Integrated 3-Layer Online Test Cheating Detection System Using YOLOv8, InsightFace, and GazeTracking Modules

Farrel Laogi Murjitama
Department of Informatics Engineering, Faculty of Telematics Energy, Institut Teknologi PLN, Jakarta

Yudhy S. Purwanto
Department of Informatics Engineering, Faculty of Telematics Energy, Institut Teknologi PLN, Jakarta

Abstract: The adoption of online tests has introduced significant challenges in maintaining academic integrity, particularly in real-time detection of cheating behaviors. This study proposes an intelligent proctoring system that automatically detects suspicious participant behavior during an online test by integrating image processing and computer vision techniques. The system integrates a YOLOv8s model based on the YOLO neural network algorithm to localize and classify facial states and suspicious objects in each video frame. This detection layer is complemented by an InsightFace face recognition module, which extracts deep facial embedding features and performs similarity matching against a registered reference image to continuously verify the identity of the participant and detect attempts at impersonation. In parallel, the GazeTracking module analyzes eye landmarks and pupil dynamics to monitor eye behavior, including blinking and significant gaze deviation, providing additional behavioral cues related to attention and potential cheating. The system consists of three detection layers: (1) YOLOv8s for object and behavior detection, (2) InsightFace for identity verification, and (3) GazeTracking for eye behavior analysis. Together, these components form a synchronized computer vision module that performs real-time analysis from live video streams, allowing the system to classify behavioral states such as abnormal head orientation, multiple faces, foreign objects, no face detected, identity mismatch, and eye closure. The experimental results show that the YOLOv8s model achieves an mAP@50 of 0.9918, a precision of 0.9856, and a recall of 0.9903 on the validation set while maintaining real-time performance at an average of 10 frames per second. The findings demonstrate that deep learning-based visual monitoring can effectively support automated online exam supervision, offering a viable computer vision-based proctoring approach.

Keywords: Online test, Cheating detection, YOLO, InsightFace, GazeTracking.

Correspondents Author:
Farrel Laogi Murjitama, Department of Informatics Engineering, Faculty of Telematics Energy, Institut Teknologi PLN, Jakarta
Email: farrel2231096@itpln.ac.id

Received February 15 2026; Revised March 17, 2026; Accepted March 19, 2026; Published March 24, 2026.

Introduction

The development of digital technology has brought about major changes in various areas of life, including the education sector (Gusniwati & Rahmawati, 2024). The face-to-face learning process is now increasingly shifting to an online learning model. Online examinations, in particular, have become increasingly prevalent, allowing students to complete academic assessments remotely (Gusai et al., 2023). However, new challenges arise in terms of academic integrity and honesty. Online examination systems that are not directly supervised could open great opportunities for cheating. Examinees may engage in actions such as collaborating with others (collusion), using additional devices (mobile phones, dual screens, or communication applications), or manipulating their identity using a proxy or facial manipulation technology such as deepfake (Hasibuan & Hendrik, 2025).

In online examinations, participants are generally supervised in simple ways, such as by using a webcam or screen recording (S. Purwanto et al. 2022). However, this system still has fundamental weaknesses , such as manual monitoring by supervisors and limited human ability to supervise many participants simultaneously.As a result, manual supervision is no longer effective when the number of participants increases or when exam participants have access to more advanced technology (S. Purwanto et al.2022). Therefore, an AI-based automated proctoring system capable of real-time supervision and detecting indications of cheating without direct intervention is needed.

Advances in computer vision and deep learning technology have opened enormous opportunities to address this issue. One rapidly developing method is face and object detection, which is the ability of a system to automatically recognize and track specific objects from images or videos (Abdurrasyid et al., 2022). The latest models , such as YOLO version 8s, are known for their lightweight and efficient architecture and fast and accurate detection capabilities. This model can identify various objects within a single image frame, including faces, hands, and mobile phones.The small ('s') variant was selected to balance detection accuracy and inference speed, making it suitable for real-time CPU-constrained proctoring environments. YOLOv8s has the potential to become the primary model in visual-based fraud detection systems (Putu Ary Sri Tjahyanti et al., 2024).

Several previous studies have used earlier versions of the YOLO algorithm (YOLOv4, YOLOv5, YOLOv4, YOLOv5). YOLOv5) to detect suspicious activities such as mobile phone use or repetitive head movements during online examinations (Essahraui et al., 2025). However, these studies are generally limited to only one aspect of behavior, such as object detection without facial expression analysis or participant identity recognition. Face recognition and motion analysis technologies, such as FaceNet, ArcFace, and MediaPipe, are also developing

rapidly, but they have not yet been fully integrated into the context of online exam monitoring systems (Korah & Varghese, 2024).

Based on these conditions, a significant research gap. Several integrated systems combine face and object detection, identity face recognition, and eye movement analysis (head, object, face, and eye analysis) to detect various forms of cheating during online examinations (Jatnika et al., 2023). Furthermore, the application of the YOLOv8s model in the context of online examinations remains limited, despite its significant improvements over predecessor models in terms of detection accuracy and inference speed (Nur Aziz Thohari et al., 2025). Another important challenge is how the system can maintain privacy and ethical use of facial data so that the technology developed does not violate the rights of exam participants (S. Purwanto et al. 2022).

Therefore, this research focuses on developing a facial analysis-based online exam cheating detection system using the YOLOv8s model, which is capable of analyzing the behavior of participants in real time through facial video recording (Putra & Mulyana, 2024). This system is designed to detect several forms of cheating behavior, such as looking away for too long, using additional devices such as mobile phones, and potential cheating based on facial identity mismatches (Purwanto & Putra, 2025). The detection process will be carried out through a combination of identity verification (face recognition) and analysis of gaze direction and facial movements (motion detection) (Potluri et al., 2023).

This study is expected to produce a prototype of an intelligent proctoring system (S. Purwanto et al.2022) that is not only capable of accurately and efficiently detecting cheating behavior but also makes a real contribution to improving academic integrity in the digital age. In addition, the results of this research are expected to serve as an initial reference in the development of a deep learning based monitoring system that is ethical, transparent, and adaptive, in accordance with the principles of academic honesty and personal data protection (Jatnika et al., 2025).

Research Method

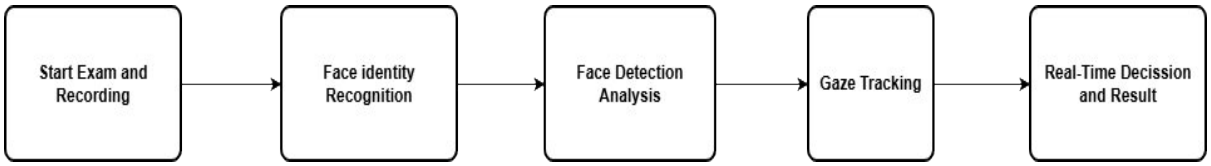

Figure 1 Integrated pipeline system

This study proposes a computer vision based proctoring framework that detects cheating behavior during online examinations by combining deep learning-based object detection with facial verification and gaze analysis. The research methodology comprises five main stages:

dataset preparation and annotation, automatic dataset labeling, YOLOv8s model training, model evaluation, and real-time system integration.

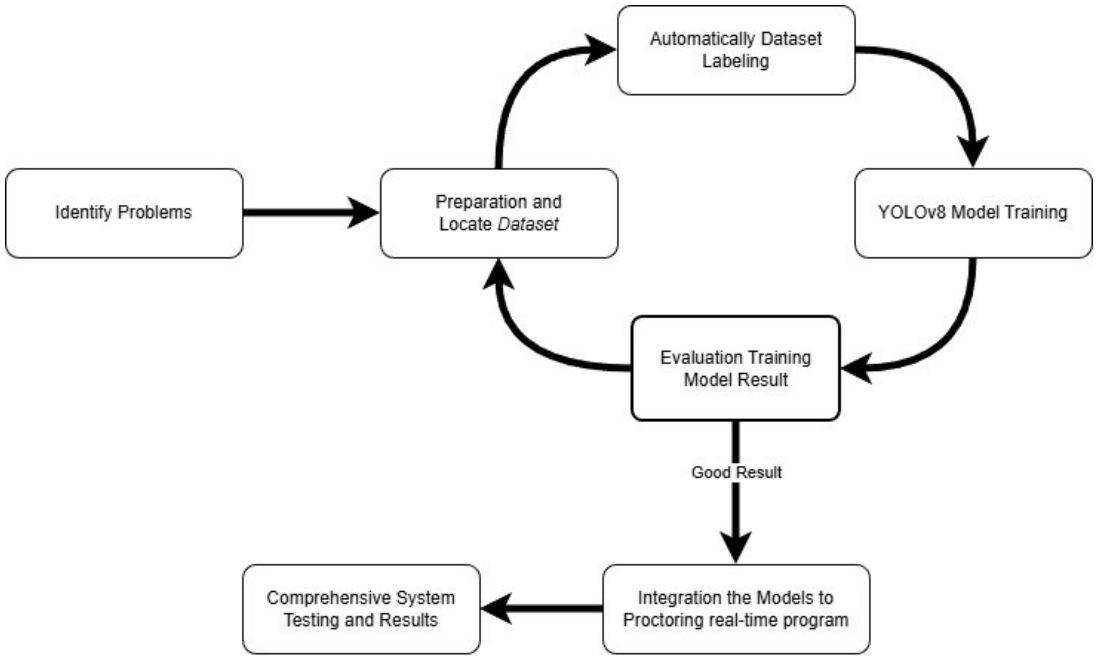

Figure 2 Overall research flow

Preparation and labeling of the dataset

Dataset source

The dataset was self-collected through simulated online examination sessions representing predefined normal and cheating behavioral scenarios. All images were obtained from simulated online examination sessions , representing predefined normal and cheating behaviors.

To improve robustness, each class was recorded under the following four lighting categories as following; light_A (bright), B (normal), C (dim), and D (dark). These categories ensure brightness and contrast diversity (Juliandy et al., 2024). The dataset represents variations in participant behavior, p. 3). lighting conditions, and environmental disturbances that may occur during online examinations (Essahraui et al., 2025). Each image was stored in PNG format and manually organized according to the behavioral categories relevant to cheating detection. Before the automatic annotation, the raw captured images were grouped according to the behavioral class and lighting conditions. Table 1 summarizes the distribution of the initial images.

Table 1 Distribution of datasets by class and lighting condition

Class	light_A	light_B	light_C	light_D	Total Images
face_normal	71	58	146	139	414
face_not_forward	89	106	41	165	401
foreign_object	255	153	302	186	896
multi_face	42	42	71	85	240
Total	457	359	560	575	1951

Therefore, the dataset contained 1,951 images in total. The lighting categories do not represent separate datasets but rather environmental variations within each class (Islam, 2023) to ensure that the model is exposed to heterogeneous illumination during training. The dataset size is relatively limited for multi-class detection training. This constraint is addressed through data augmentation (mosaic, horizontal flip) and transfer learning from a pretrained YOLOv8s base model. The study did not include eye movement and face identity verification as dataset classes of the YOLOv8s model because object detection models do not handle such tasks optimally (Gündüz & Işık, 2023). Instead, these features were separately implemented using specialized libraries with InsightFace for face recognition and GazeTracking for eye behavior analysis.

Automatic labeling

The manual labeling of large amont image datasets is time-consuming and prone to inconsistency. Therefore, we implemented an automated labeling method that combines face and object detection models using a hybrid approach.The pipeline uses the InsightFace framework for facial detection and landmark-based pose estimation (Potluri et al., 2023). The Buffalo_L model within the InsightFace framework was used to provide robust facial detection and 5-point landmark estimation. Only images that satisfied strict filtering criteria, including minimum face size, frontal pose constraints, and lighting-adjusted confidence thresholds, were retained for labeling. CLAHE was applied as a preprocessing step to normalize illumination for images affected by brightness and contrasts (Okyere-Gyamfi et al., 2025).

$$I' = \text{CLAHE}(I) \tag{1}$$

Variable description: I: original input image (raw image) from the webcam, I': Image of contrast enhancement after CLAHE, CLAHE (·): function of Contrast Limited Adaptive Histogram Equalization. CLAHE was selected over global histogram equalization because it adaptively enhances local contrast in non-uniform and low-light regions without overamplifying noise, which is critical for reliable face detection under dim and dark lighting conditions (light_C, light_D). This is important to ensure that the face detection model works stably under different lighting conditions. Then, the valid samples were automatically

converted into a bounding box format and stored as paired image label files (Potluri et al., 2023). This method transforms image captures into a structured training dataset with minimal human intervention:

$$x_c = \frac{x_1 + x_2}{2W} \quad (2)$$

$$y_c = \frac{y_1 + y_2}{2H} \quad (3)$$

$$w = \frac{x_2 - x_1}{W} \quad (4)$$

$$h = \frac{y_2 - y_1}{H} \quad (5)$$

Variable description: x_1, y_1 : coordinates of the top left corner of the bounding box (pixel), x_2, y_2 : bottom right corner coordinates of the bounding box (pixels), W: image width (pixels), H : image height (pixels), x_c:center coordinates of the bounding box on the x-axis (normalized from 0 to 1), y_c : center coordinates of the bounding box on the y-axis (normalized 0–1), w : relative bounding box width relative to the image, and h: bounding box height relative to the image. The generated labels are automatically stored in the dataset directory structure. This automated process ensures labeling quality while reducing human mistake bias.

Dataset Splitting

After labeling, all images were consolidated into a unified dataset and automatically split into training and validation datasets specifically optimized for supervised training. The dataset was divided with a ratio of 80% for training and 20% for validation with fixed-seed randomized shuffling to ensure reproducibility (Septiandi et al., 2021).

Table 2 Class distribution of datasets

Class	Description	Total Images	Training	Validation
face_normal	The participant is facing forward in a normal posture.	410	328	82
face_not_forward	Participant looking away or turning sideways.	385	311	74
multi_face	More than one face is visible in the frame.	202	154	48
foreign_object	The presence of unauthorized objects (e.g., smartphone).	323	263	60
Total		1320	1056	264

The curated dataset distribution is summarized in above. No independent test set was held out section. Therefore, the reported performance metrics reflect validation set performance, and formal generalization to fully unseen data cannot be established from this study alone, a limitation addressed in the Discussion. The filtering process improves dataset consistency by ensuring that each retained image clearly represents its intended behavioral class (Khoiriyah et al., 2025). Images containing occlusions, ambiguous poses, insufficient face visibility, or detection failures were excluded during filtering. The curated set of 1,320 images represents 67.7% of the 1,951 raw images, reflecting the strictness of the filtering criteria.

YOLOv8s Model Training

The detection module is based on the YOLOv8s model, which is fine-tuned on the curated dataset through transfer learning (TL) from pretrained weights (Ajayi et al., 2024). TL was applied to accelerate convergence and leverage visual features already encoded in the pretrained backbone. The following is the composite training loss:

$$L = L_{box} + L_{obj} + L_{cls} \quad (6)$$

Variable description: L: Minimized total loss during training, *L*: bounding box regression loss (object position error), L_{obj}: loss of objecthood (belief in the existence of objects), *L*: Classification of loss of object classes. TL was performed using GPU acceleration with mixed precision inference to accelerate convergence and leverage pretrained feature representations to maximize hardware utilization (Kljucaric & George, 2023). Due to hardware memory constraints, a batch size of 4 was used with the AdamW optimizer to regularize the optimization process with cosine decay learning rate scheduling and early stopping (patience = 50) were applied to compensate for the small batch effect and ensure small-batch training stability. For the YOLOv8s model, the training configuration must be adjusted for the maximum result as follows:

Table 3. YOLOv8 training configuration

Parameter	Value
Input (dataset) resolution	640 × 640
Epochs	200
Data caching	Disk
Model	yolov8s.pt
Optimizer	AdamW
Early stopping patience	50

Model Evaluation

Model performance was evaluated on the validation dataset using standard object detection metrics, including precision, recall, F1 score, and mean Average Precision (mAP). Confusion matrices were generated to analyze the class-wise prediction behavior (Drantantiyas et al., 2023). Model performance is evaluated using precision, recall, and F1 score (Putra & Mulyana, 2024)as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{8}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \tag{9}$$

Variable description: True Positive (TP): correct detection of existing objects, False positive (FP): incorrect detection of objects that do not exist, FN (False Negative): objects that fail to be detected, Precision: accuracy of detection, Recall: detection completeness, F1: average harmonic precision and recall. The evaluation is solely conducted on the validation split. The limitations of this evaluation protocol include: absence of statistical confidence intervals, lack of robustness testing under adversarial or extreme lighting conditions, and no evaluation of data from external sources or independent participants. These results provide clear directions for future validation.

Facial Recognition Module (FRM)

A face recognition module verifies participant identity throughout the examination session using deep facial embeddings from InsightFace Buffalo_L (S. Purwanto et al. 2022). The module uses deep facial embeddings in a one-vs-reference verification mode: a single reference embedding is registered at the start of the session, and each detected face during the session is compared against this reference. This component was evaluated through observational testing.Formal quantitative metrics, including receiver operating characteristic curve analysis, FAR, FRR, and threshold sensitivity analysis, were not conducted. This constitutes a methodological limitation that is explicitly acknowledged.

A reference image of the authorized participant is converted into a high-dimensional embedding vector. The detected faces are compared with the reference embedding during real-time operation using cosine similarity (Terampe & Arif Pramudwiatmoko, 2025).

$$\text{Similarity} = \| a \| \| b \| \tag{10}$$

Variable description, a: reference face embedding vector, b: Facial embedding vector from the real-time frames, ‖ a ‖: norm (length) of vector a, ‖ b ‖: norma vector b, Similarity: cosine similarity value (range -1 to 1). Face identity verification compares the detected embedding against the reference using the following decision rule:

$$\text{Face match} = \text{Cosine Similarity value} \geq \text{Threshold} \quad (11)$$

The threshold is the minimum similarity value for identity matching. A face mismatch violation is triggered if the cosine similarity value falls below a threshold value.This mechanism ensures identity consistency throughout the examination session.

Gaze-Tracking Module

The gaze-tracking module analyzes eye direction and blink-state using facial landmark detection to monitor the participant's attention (Liu et al., 2024). The module classifies gaze into three states: neutral (forward-facing), looking away, and blinking. Similar to the face recognition module, the gaze tracking performance was characterized through observational testing. Systematic quantitative metrics for blink detection accuracy, gaze deviation detection rate, and false positive rate were not measured, which is a limitation of the current evaluation.

The module applies pupil localization relative to eye corner landmarks to estimate horizontal and vertical gaze directions. Blink detection is implemented using the EAR criterion. When the vertical-to-horizontal eye landmark distance ratio falls below a predefined threshold for a minimum number of consecutive frames, a blink event is registered (Murjitama et al., 2025). A VIOLATION is triggered when either 'eye_looking_away' or 'blinking' is detected persistently beyond a configurable temporal threshold. A blink event is registered when EAR falls below a predefined threshold (EAR_threshold) for a minimum number of consecutive frames.A VIOLATION flag is triggered when either 'eye_looking_away' or 'blinking' persists beyond a configurable temporal duration threshold, indicating sustained inattention rather than momentary eye movement. The accuracy of this module depends on the facial landmark localization quality, which is sensitive to webcam resolution, camera angle, and ambient lighting conditions. Formal quantitative characterization of this module, including the blink detection accuracy rate, gaze deviation detection rate, confusion rate between gaze states, and false positive rate under neutral conditions, was not conducted in this study and is identified as a direction for future evaluation.

Integration of the Real-Time Proctoring System

The trained YOLOv8 model was integrated into a real-time proctoring application that processes live webcam streams. The system combines three modules: (1) YOLO-based

behavior detection, (2) facial identity verification using InsightFace embeddings, and (3) gaze tracking for eye movement analysis. These modules operate asynchronously using interval-based inference scheduling to maintain high frame rates. Each incoming video frame undergoes selective analysis based on predefined intervals. Heavy computations , such as object detection and facial recognition, are executed periodically, while cached results are reused between intervals and reduce latency.

Cosine similarity was used to compare the detected face embedding with a reference image (Kim et al., 2023). Violations are triggered when similarity falls below a defined threshold or abnormal behaviors are detected. Several optimization strategies were implemented to ensure stable high-speed processing and smooth real-time visualization while maintaining detection reliability: GPU-accelerated inference, mixed-precision computation, frame caching and asynchronous rendering, reduced buffer latency, and selective frame analysis (SFA).

Results and Discussion

Performance of the training model

The YOLOv8s model was trained over 200 epochs on 1,056 curated training images across four behavioral classes, with 264 images reserved for validation. This training setup directly addresses the first research objective: developing a real-time behavioral classification module capable of distinguishing between normal and cheating-related face and object states. The training process shows stable convergence across all loss components, as shown below.

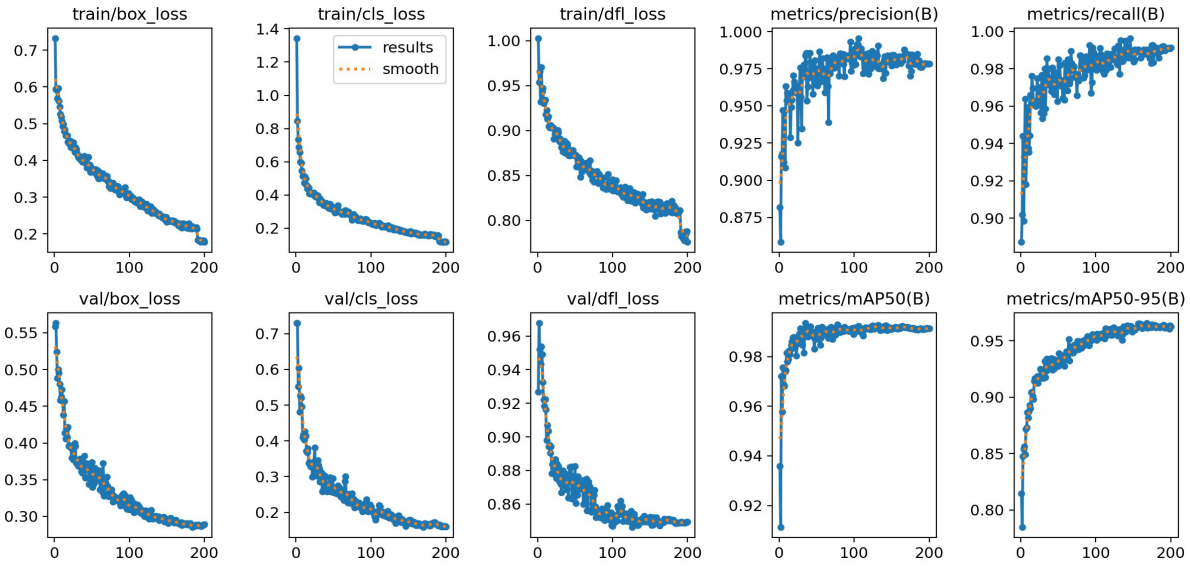

Figure 3 results.png as YOLOv8 training curves

Both training and validation losses decreased consistently throughout training. At the best-performing checkpoint (epoch 168), the training and validation box losses were 0.222 and 0.289, respectively, with a ratio of approximately 1.30, whereas the classification losses were

nearly equal (train: 0.161, val: 0.164). This small and consistent gap between training and validation losses provides supporting quantitative evidence that severe overfitting did not occur, although formal external validation would be required to confirm generalization to unseen populations. The loss curves flattened in the final training epochs, indicating stable convergence. The mean average precision summarizes the overall detection quality as follows:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (12)$$

Variable description : N: number of object classes (N = 4), AP_i : Average Precision for class I, mAP : mean of AP values across all N classes. Then, the mAP is computed as the arithmetic mean of AP values across all classes, representing the overall detection quality. A high mAP indicates that the detector achieves high accuracy and consistency across multiple behavioral classes. Here, N is the number of classes and $AP_i$is the average precision for class i. The best performing epoch was recorded at epoch 168, which was selected based on the maximum mAP@50. Table 1 summarizes the results of quantitative performance.

Table 4 Best training performance metrics

Metric	Value
Precision	0.9856
Recall	0.9903
mAP@50	0.9918
mAP@50–95	0.9656
Best epoch	168 / 200

These metrics reflect the strong detection performance of the validation set under the described controlled training conditions. The close alignment of mAP@50 (0.9918) and mAP@50–95 (0.9656) indicates that the model maintains consistent localization accuracy across varying IoU thresholds. As noted, the performance is reported on the validation split only an independent test set was not used, and generalization claims to heterogeneous real-world conditions are appropriately reserved for future work.

Visual model evaluation

The performance of the proposed model is evaluated using standard OD metrics. Precision measures the proportion of correct detections among all detections. A high precision value indicates that most detections produced by the model are accurate and that the system rarely generates false alarms. The outputs generated during validation, including labeled samples, predicted samples, and confusion matrices , provide clear evidence of the model's reliability.

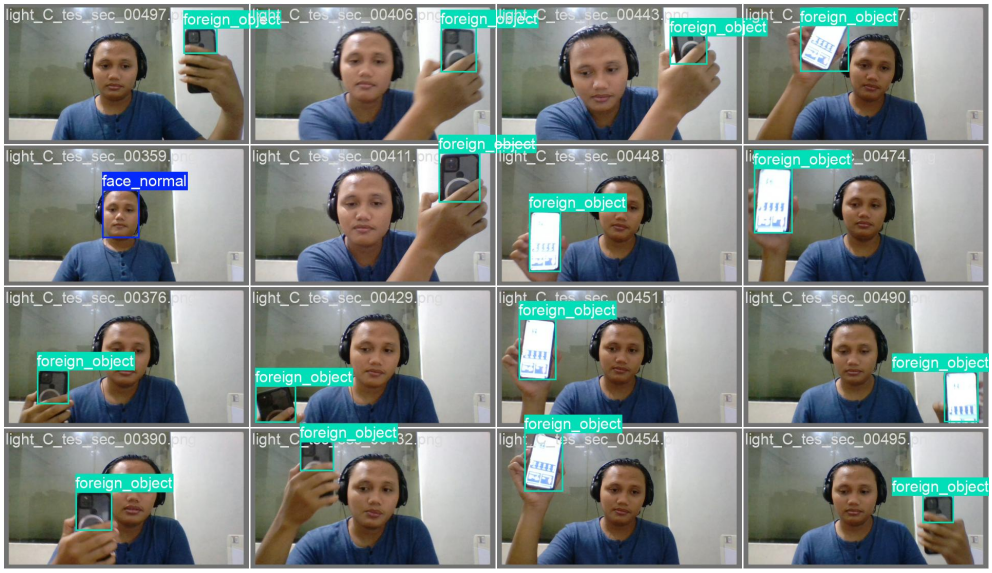

Figure 4 Validation of samples with ground truth labels (val_batch0_labels)

Ground truth annotation samples confirm that the dataset was consistently labeled with clearly defined bounding boxes and correct class assignments across behavioral categories. The bounding boxes and class labels are clearly defined, providing a reliable reference for evaluating prediction quality.

Figure 5 Validation of samples with model predictions (val_batch0_pred)

Model predictions show strong spatial alignment with ground truth labels, with accurately localized bounding boxes and consistent class assignments, confirming that the YOLOv8s model learned discriminative features for all four behavioral classes. The bounding boxes are accurately localized, and class assignments are consistent, indicating that the YOLOv8s model successfully learned the object features during training.

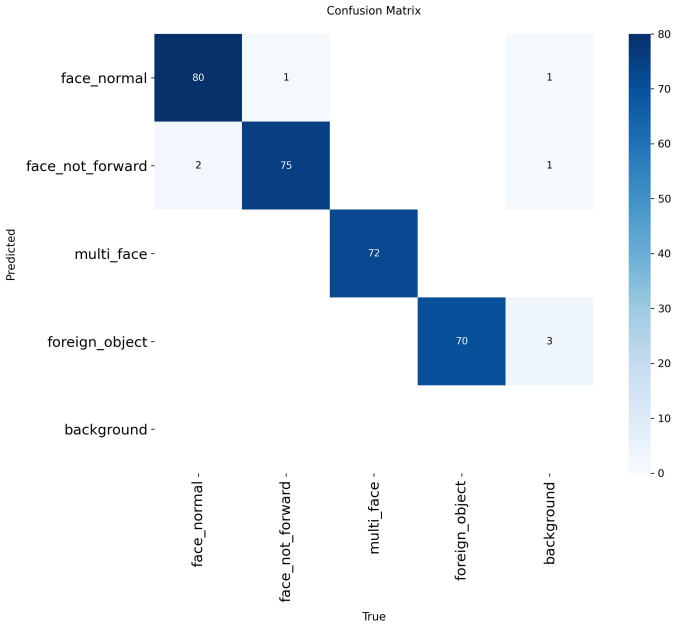

Figure 6 the Confusion matrix

The confusion matrix's diagonal dominance confirms strong class-level accuracy. Minor confusion is observed between 'face_normal' and 'face_not_forward'. This is attributable to the visual similarity of near-frontal head orientations: a marginally off-axis face may produce features resembling both the normal and non-forward' class distributions under certain lighting conditions or with slight camera angles. This confusion could occasionally produce false VIOLATION flags for low-severity errors for participants with slight head tilts in a practical proctoring context given that proctors can review flagged events.Nonetheless, this highlights the importance of stricter annotation boundaries between two classes for iterating the dataset.

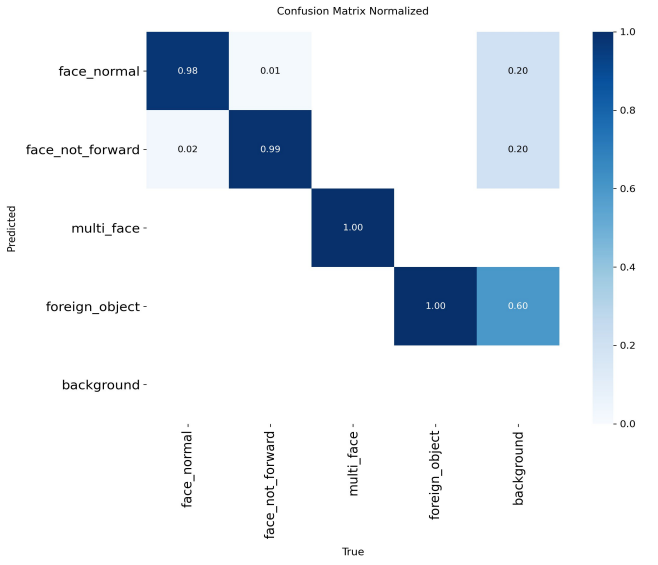

Figure 7 Normalized confusion matrix (NCM)

The normalized confusion matrix reveals that the foreign_object of the most populated class (896 raw images, 323 after curation) achieves near-perfect recall without apparent dominance bias, suggesting that the model learned sufficiently discriminative features for all four classes despite the class imbalance. Occasional confusion between 'multi_face' and 'foreign_object' may occur when objects are positioned near the face region, creating a partial spatial overlap. This class imbalance was not explicitly corrected through oversampling or loss weighting . Future studies should investigate whether balanced sampling can further reduce residual confusion. Overall, the confusion matrix analysis confirms that the interclass misclassification rates are low across all classes, supporting the trained model's suitability for behavioral detection in proctoring applications.

Face recognition performance

Each detected face was compared to the stored reference embedding via cosine similarity during operation.The cosine similarity threshold was empirically set to 0.5 through iterative observational testing: values above 0.5 consistently corresponded to the registered participant, while values below 0.5 reliably indicated unauthorized or occluded faces in the test environment. It is explicitly acknowledged that no systematic FAR/FRR analysis, ROC curve evaluation, or similarity score distribution analysis was conducted in this represents a significant limitation for security-sensitive deployment and is a priority for future evaluation.

Observational testing demonstrated that the face recognition module demonstrated consistent identity verification behavior under the experimental conditions. When the registered participant's face was clearly visible and forward-facing, similarity scores consistently exceeded 0.5, resulting in stable NORMAL identity status). The scores dropped below 0.5 when an unregistered or significantly occluded face was presented, triggering a mismatch warning (Figure 1.

Figure 8 Face mismatch detection of warning indications (FMD)

This behavior confirms that the embedding-based verification method can distinguish registered from unregistered participants under the tested conditions. However, the reliability of the chosen threshold under diverse real-world conditions (e.g., varied lighting, expressions, camera angles, and demographic diversity) cannot be characterized without formal FAR and FRR measurements.

Gaze tracking and analysis of eye behavior

Observational testing of the gaze tracking module showed that prolonged blink events and major gaze deviations from the screen were consistently detected during the test sessions. Blink detection was the most stable behavior for identifying sustained eye closure, providing reliable detection of disengagement events.

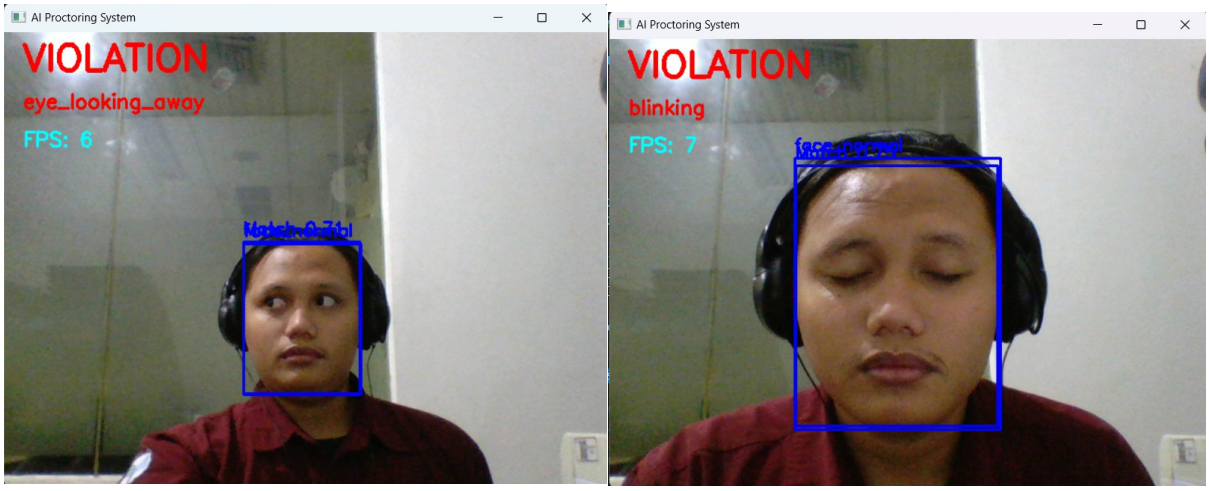

Figure 9 Detection of eye looking away and blinking behavior during real-time monitoring

Fine-grained gaze direction detection exhibited greater variability, particularly under dim lighting and at oblique camera angles. Large gaze deviations were reliably detected until subtle gaze shifts were inconsistently detected. This study did not measure systematic quantitative metrics, including blink detection accuracy rate, gaze deviation detection rate, and false positive rate, which limits the performance of this module's formal characterization. Whether gaze tracking improves the overall system violation detection rate compared with a YOLO+InsightFace baseline was not empirically tested. This comparative evaluation is identified as a key direction for future research.

Integration of the Real-Time Proctoring System

The trained YOLOv8s model is integrated into a real-time proctoring program that combines face recognition and gaze tracking. The system processes webcam input and evaluates behavioral compliance in real time.

According to performance benchmarking, the system operates at an average of approximately 10 frames per second, with observed values ranging between 6 and 15 frames per second. The detection categories targeted in this system were sustained behavioral violations, such as changes in head orientation, presence of persistent foreign objects, and gaze deviation lasting multiple seconds. 10 FPS provides sufficient temporal sampling. However, brief transient events (rapid phone glimpses and momentary blinks) may occasionally be missed at the lower end of the FPS range, indicating a tradeoff in detection coverage. Table 1 summarizes the hardware specifications and performance results.

Table 5 Real-Time Benchmark Performance

Metric	Value
Average FPS	10
Minimum FPS	6
Maximum FPS	15
Hardware — CPU	Ryzen 7 (gaming laptop)
Hardware — RAM	8 GB
Hardware — GPU	NVIDIA GPU RTX 3050 Ti, CUDA device 0
GPU — VRAM	4 GB
Primary bottleneck	CPU scheduling and I/O (GPU utilization remains low)
Inference strategy	Interval-based scheduling (async module execution)

GPU use remained consistently low during testing, indicating that performance is CPU-bound rather than GPU-bound. The interval-based inference strategy that executes each detection module at staggered frame intervals is essential for maintaining stability on single GPU gaming hardware. This strategy reduces peak GPU load but introduces inter-frame latency per module at 10 FPS. The YOLO detection interval determines the maximum temporal resolution for behavioral classification, which is acceptable for sustained violation detection but insufficient for sub-second event capture. Module-level latency profiling was not separately measured in this study.

The final system aggregates signals from all modules to determine the global status of either ‘NORMAL’ or ‘VIOLATION’. This OR-fusion strategy maximizes violation recall but may produce false positives when individual module errors coincide. Each violation trigger type was consistently activated during experimental testing when the corresponding anomalous behavior was intentionally presented. Instead of relying solely on object detection and recognition, the system evaluates identity verification object detection and attention pattern simultaneously, resulting in more comprehensive proctor solution.

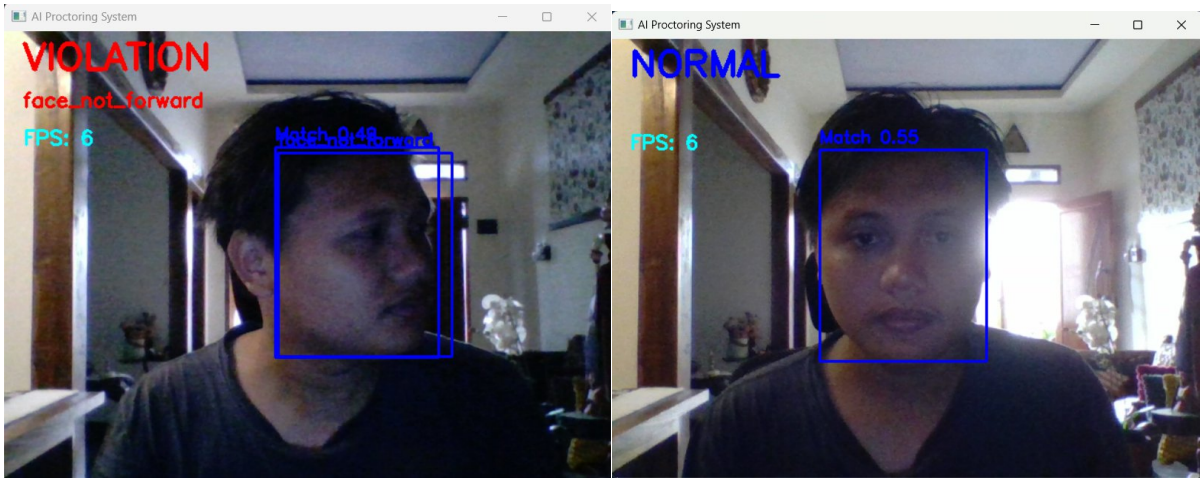

Figure 10 Capture of violation and normal behavior status

The design enables responsive real-time monitoring under the tested single-user hardware configuration. Scalability to multi-user or server-hosted environments was not evaluated and represents a significant open research question. Contextualizing these results within the current state of the art: Single-modality YOLO-based proctoring systems achieve object detection for specific violation types but do not address identity verification or attentional monitoring. Standalone face recognition systems address impersonation but cannot detect object-based cheating or gaze behavior. The proposed three-layer architecture simultaneously addresses all three detection dimensions within a single real-time pipeline.

Regarding inference efficiency, the interval-based scheduling enables 10 FPS operation on a Ryzen 7 procie gaming laptop , which is comparable to or exceeds the reported inference speeds of multi-GPU-dependent systems for single-user monitoring. Direct quantitative accuracy comparison is constrained by differences in dataset classes, collection conditions, and evaluation protocols across studies. However, the mAP@50 of 0.9918 for the YOLO layer is comparable to or superior to YOLOv5-based proctoring systems reported in recent literature, acknowledging that these comparisons are performed on different datasets.

Conclusions

This study developed and evaluated a prototype of a functional intelligent online exam proctoring system that integrates three detection layers. YOLOv8s for face and object behavioral classification, InsightFace for embedding-based identity verification, and GazeTracking for eye behavior monitoring. The system addresses multiple violation categories within a unified real-time pipeline, such as abnormal head orientation, multiple faces, identity mismatch, and abnormal eye behavior. The YOLOv8s model achieved strong detection performance on the validation set, with a peak at epoch 168 recording precision of 0.9856, recall of 0.9903, mAP@50 of 0.9918, and mAP@50–95 of 0.9656. The quantitative

convergence of training and validation losses at the best checkpoint is reflected in a box loss ratio of approximately 1.30. This study provides supporting evidence against severe overfitting on the training data, and the results of confusion matrix analysis indicate minimal interclass misclassification. The face recognition module demonstrated a clear behavioral separation between registered and unregistered participants, with identity-confirmed sessions consistently producing cosine similarity scores at or above 0.49, while an unregistered participant produced a score as low as 0.03 against the reference embedding. The gaze tracking module reliably detected prolonged blink events and large gaze deviations under the tested conditions. The integrated system operated at an average of approximately 10 FPS on consumer-grade hardware through an interval-based inference scheduling strategy, demonstrating feasibility for single-user real-time proctoring without specialized infrastructure.

However, several limitations must be acknowledged. The dataset was collected from a limited number of subjects under controlled indoor conditions, restricting the formal assessment of generalization to heterogeneous real-world deployment scenarios. No independent test set was held out, meaning all reported performance metrics reflect validation-split results only, and generalization to fully unseen populations cannot be formally established from this study alone. Furthermore, the face recognition and gaze tracking modules were evaluated only through observational prototype testing. In this study, formal quantitative metrics, including FAR, FRR, ROC curve analysis, blink detection accuracy rate, gaze deviation detection rate, and confusion rate between gaze states, were not measured.No systematic hyperparameter optimization or comparative model experiments were conducted, and system scalability under multi-user or adversarial conditions was not evaluated.

Future work should address these limitations by expanding the dataset to include multiple subjects with diverse demographic characteristics to improve model generalization, and by conducting formal quantitative evaluation of the InsightFace and GazeTracking modules through controlled experiments with annotated ground-truth video sequences. Systematic hyperparameter tuning and comparative model experiments, such as ablation studies between YOLOv5 and YOLOv8s should be performed to formally justify architecture choices. Additionally, integrating higher-resolution cameras and exploring GPU-optimized deployment pipelines are recommended to improve inference speed, while evaluation under real online examination conditions with multiple concurrent users would provide a more rigorous assessment of system scalability and detection robustness beyond the current controlled prototype setting.

References

Abdurrasyid, indrianto, & susanti, m. N. I. (2022). Face detection and global positioning system on a walking aid for people with blindness bulletin of electrical engineering and informatics, 11(3), 1558–1567. Https://doi.org/10.11591/eei.v11i3.3429

Ajayi, o. G., ibrahim, p. O., & adegboyega, o. S. (2024). Effect of hyperparameter tuning on the performance of yolov8 for multi crop classification on uav images. Applied sciences (switzerland), 14(13). Https://doi.org/10.3390/app14135708

Drantantiyas, n. D. G., yulita, w., ridwan, n. T., ramadhani, u. A., kesuma, r. I., rakhman, a. Z., bagaskara, r., miranto, a., & mufidah, z. (2023). Performasi deteksi jumlah manusia menggunakan yolov8. Jasiek (jurnal aplikasi sains, informasi, elektronika dan komputer), 5(2), 63–68. Https://doi.org/10.26905/jasiek.v5i2.11605

Essahraui, s., lamaakal, i., maleh, y., makkaoui, k. El, bouami, m. F., ouahbi, i., almousa, m., alqahtani, a. A. S., & abd el-latif, a. A. (2025). Deep learning models for detecting cheating in online exams. Computers, materials and continua, 85(2), 3151–3183. Https://doi.org/10.32604/cmc.2025.067359

Gündüz, m. Ş., & işik, g. (2023). A new yolo-based method for social distancing from real-time videos. Neural computing and applications, 35(21), 15261–15271. Https://doi.org/10.1007/s00521-023-08556-3

Gusai, o. P., rani, a., & yadav, p. (2023). Global higher education and the covid-19 pandemic (1st editio, p. 19). Apple academic press. Https://www.taylorfrancis.com/chapters/edit/10.1201/9781003328582-8/covid-19-era-shift-conventional-education-system-learning-virtual-learning-distance-education-om-prakash-gusai-ankur-rani-preeti-yadav

Gusniwati, m., & rahmawati, e. Y. (2024). Pengaruh kemampuan awal dan minat belajar terhadap hasil belajar kalkulus mahasiswa teknik informatika itpln jakarta. Original research, 80, 37–44.

Hasibuan, e., & hendrik, b. (2025). Perbandingan metode deep learning dalam deteksi kekerasan fisik berbasis video : studi literatur pada cnn ,. 0738(4), 980–988.

Islam, m. M. (2023). Real-time dataset of pond water for fish farming using iot devices. Data in brief, 51, 4–10. Https://doi.org/10.1016/j.dib.2023.109761

Jatnika, h., luqman, l., nur, m., susanti, i., andriyani, p., & wibisono, m. (2025). Enrichment : journal of multidisciplinary research and development application of multiple linear regression (mlr) method in. 2(12), 1–10.

Jatnika, h., rifai, m. F., & primadhani, v. R. (2023). Application of iso/iec 9126 on quality measurement of web-based records management applications at itcc itpln. Jurnal scientia, 12(03), 2665–2676.

Https://www.academia.edu/download/111466674/application_of_isoiec_9126.pdf

Juliandy, c., wong, n. P., & darwin. (2024). Modeling face detection application using convolutional neural network and face-api for effective and efficient online attendance tracking. Jurnal online informatika, 9(1), 10–17. Https://doi.org/10.15575/join.v9i1.1203

Khoiriyah, h., abdillah, f., nurfal aziz, a., & gede wiryawan, i. (2025). Telematika violence and robbery detection system using yolov5 algorithm based on iot technology. Telematika, 18(2), 121–133. Http://ejournal.amikompurwokerto.ac.id/index.php/telematika/http://dx.doi.org/10.35671/telematika.v18i2.3088

Kim, h. Bin, choi, n., kwon, h. J., & kim, h. (2023). Surveillance system for real-time high-precision recognition of criminal faces from wild videos. Ieee access, 11(june), 56066–56082. Https://doi.org/10.1109/access.2023.3282451

Kljucaric, l., & george, a. D. (2023). Deep learning inferencing with high-performance hardware accelerators. Acm transactions on intelligent systems and technology, 14(4). Https://doi.org/10.1145/3594221

Korah, p., & varghese, c. (2024). Ir face detection and recognition using yolov8 and facenet. Ignitarium. Https://ignitarium.com/ir-face-detection-and-recognition-using-yolov8-and-facenet/

Liu, j., chi, j., & yang, z. (2024). A review on personal calibration issues for video-oculographic-based gaze tracking. Frontiers in psychology, 15. Https://doi.org/10.3389/fpsyg.2024.1309047

Murjitama, f. L., p, u. P. S., widodo, s. A., dwijayanti, s. A., d, q. F., & purwanto, y. S. (2025). Blink detection sensor sebagai pembantu komunikasi pasien stroke berat berbasis internet of things (iot). 9(2), 1952–1958.

Nur aziz thohari, a., fathul lathief, m., triyono, l., & santoso, k. (2025). Deteksi kecurangan ujian pada ruangan tertutup menggunakan algoritma yolov8. Journal of computer science and informatics engineering, 4(2), 61–71. Https://doi.org/10.55537/cosie.v4i2.1100

Okyere-gyamfi, s., asante, m., peasah, k. O., missah, y. M., & akoto-adjepong, v. (2025). Contrast limited adaptive histogram equalization (clahe) and colour difference histogram (cdh) feature merging capsule network (ccfmcapsnet) for complex image recognition. Plos one, 20(10 october), 1–27. Https://doi.org/10.1371/journal.pone.0335393

Potluri, t., venkatramaphanikumar, s., & venkata krishna kishore, k. (2023). An automated online proctoring system using attentive-net to assess student mischievous behavior. Multimedia tools and applications, 82(20), 30375–30404.

Https://doi.org/10.1007/s11042-023-14604-w

Purwanto, y. S., & putra, r. I. (2025). Real-time multi-screen cheating detection using k-means clustering. Journal of information systems and informatics, 7(3), 2758–2779. Https://doi.org/10.51519/journalisi.v7i3.1262

Putra, r. F., & mulyana, d. I. (2024). Optimasi deteksi objek dengan segmentasi dan data augmentasi pada hewan siput beracun menggunakan algoritma you only look once (yolo). Jurnal jtik (jurnal teknologi informasi dan komunikasi), 8(1), 93–103. Https://doi.org/10.35870/jtik.v8i1.1391

Putu ary sri tjahyanti, l., santo gitakarma, m., & korespondensi, p. (2024). Exam fraud detection system using yolo: identifying mobile phone usage and suspicious interactions. Jurnal komputer dan teknologi sains (komteks), 3(2), 25–32.

S. Purwanto, y., farid rifai, m., & jatnika, h. (2022). A comparison between offline and multimodal online platforms at english standardization tests for college students. Proceedings of the international conference on sustainable innovation on humanities, education, and social sciences (icosi-hess 2022), 178–192. Https://doi.org/10.2991/978-2-494069-65-7

Septiandi, l. A., yuniarno, e. M., & zaini, a. (2021). Deteksi kedipan dengan metode cnn dan percentage of eyelid closure (perclos). Jurnal teknik its, 10(1). Https://doi.org/10.12962/j23373539.v10i1.61174

Terampe, g. C., & arif pramudwiatmoko. (2025). Facial recognition performance evaluation with yolov8, arcface, and svm in a contactless employee attendance system. Jurnal riset informatika, 8(1), 108–120. Https://doi.org/10.34288/jri.v8i1.465