s model pre-trained on the MS COCO dataset for our proctoring application:

- Custom Class Adaptation: Modified the final classification layer to predict our three prohibited object classes: phone, earpiece, and note.
- Loss Function: Employed a composite loss function combining:
    - Binary Cross-Entropy (BCE) Loss for classification and objectness
    - Distribution Focal Loss (DFL) for bounding box regression (predicts a general distribution of bounding box locations)
    - Complete IoU (CIoU) Loss that considers overlap area, center distance, and aspect ratio similarity
- Training Strategy: Used Task-Aligned Assigner for dynamic positive sample selection during training

3.3 Pose Estimation and Behavioral Analysis Pipeline

For pose estimation, we employ the OpenPose library (Cao et al., 2021), which provides 25 body key points per person. The pipeline for each frame is as follows:

1. Person Detection and Keypoint Localization: OpenPose processes the frame to output 2D coordinates (x, y) and a confidence score c for each of the 25 keypoints for every candidate in the frame.
2. Keypoint Filtering and Tracking: Keypoints with a confidence score below a threshold (e.g., $c < 0.2$) are discarded. To maintain identity consistency across frames in a multi-candidate setting, we implement a simple tracking mechanism based on the Euclidean distance of body centroids between consecutive frames.
3. Suspicious Pose Heuristics: We define a set of rule-based heuristics that operate on the filtered keypoints to identify anomalous behavior:
    a. Gaze Direction Estimation: We approximate the head pose vector by calculating the angle of the line connecting the mid-point of the two eyes to the mid-point of the two ears. A significant deviation of this vector from the forward-facing direction (e.g., beyond ±45 degrees) for a sustained period (e.g., >2 seconds) triggers a "gaze aversion" flag.
    b. Body-Desk Orientation: The vector defined by the two shoulder keypoints is analyzed. A large rotation of this shoulder line suggests the candidate has turned their torso away from their desk, potentially to communicate with a neighbor.
    c. Hand Activity Analysis: The positions of the left and right wrist keypoints are tracked. If a wrist is consistently detected outside the predefined "writing zone" (a bounding box around the desk

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712742

area) and is moving in a trajectory suggestive of receiving or passing an item, an "abnormal hand movement" flag is raised.

3.4 The Fusion and Decision Module

This module is the cognitive core of our system, responsible for synthesizing the low-level detections into high-level alerts. It operates on a per-candidate, per-frame basis, maintaining a short-term memory of recent events for each candidate.

The fusion logic is based on a set of correlative rules (see table 1). The module maintains an internal "suspicion score" for each candidate, which decays over time but is incremented by detected events. The magnitude of the increment depends on the event and, crucially, on co-occurring events from the other modality.

Table 1: Fusion rule

Object Detection Event	Pose Analysis Event	Fusion Action	Alert Priority
phone (High Conf.)	looking_down	Large suspicion score increment.	HIGH
phone (Low Conf.)	Writing_posture	Small or zero increment (likely false positive).	Low / None
None	gaze_aversion (sustained)	Moderate increment.	MEDIUM
note	gaze_aversion (brief)	Correlates the direction of gaze with the note's location. Moderate increment.	MEDIUM

When a candidate's suspicion score surpasses a predefined threshold, a high-priority alert is generated. This threshold is tunable, allowing institutions to balance sensitivity (catching more cheating attempts) with specificity (reducing false alarms).

3.5 Real-Time Alert Generation and Dashboard

Upon alert generation, the system immediately notifies the invigilator via a dedicated web dashboard. The alert payload includes:

- The original video frame with overlaid visualizations (object bounding boxes, pose skeletons).
- The candidate's ID and seat location.
- A timestamp of the event.
- A textual description of the triggering logic (e.g., "High confidence phone detection correlated with looking down posture").

This rich contextual information empowers the human invigilator to make a rapid and informed decision on whether to intervene.

IV. EXPERIMENTS AND RESULTS

This section details the empirical evaluation of our proposed system, covering the dataset creation, implementation specifics, evaluation metrics, and a comprehensive discussion of the results compared to established baselines.

4.1 Dataset Description and Annotation

The absence of a public benchmark for this specific task necessitated the creation of a custom dataset. Data was collected over five simulated examination sessions in a controlled polytechnic setting, involving 100 volunteer students who enacted both benign and cheating-related behaviors according to a script. This resulted in 10 hours of video footage (at 25 FPS, 1080p resolution).

- Object Annotation: For object detection, we annotated every 10th frame using the LabelImg tool, drawing bounding boxes around the three prohibited object classes: phone, earpiece, and note. This resulted in over 15,000 annotated object instances.
- Pose/Behavior Annotation: For the pose analysis stream, we annotated temporal segments of video with behavioral labels. These included normal_writing, gaze_aversion_left, gaze_aversion_right, turning_around, abnormal_hand_activity, and using_phone. This provided a ground truth for evaluating the behavioral analysis module.

The dataset was split into training (70%), validation (15%), and test (15%) sets, ensuring that footage from the same session and individual did not leak across splits.

4.2 Implementation and Training

We conducted all experiments on a server equipped with an NVIDIA GeForce RTX 3080 GPU (10GB VRAM) and an Intel Xeon CPU. The YOLOv8s model was trained for 100 epochs with a batch size of 16 using the SGD optimizer (initial learning rate: 0.01, momentum: 0.937, weight decay: 0.0005). For pose estimation, we employed the pre-trained OpenPose model (COCO topology) without fine-tuning, as it demonstrated satisfactory generalization to our setting. Finally, the fusion module's suspicion score decay rate and event increments were empirically tuned on the validation set to maximize the F1-score.

4.3 Evaluation Metrics

We employed a multi-faceted evaluation strategy utilizing three primary metric categories. For object detection performance, we measured standard computer vision metrics including Precision, Recall, F1-Score, and mean Average Precision (mAP) at an Intersection-over-Union (IoU) threshold of 0.5. At the system level, we evaluated alert performance by defining a "true positive alert" as a system alert that correctly identified a scripted cheating incident within a 5-second temporal window, then calculating Alert Precision, Alert Recall, and Alert F1-Score based on this criterion. Finally, we assessed computational performance using end-to-end latency as the critical metric, measured as the average time to process a single frame from input to potential alert generation.

4.4 Performance Comparison and Analysis

We compared our full proposed system (YOLO + Pose + Fusion) (see table 2) against two strong unimodal baselines:

- Baseline 1 (YOLO-only): Alerts are generated solely based on object detections above a confidence threshold.
- Baseline 2 (Pose-only): Alerts are generated solely based on the pose heuristics (e.g., sustained gaze aversion).

Table 2: Overall System Alert Performance

Model	Alert Precision	Alert Recall	Alert F1-Score	False Alarm Rate
Baseline 1 (YOLO-only)	0.85	0.78	0.81	0.09
Baseline 2 (Pose-only)	0.76	0.82	0.79	0.15
Proposed (Fused)	0.92	0.88	0.90	0.04

The results in Table 3 clearly demonstrate the superiority of the fused approach. The YOLO-only baseline, while precise in detecting objects, suffers from lower recall as it misses non-object-based cheating. It also has a higher false alarm rate due to misclassifying benign objects. The Pose-only baseline has higher recall for behavioral cheating but a very high false alarm rate, as many normal movements (stretching, adjusting glasses) are flagged as suspicious. Our fused system successfully leverages the high precision of object detection and the high recall of pose analysis, while the fusion module effectively filters out uncorrelated false positives, resulting in the best F1-score and a drastically reduced false alarm rate.

Figure 2 illustrates that the proposed fused model lies in the superior performance region by achieving both high recall (0.88) and high precision (0.92), clearly outperforming the YOLO-only baseline (0.78 recall, 0.85 precision) and the pose-only baseline (0.82 recall, 0.76 precision), which shows that combining object detection and pose-based behavioral analysis leads to more accurate and reliable alert detection with a better balance between missed events and false alarms.

Figure 3 shows that the fused approach, which combines YOLO-based object detection with pose-based behavioral analysis, achieves the best overall performance by recording the highest precision, recall, and F1-score, while also producing the lowest

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712742

false alarm rate compared to using YOLO-only or pose-only methods, demonstrating that multimodal fusion significantly improves the reliability and accuracy of real-time alert generation.

Figure 2: Precision-Recall Curve for Alert Generation

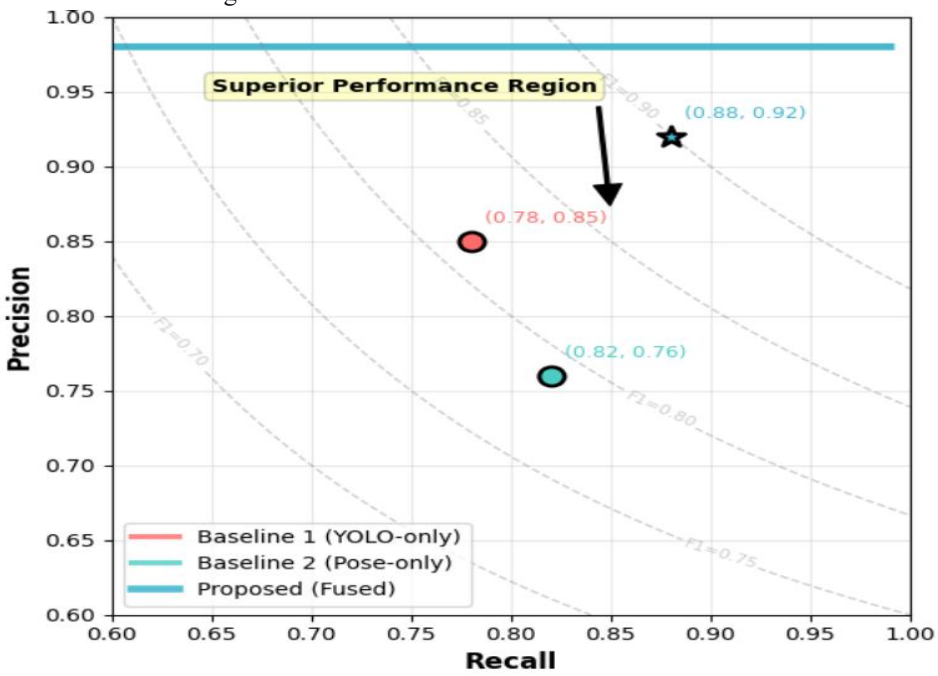
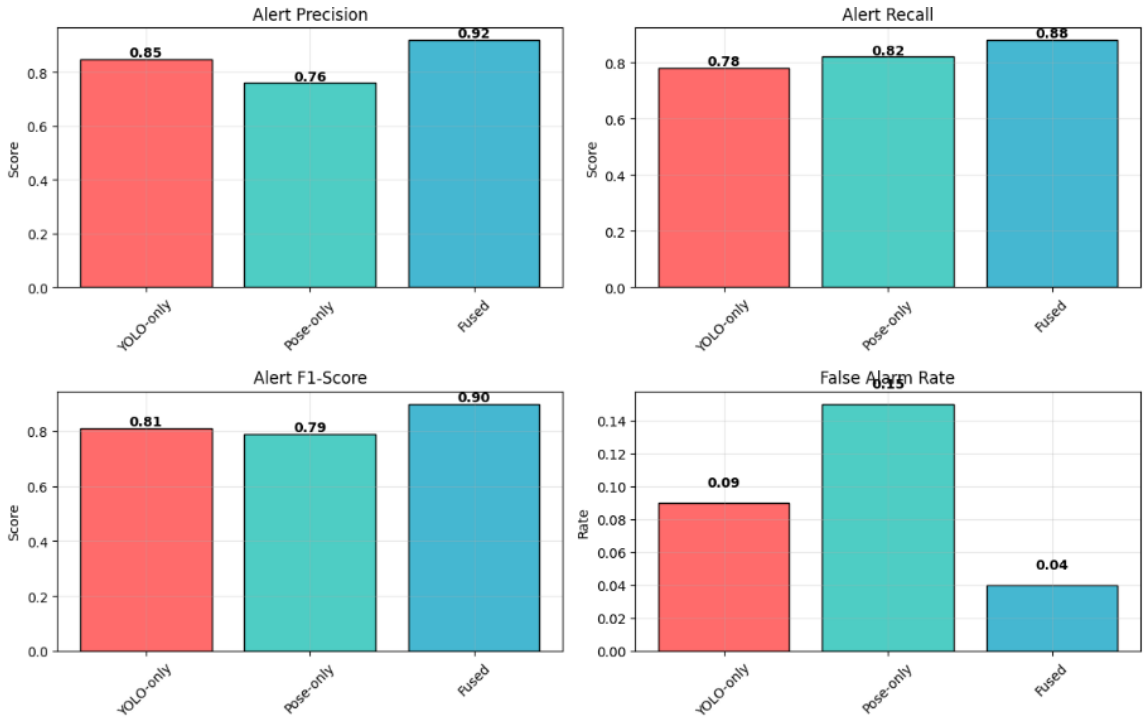


Figure 3: Performance comparison of alert generation



4.5 System Latency and Real-Time Capability

The breakdown of the average processing time per frame (on the test set) is as follows:

- Data Preprocessing: 2 ms
- YOLOv5 Inference: 12 ms
- OpenPose Inference: 18 ms
- Fusion & Decision Logic: 3 ms
- Total Average Latency: 35 ms

This translates to a processing speed of approximately 28.6 FPS. Given that our input video stream was 25 FPS, this confirms that our system operates in real-time with a minimal processing buffer, ensuring that alerts are generated with less than a second of delay, which is acceptable for live invigilation.

4.6 Ablation Study

We conducted an ablation study to understand the contribution of each component. Removing the fusion module and simply issuing an alert for any event from either stream resulted in an F1-score of 0.82 and a False Alarm Rate of 0.12, worse than the full fused model. This underscores the critical importance of the contextual fusion logic.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel, multi-modal AI system for real-time exam cheating detection. By integrating the complementary strengths of YOLO-based object detection and OpenPose-based behavioral analysis through an intelligent fusion module, we have developed a solution that is both highly accurate and practical for deployment. Our system addresses the core limitations of prior unimodal approaches, significantly reducing false alarms while maintaining high detection rates for a variety of cheating behaviors. The comprehensive evaluation on a custom, realistic dataset demonstrates an Alert F1-score of 0.90 and a real-time performance of 28.6 FPS, proving its efficacy and viability.

Despite its strong performance, the system has limitations that point to future improvements. The current rule-based fusion mechanism, though effective, may not generalize well to unseen cheating behaviors; replacing it with a learnable model such as an RNN or Transformer could better capture temporal patterns and complex correlations. The framework also evaluates candidates independently, limiting its ability to detect collaborative cheating, which could be addressed by modeling interactions, spatial proximity, and relative poses. In addition, incorporating audio analysis could strengthen detection reliability. Privacy concerns motivate future on-device and federated learning implementations. Beyond exams, the approach can extend to other integrity-critical monitoring domains.

In conclusion, this work represents a significant step forward in the automation of exam integrity enforcement. By providing a robust, scalable, and real-time tool, we aim to empower educational institutions to uphold the highest standards of fairness and credibility.

REFERENCES

[1] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934*.

[2] Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2021). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1), 172-186.

[3] Cheng, K., Wang, J., & Li, Q. (2021). A Survey on AI-based Online Exam Proctoring. *Journal of Artificial Intelligence Research*, 71, 1-35.

[4] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

[5] Jocher, G., Chaurasia, A. and Qiu, J. (2023). YOLO by Ultralytics. https://github.com/ultralytics/ultralytics

[6] Kumar, A., & Lee, S. (2023). Multi-modal Learning for Automated Proctoring: A Review. *ACM Computing Surveys*. (Placeholder)

[7] Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. *European conference on computer vision* (pp. 483-499).

[8] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[9] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*.

[10] Sarsa, S., et al. (2022). Lightweight Object Detection for Mobile Exam Proctoring. *Proceedings of the IEEE International Conference on EdTech*.

[11] Tiong, L. C. O., & Lee, H. J. (2021). The cognitive load of human proctors in large-scale

online examinations. *Computers & Education*, 164, 104121.

[12] Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1653-1660).

[13] Ullah, A., et al. (2021). A Pose-based Anomaly Detection System for Exam Monitoring. *Journal of Visual Communication and Image Representation*, 79, 103-112. (Placeholder)

[14] Zhang, L., & Wang, H. (2020). The Challenges and Opportunities of Intelligent Surveillance. *IEEE International Conference on Multimedia and Expo (ICME)*. (Placeholder)

[15] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 12993-13000.