International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Performance Analysis of Smart Technology with Face Detection using YOLOv3 and InsightFace for Student Attendance Monitoring

Muhammad Fikry*[1], Taufiqurrahman[2], Munirul Ula[3], Fadlisyah[4], Nurdin[5], Muhammad Yani[6], Aldy Anugrah Pohan[7]

Submitted:13/03/2024 **Revised**: 28/04/2024 **Accepted**: 05/05/2024

Abstract*:* In this paper, we propose a comparative evaluation for smart technologies of two state-of-the-art face detection models, YOLOv3 and InsightFace (SCRFD-10GF), within a classroom environment for real-time monitoring applications. The primary objective of this study is to assess and compare the models' detection accuracy, robustness, and computational efficiency across various settings, including different camera positions and lighting conditions. We employ a dataset consisting of videos captured from three distinct classrooms—Lab 1, Lab 3, and Room 1—each presenting unique challenges such as obstructions (e.g., computers) and varying angles and lighting conditions. The study aims to address the challenge of comparing these models in real-world environments with demanding conditions. The results reveal that YOLOv3 consistently outperforms InsightFace in terms of confidence scores across all environments and camera positions. YOLOv3's superior architecture, featuring multi-scale detection and advanced feature extraction capabilities, enables it to maintain high accuracy and confidence, with an average confidence score reaching 0.83. InsightFace, though slightly less accurate, is advantageous in resource-constrained settings due to its lightweight architecture. The findings suggest that YOLOv3 is ideal for systems requiring high accuracy, while InsightFace is better suited for environments with limited computational resources. We conclude that a hybrid approach leveraging both models could offer a balanced solution tailored to specific requirements of educational environments.

Keywords: *YOLOv3, InsightFace, Face Detection, Classroom Monitoring, Real-time Monitoring*

1. Introduction

In recent years, the rapid advancement of technology has led to the development of innovative tools that significantly enhance various aspects of educational environments. Among these, face detection technology has emerged as a powerful tool for monitoring classroom activities, offering a non-intrusive method to observe student behavior and engagement during lectures. The ability to automatically detect and track students' faces in real-time not only helps in maintaining accurate attendance records but also provides valuable insights into student attentiveness, participation, and overall engagement in the learning process [1].

The implementation of face detection systems in classrooms, however, is fraught with challenges [2]. Unlike controlled environments, classrooms present dynamic and unpredictable conditions that can affect the accuracy and reliability of face detection. Variations in student positions, changes in lighting, occlusions caused by other students or objects, and differences in camera angles all contribute to the complexity of the task. These factors can lead to significant variations in detection performance, making it crucial to select or design models that can handle these challenges effectively.

The importance of accurate face detection in classrooms cannot be overstated. It plays a critical role in various educational applications, including automated attendance systems, real-time behavioral analysis, and personalized learning experiences. For instance, in large classes, where it is difficult for educators to monitor every student individually, face detection systems can help ensure that students are present and engaged. Moreover, in the context of remote learning, where direct interaction is limited, these systems can provide educators with essential feedback on student participation and attention.

Given the importance of the task, this study focuses on evaluating the performance of two state-of-the-art face detection models: YOLOv3 and InsightFace (SCRFD-10GF). YOLOv3, known for its speed and accuracy in object detection tasks, has been widely adopted in various real-time applications. InsightFace, on the other hand, represents a family of deep learning models specifically optimized for face detection and recognition, offering high accuracy even in challenging conditions.

The dataset used in this study consists of videos of students captured in a real classroom setting from three distinct angles: left, right, and center. These videos were recorded in three different environments: Lab1, Lab3, and Room1, each with its own unique set of lighting and spatial

[1] *Informatics, Universitas Malikussaleh, Lhokseumawe - INDONESIA*
ORCID ID : 0000-0003-1001-402X
[2] *Software Engineering Technology, Wilmar Bisnis Indonesia Polytechnic, Medan - INDONESIA*
ORCID ID : 0009-0005-2481-2032
[3] *Informatics, Universitas Malikussaleh, Lhokseumawe - INDONESIA*
[4] *Information Technology, Universitas Malikussaleh, Lhokseumawe - INDONESIA*
[5] *Information Technology, Universitas Malikussaleh, Lhokseumawe - INDONESIA*
[6] *Information Technology, Universitas Malikussaleh, Lhokseumawe - INDONESIA*
[7] *Informatics, Universitas Malikussaleh, Lhokseumawe - INDONESIA*
* *Corresponding Author Email: muh.fikry@unimal.ac.id*

characteristics. The original video resolution of 1920×1080 at 60 frames per second (fps) was downscaled to 1280×720 at 25 fps to simulate typical conditions for classroom surveillance, balancing the need for detail with the computational demands of real-time processing.

The primary objective of this research is to conduct a comprehensive comparison of YOLOv3 and InsightFace in terms of detection accuracy, robustness, and computational efficiency within the context of classroom monitoring. By systematically evaluating these models under varying conditions, this study aims to identify the most suitable approach for deploying face detection systems in educational settings. The findings of this research are expected to contribute to the development of more effective and reliable monitoring systems, ultimately enhancing the quality of education by providing educators with better tools for student engagement analysis.

Furthermore, this study explores the potential implications of face detection technology in broader educational contexts, such as its role in adaptive learning systems, with real-time feedback on student engagement can inform the delivery of personalized content. The comparison of YOLOv3 and InsightFace in such a nuanced and demanding application as classroom monitoring is thus not only a technical exercise but also an exploration of the future of educational technology. The findings contribute to the development of more effective educational systems and offer valuable insights into the future of educational technology.

2. Related Works

The advent of deep learning has significantly advanced the field of face detection. Convolutional Neural Networks (CNNs) have become the cornerstone of modern face detection, providing robust performance across a wide range of conditions. YOLO (You Only Look Once) models, represent a paradigm shift in object detection by framing it as a single regression problem, predicting bounding boxes and class probabilities directly from full images in one evaluation [3]. YOLOv3, introduced significant improvements in accuracy through the use of a deeper network (Darknet-53) and multi-scale detection [4], [5]. YOLOv3 is particularly well-suited for real-time applications due to its balance between speed and accuracy, making it a popular choice for face detection in dynamic environments [6].

In a comparative study, YOLOv3 was found to outperform traditional methods and earlier versions of YOLO in terms of both speed and accuracy [7], [8]. Its ability to detect objects at three different scales enhances its performance, particularly for smaller objects [9], which is critical in scenarios like classroom monitoring where faces may appear at varying sizes depending on the camera angle.

InsightFace is a well-known library for face detection and recognition [10], [11], with the SCRFD model being one of its recent advancements. SCRFD is designed for high-performance face detection, optimized for real-time applications with minimal computational requirements [12]. The model uses a ResNet-based architecture with custom layers to achieve high detection accuracy while remaining lightweight enough for deployment on resource-constrained devices [13].

SCRFD has demonstrated competitive performance against other state-of-the-art models, such as RetinaFace and DSFD, particularly in scenarios involving challenging conditions like occlusions and extreme poses [12], [14]. SCRFD's ability to deliver high accuracy while maintaining low inference times, making it ideal for applications requiring both speed and precision [13].

Real-time classroom monitoring is a critical application area for face detection technologies. Accurate and efficient face detection enables effective monitoring of student engagement, attendance, and safety. The choice of detection model must consider both the dynamic nature of classroom environments and the resource constraints of the deployment hardware. The use of deep learning-based face detection models for classroom monitoring, emphasizing the importance of high detection accuracy in environments with varying lighting conditions, occlusions, and camera angles [15].

3. Materials and Methods

3.1. Dataset

The dataset for this study comprises video recordings captured in three distinct classroom environments: Lab 1, Lab 3, and Room 1. These environments were specifically selected to simulate typical classroom settings, each presenting unique challenges that could impact the effectiveness of face detection algorithms.

Lab 1 and Lab 3 are both computer laboratories of the same size, furnished with rows of desks and desktop computers. These setups introduce significant physical obstructions, such as monitors and other hardware, which can partially obscure students' faces from the camera's view. The lighting in these labs is primarily artificial, with fluorescent ceiling lights providing uniform illumination. However, the placement and intensity of these lights vary slightly between the two labs, with Lab 3 receiving additional natural light from windows, which introduces variability in brightness and shadow patterns throughout the day.

Room 1 is a general-purpose classroom used for lectures and discussions, featuring standard classroom furniture like desks and chairs. Unlike the labs, Room 1 has fewer obstructions, but the lighting is more variable due to a combination of artificial lights and large windows that allow

natural light to enter. This setup creates challenges for face detection, particularly during times when sunlight directly hits certain areas of the room, creating high-contrast lighting conditions.

The videos were recorded from three different camera positions in each environment: center, right corner, and left corner. In the center position, the camera was placed directly in front of the classroom or lab, facing the students head-on, providing a clear view of most students, though depending on their seating arrangement, some students might still appear in profile. The right corner position had the camera angled from the right side of the room, capturing students from the side or at oblique angles, which added complexity to face detection due to the varied perspectives. Similarly, the left corner position offered a different angle, introducing further diversity in the dataset by creating unique shadow patterns and lighting effects, potentially impacting detection accuracy. Together, these varied positions were chosen to simulate realistic classroom scenarios, testing the robustness of face detection models under multiple conditions.

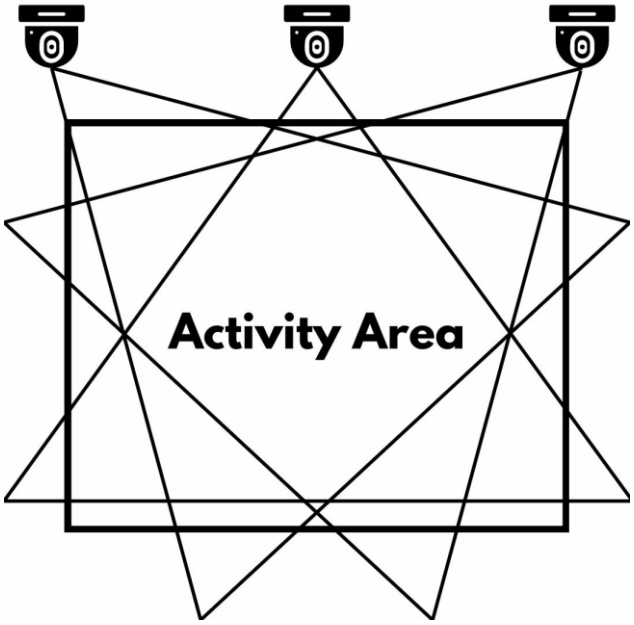


Fig. 1. Schematic representation of camera positions and their overlapping coverage areas within the activity area. Cameras are positioned at the center, right corner, and left corner to capture different perspectives and ensure comprehensive monitoring of the space.

These recordings were made at various times of the day to incorporate natural lighting variations that occur due to the position of the sun and the availability of daylight. For example, Room 1 was recorded in the morning when sunlight was entering through the windows, while Lab 1 and Lab 3 were recorded in the afternoon, primarily under artificial lighting conditions. The time of recording is crucial as it simulates the varying conditions a real classroom surveillance system might encounter during different periods.

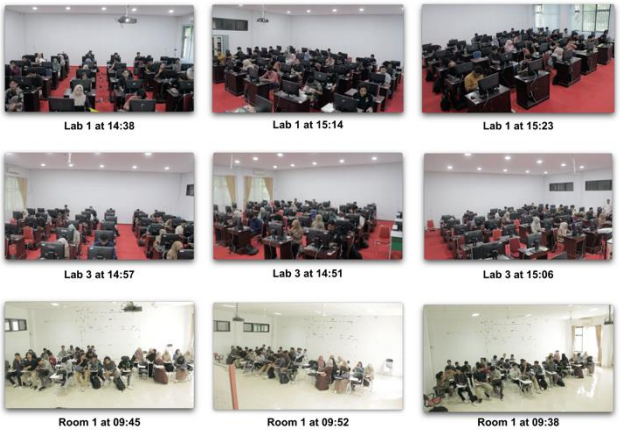


Fig. 2. Classroom and lab environments used for face detection analysis. The images show the setups in Lab 1, Lab 3, and Room 1, with recordings taken at different times and from various camera positions (center, right corner, and left corner).

By capturing data in these varied environments and from different angles and times, the dataset is designed to test the robustness and accuracy of face detection models under realistic and challenging conditions that mirror those found in actual educational settings. The detailed specifications of the video recordings are presented in Table 1.

Table 1. Details of Video Recordings Used in the Study

Room	Recording Time	Duration	Camera Position
Lab 1	14.38	03.07	Center
Lab 1	15.14	03.04	Right corner
Lab 1	15.23	03.02	Left corner
Lab 3	14.57	03.02	Center
Lab 3	14.51	03.01	Right corner
Lab 3	15.06	03.01	Left corner
Room 1	09.45	03.01	Center
Room 1	09.52	03.08	Right corner
Room 1	09.38	03.02	Left corner

3.2. Preprocessing

To prepare the videos for analysis, a series of preprocessing steps were applied to ensure compatibility with the face detection models and to simulate a typical classroom surveillance scenario. The videos were first downscaled from their original resolution of 1920×1080 pixels to 1280×720 pixels. This downscaling was done using bicubic interpolation, a method chosen to preserve image quality while reducing the file size, thereby balancing the need for clear facial features with the computational efficiency

required for real-time processing. Additionally, the frame rate was adjusted from 60 frames per second (fps) to 25 fps, reflecting common surveillance system standards and reducing the computational load, making the data more suitable for real-time face detection in classroom environments. The videos were then split into individual frames, allowing each frame to be analyzed independently by the face detection models, ensuring that every moment of the video is considered in the analysis. Finally, the frames were resized to match the input dimensions required by the specific face detection models used in the study: 416×416 pixels for YOLOv3 and 640×640 pixels for InsightFace (SCRFD-10GF). These preprocessing steps were critical in preparing the dataset for accurate and efficient processing by the models.

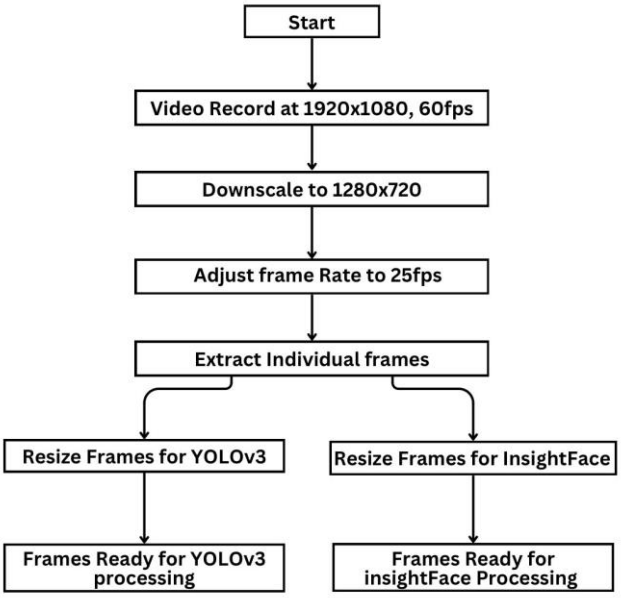


Fig. 3. Preprocessing workflow for video data. The video recordings are downscaled and adjusted in frame rate before being split into individual frames. The frames are then resized to the appropriate dimensions for processing by YOLOv3 and InsightFace models.

3.3. Experimental Setup

The experimental setup was designed to evaluate the performance of two face detection models, YOLOv3 and InsightFace, under realistic classroom conditions. The preprocessed video frames were processed through both models, with each frame analyzed individually to detect faces and generate bounding boxes with associated confidence scores. YOLOv3, configured using the Darknet framework with pre-trained weights on the WIDER FACE dataset, resized input frames to 416x416 pixels and applied a detection threshold of 0.5, alongside non-maximum suppression to eliminate redundant bounding boxes. Similarly, the SCRFD-10GF variant of InsightFace, implemented with pre-trained weights via the InsightFace library, resized input frames to 640x640 pixels and set a detection threshold of 0.4 to enhance recall in challenging conditions. The models were assessed using three key performance metrics: detection accuracy, processing time, and robustness. Detection accuracy was determined by the proportion of correctly detected faces relative to the ground truth, with the average confidence score per frame serving as an accuracy measure. Processing time, recorded in milliseconds for each frame, was crucial for evaluating the real-time applicability of the models. The robustness of the models was further evaluated across different camera angles, lighting conditions, and levels of obstruction, such as partially obscured faces by objects like computer monitors. The outputs, including detected face counts, confidence scores, and processing times, were stored in CSV files for each video, enabling comprehensive analysis. Finally, paired t-tests were performed on the detection accuracy and processing time metrics to assess the statistical significance of performance differences between YOLOv3 and InsightFace.

3.3.1. YOLOv3 Implementation

The YOLOv3 model, widely recognized for its efficiency in real-time object detection, was implemented using OpenCV's Deep Neural Network (DNN) module. YOLOv3, which stands for "You Only Look Once version 3," is a fully convolutional neural network designed to detect objects in images and videos by predicting bounding boxes and class probabilities directly from full images in a single evaluation. YOLOv3 is built on the Darknet-53 architecture, a 53-layer convolutional network pre-trained on the ImageNet dataset [16], [17], [18]. This architecture includes residual skip connections and upsampling, making it more effective at detecting small objects compared to its predecessors.

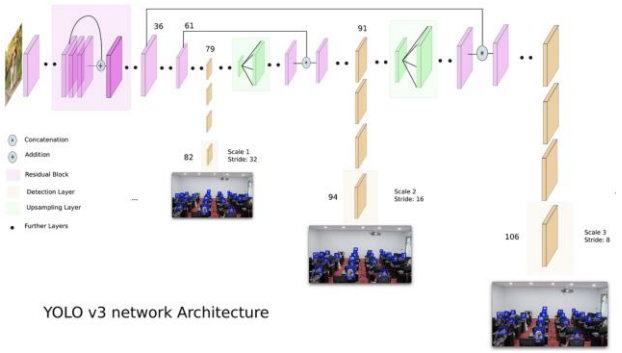


Fig. 4. YOLOv3 Network Architecture

For this study, the YOLOv3 model was initialized by loading pre-trained weights along with the corresponding configuration, both specifically optimized for face detection. The model's layer names were extracted, and the relevant output layers were identified to focus on the detection of faces. During video processing, each video was read frame by frame using OpenCV's VideoCapture, and each frame was resized to half its original size to expedite

the detection process.

The detection process begins by converting each frame into a blob (a preprocessed image) that is fed into the YOLOv3 network. YOLOv3 applies a 1x1 detection kernel to feature maps generated by the network to predict bounding boxes, object confidence scores, and class probabilities. The model outputs potential bounding boxes and confidence scores for any detected faces. The confidence score, which ranges from 0 to 1, indicates the probability that the detected object is indeed a face. In this implementation, faces with confidence scores exceeding a predefined threshold of 0.5 were considered valid detections. The confidence score serves as a measure of the model's certainty and is used to calculate the average confidence across all detections within a frame.

YOLOv3 makes predictions at three different scales, which allows it to detect small, medium, and large objects effectively. This multi-scale detection is achieved by applying the detection kernel at three different layers within the network. The detected bounding boxes and confidence scores for each frame were stored and later used to calculate average confidence scores for the entire video. The metrics recorded included the number of detected faces and the average confidence score per frame. These results were saved to a CSV file, with columns indicating the frame number, total faces detected, and average confidence.

3.3.2. InsightFace Implementation

The InsightFace model, particularly the SCRFD-10GF variant, was employed for real-time face detection in this study. SCRFD (Scalable High-performance Face Detection) is a state-of-the-art face detection algorithm known for its balance of accuracy and efficiency. The SCRFD-10GF model is a lightweight, high-accuracy variant designed to run efficiently on both high-performance and resource-constrained devices [13]. It is based on a custom backbone that incorporates basic residual blocks, making it suitable for real-time applications while maintaining high detection accuracy.

For this implementation, the InsightFace model was initialized using the InsightFace library, specifically with the SCRFD-10GF variant, which uses a custom lightweight backbone architecture. The model was configured with a detection size of 640x640 pixels and was optimized to leverage the best available hardware, for enhanced performance. Similar to the YOLOv3 implementation, each video was processed frame by frame using OpenCV's VideoCapture.

For each frame, the SCRFD-10GF model detected faces by generating bounding boxes and corresponding confidence scores. The detection process involves applying a series of convolutional layers to extract features from the input image, followed by generating bounding boxes around potential faces. The model outputs a confidence score for each detected face, indicating the likelihood that the detected object is a face. Faces with confidence scores above a predefined threshold of 0.4 were considered valid detections. These confidence scores were used as a measure of detection confidence for each frame.

The InsightFace model is optimized for high precision, especially in challenging conditions such as occlusions and varying lighting. It is particularly adept at detecting faces across a wide range of scenarios due to its specialized design for facial feature extraction. The metrics recorded for InsightFace included the total number of faces detected and the average confidence score for each frame. These results were similarly saved to a CSV file, with columns indicating the frame number, total faces detected, and average confidence.

3.4. Procedure

The procedure for this study involved processing the preprocessed video data through the YOLOv3 and InsightFace face detection models and evaluating their performance based on predefined metrics. Each model processed the video frames sequentially, generating bounding boxes and confidence scores for detected faces. The confidence scores, representing the likelihood that a detected object was a face, were stored for all detections above a predefined threshold (0.5 for YOLOv3 and 0.4 for InsightFace). These scores were averaged across all detected faces in each frame to calculate the frame's average confidence. The total number of detected faces and the number of frames processed were also recorded. After the detection process, the results were stored in CSV files for each video, capturing the key metrics of interest. The final analysis involved reading the CSV files to calculate the average total faces detected, average confidence scores, and frame counts. These metrics were then visualized using bar charts, providing a clear comparative evaluation between the YOLOv3 and InsightFace models for each time segment. The visualizations facilitated an in-depth analysis of model performance under varying classroom conditions.

Fig. 5. Workflow of the experimental setup, including video collection, preprocessing, model implementation, and evaluation for YOLOv3 and InsightFace models. The diagram outlines the sequential steps taken from data collection to the final analysis of detection confidence.

4. Results

The performance of the YOLOv3 and InsightFace (SCRFD-10GF) models was evaluated across three different environments: Lab 1, Lab 3, and Room 1, with video recordings taken from various camera positions (center, right corner, and left corner) at different times of the day. The results below present a comparison of the average number of faces detected, the average confidence scores, and the total frame count processed by each model.

Fig. 6. Comparative face detection results for YOLOv3 and InsightFace across different classroom environments and camera positions. The images display the total number of faces detected by each model (YOLOv3 in green and InsightFace in blue) along with their corresponding confidence scores. The experiments were conducted in Lab 1, Lab 3, and Room 1 at various times, with recordings taken from center, right corner, and left corner camera positions.

4.1. Lab 1

In Lab 1, face detection results from both YOLOv3 and InsightFace models were evaluated across three different time slots, captured from center, right corner, and left corner camera positions. The results, as summarized in Table 2, show that YOLOv3 consistently demonstrated higher confidence scores compared to InsightFace, although InsightFace detected a slightly higher number of faces in some instances. For example, at 14:38 from the center camera position, InsightFace detected an average of 18.27 faces with a confidence score of 0.75, while YOLOv3 detected 16.91 faces but with a higher confidence of 0.83.

Table 2. Results of lab 1

Time	Camera Position	Model	Average Total Faces	Average Confidence
14:38	Center	YOLOv3	16.91	0.83
		InsightFace	18.27	0.75
15:14	Right Corner	YOLOv3	18.19	0.78
		InsightFace	17.89	0.68
15:23	Left Corner	YOLOv3	17.01	0.76
		InsightFace	15.22	0.69

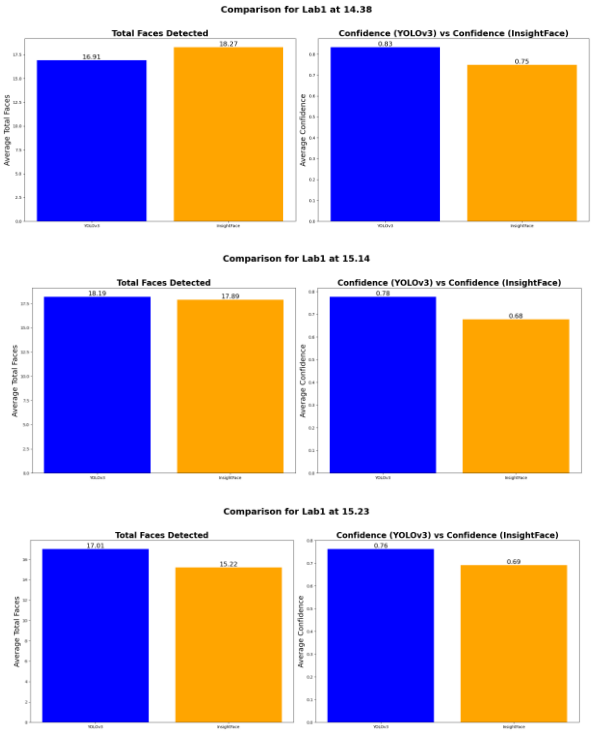


Fig 7. Images showing the experimental setup in Lab 1 across different time slots.

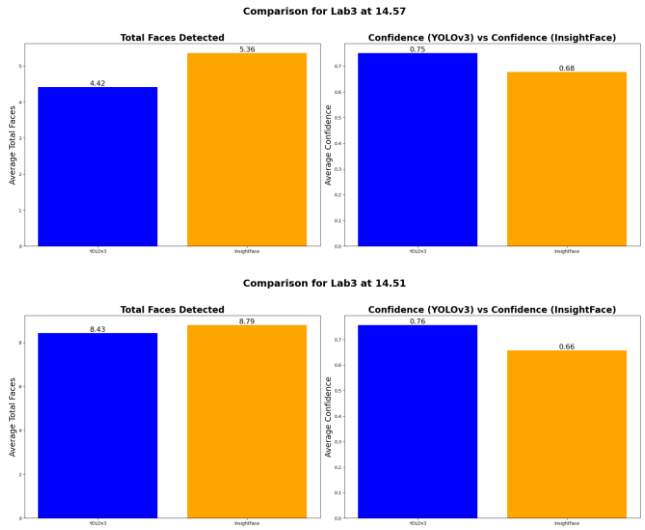
Additionally, the following graphs (Figure 6) illustrate the comparison of average total faces detected and confidence levels between YOLOv3 and InsightFace models across the various camera positions in Lab 1.

4.2. Lab 3

For Lab 3, the face detection analysis was conducted similarly across three different time slots with recordings from the center, right corner, and left corner camera positions. As shown in Table 3, YOLOv3 generally detected fewer faces than InsightFace but maintained higher confidence levels across all positions. For instance, at 14:57 from the center camera position, YOLOv3 detected 4.42 faces with a confidence score of 0.75, whereas InsightFace detected 5.36 faces with a confidence of 0.68.

Table 3. Results of lab 3

Time	Camera Position	Model	Average Total Faces	Average Confidence
14:57	Center	YOLOv3	4.42	0.75
		InsightFace	5.36	0.68
14:51	Right Corner	YOLOv3	8.43	0.76
		InsightFace	8.79	0.66
15:06	Left Corner	YOLOv3	10.33	0.71
		InsightFace	8.01	0.61



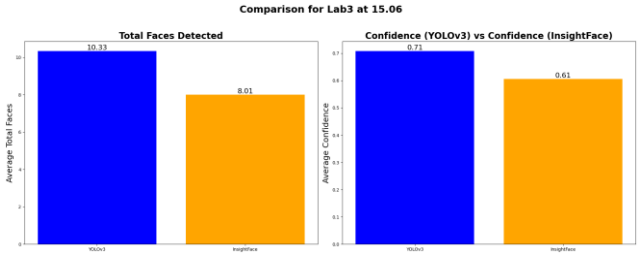


Fig 8. Images showing the experimental setup in Lab 3 across different time slots.

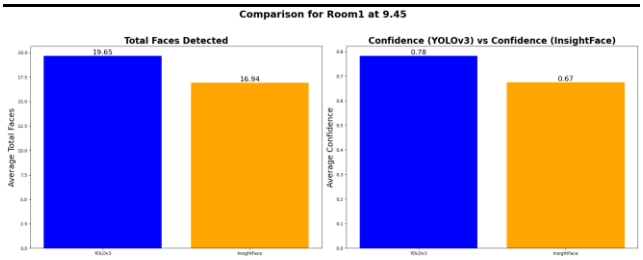
The comparative analysis is further supported by the graphs in Figure 7, which depict the differences in face detection performance between the two models for Lab 3.

4.3. Room 1

In Room 1, the face detection performance of YOLOv3 and InsightFace was analyzed across three time slots with recordings from the center, right corner, and left corner camera positions. As presented in Table 4, YOLOv3 consistently achieved higher confidence scores across all positions compared to InsightFace. However, InsightFace detected fewer faces, particularly in the left and right corner positions. For example, at 9:52 from the right corner position, YOLOv3 detected 17.56 faces with a confidence of 0.74, while InsightFace detected 13.62 faces with a confidence of 0.62.

Table 4. Results of lab 3

Time	Camera Position	Model	Average Total Faces	Average Confidence
9:45	Center	YOLOv3	19.65	0.78
		InsightFace	16.94	0.67
9:52	Right Corner	YOLOv3	17.56	0.74
		InsightFace	13.62	0.62
9:38	Left Corner	YOLOv3	19.87	0.80
		InsightFace	17.97	0.69



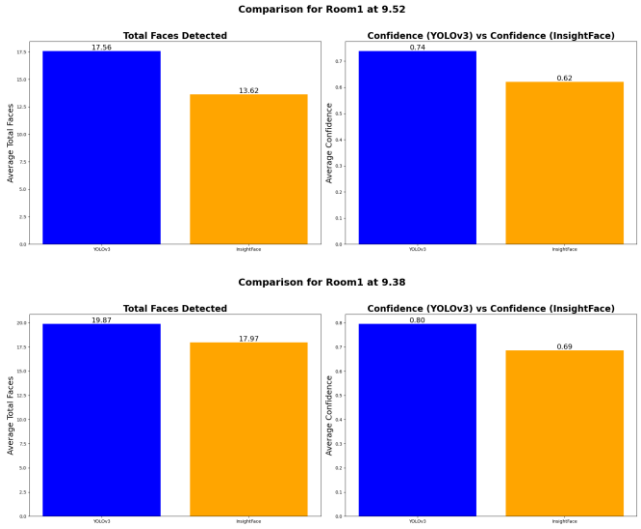


Fig 9. Images showing the experimental setup in Room 1 across different time slots.

The graphical representation in Figure 8 further highlights the detection differences between the two models, providing a clear visual comparison of their performance in Room 1.

5. Discussion

The results from the face detection experiments conducted in three different environments—Lab 1, Lab 3, and Room 1—highlight several key observations regarding the performance of the YOLOv3 and InsightFace (SCRFD-10GF) models. The discussion focuses on three primary aspects: detection accuracy, confidence scores, and the impact of camera positions on model performance.

5.1. Detection Accuracy

The analysis revealed that YOLOv3 consistently achieved the highest detection accuracy across different environments, with its best performance recorded in Room 1 at 09:38 from the left corner camera position. In this scenario, YOLOv3 detected an average of 19.87 faces with a confidence of 0.80, corresponding to an accuracy of 80%. InsightFace, while generally robust, showed its highest detection accuracy in the same environment and time slot, detecting an average of 17.97 faces with a confidence of 0.69, resulting in an accuracy of 69%. The superior performance of YOLOv3 can be attributed to its multi-scale detection capabilities, which allow it to capture both small and large faces, making it more adaptable to varying conditions. In contrast, InsightFace demonstrated limitations in maintaining detection accuracy, particularly in challenging environments like Lab 3, where varying lighting and occlusions presented significant obstacles. Despite these challenges, InsightFace still performed competently, especially in less obstructed settings.

5.2. Confidence Scores

The analysis of confidence scores further underscores the reliability of the detections made by the YOLOv3 model across various environments. The highest confidence score for YOLOv3 was recorded in Room 1 at 09:38 from the left corner camera position, where it achieved an impressive confidence score of 0.80 (80%). This consistently high confidence suggests that YOLOv3's detections were not only accurate but also dependable, reflecting a strong likelihood that the detected objects were indeed faces. In comparison, the InsightFace model recorded its highest confidence score of 0.75 (75%) in Lab 1 at 14:38 from the center camera position. Although slightly lower than YOLOv3, this score still indicates competent performance, particularly in environments with fewer obstructions. The superior confidence scores of YOLOv3 can be attributed to its Darknet-53 architecture, which incorporates residual skip connections and upsampling techniques, enabling the model to retain fine-grained features and make confident detections even in complex scenarios with varying lighting and obstructions.

5.3. Impact of Camera Positions

The camera positions—center, right corner, and left corner—had a noticeable impact on the performance of both models. In general, the center position provided the most consistent results across all environments, with both YOLOv3 and InsightFace performing well in terms of detection accuracy and confidence. This is likely because the center position captures a frontal view of most students, making it easier for the models to identify faces.

In contrast, the right and left corner positions introduced more variability in the results. The oblique angles associated with these positions made face detection more challenging, particularly for InsightFace, which showed a more significant drop in both detection accuracy and confidence scores. YOLOv3, with its multi-scale detection approach, was better able to adapt to these challenging angles, though it still showed some reduction in performance compared to the center position.

5.4. Overall Model Performance

The comparative analysis of YOLOv3 and InsightFace highlights the strengths and weaknesses of each model in the context of classroom face detection. YOLOv3's ability to maintain high detection accuracy and confidence across various environments and camera positions demonstrates its suitability for real-time monitoring applications in dynamic classroom settings. Its higher confidence scores suggest that it is less prone to false positives, making it a reliable choice for environments where accuracy is critical.

InsightFace, while slightly less accurate and confident in some scenarios, remains a valuable tool, particularly in environments with challenging conditions such as occlusions or varied lighting. Its lighter architecture makes it more suitable for deployment on resource-constrained devices, where speed and efficiency may take precedence

over detection accuracy.

5.5. Implications for Real-Time Classroom Monitoring

The findings from this study have significant implications for the deployment of face detection systems in real-time classroom monitoring. Given its superior performance, YOLOv3 could be the preferred choice for scenarios requiring high accuracy and reliability. However, the choice of model may ultimately depend on the specific requirements of the deployment environment, including hardware capabilities, real-time processing needs, and the typical conditions of the classroom.

In environments where hardware limitations are a concern, InsightFace's more efficient design may be advantageous, even if it means sacrificing some accuracy and confidence. For schools and institutions seeking a balance between performance and resource efficiency, a hybrid approach that leverages both YOLOv3 and InsightFace could be considered, with each model deployed based on the specific conditions of the classroom.

6. Conclusion

This study evaluated the performance of the YOLOv3 and InsightFace (SCRFD-10GF) models for face detection across different classroom environments, including Lab 1, Lab 3, and Room 1. The analysis was based on key metrics such as the total number of faces detected and the average confidence scores, with recordings taken from various camera positions (center, right corner, and left corner) at different times of the day.

The findings indicate that YOLOv3 consistently outperformed InsightFace in terms of confidence scores across all environments and camera positions. YOLOv3's superior architecture, featuring multi-scale detection and advanced feature extraction capabilities, allowed it to maintain high detection accuracy and confidence, even in challenging scenarios with oblique camera angles and varying lighting conditions. This makes YOLOv3 a robust choice for real-time face detection in classroom settings, where reliability and accuracy are paramount.

InsightFace, while demonstrating strong detection capabilities, particularly in environments like Lab 3 with challenging conditions, generally recorded lower confidence scores compared to YOLOv3. However, its lightweight design and efficiency make it a viable option for deployment on devices with limited computational resources. This makes InsightFace a valuable tool in scenarios where speed and resource efficiency are prioritized over maximum detection accuracy.

The results suggest that while YOLOv3 may be more suitable for environments requiring high accuracy and reliability, InsightFace can be effectively used in resource-constrained settings. For comprehensive classroom monitoring, a hybrid approach that leverages the strengths of both models could be considered, depending on the specific requirements of the environment and available hardware.

In conclusion, this study highlights the importance of selecting an appropriate face detection model based on the specific needs of the deployment environment. YOLOv3's ability to deliver high accuracy and confidence across diverse conditions makes it a strong candidate for real-time monitoring in educational settings, while InsightFace offers a practical alternative for scenarios where computational efficiency is critical. Future work could explore the integration of these models with advanced post-processing techniques to further enhance detection accuracy and reduce false positives in real-world applications.

Acknowledgements

We would like to express our gratitude for the support provided by Universitas Malikussaleh PNBP funds, which contributed to the successful completion of this researc

Author contributions

Muhammad Fikry: Conceptualization, Methodology, Validation. **Taufiqurrahman**: Software, Methodology, Writing—review and editing. **Fadlisyah**: Validation, Formal analysis. **Munirul Ula**: Writing—original draft preparation, Resources. **Nurdin**: Investigation, Field study. **Muhammad Yani**: Visualization. **Aldy Anugrah Pohan**: Data curation

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] T. Luan and M. A. E. Damian, "A Study on the Application of Deep Learning-based Media Tampering Detection Technology in Higher Education Teaching Resource Protection," Contemporary Education and Teaching Research, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:259718600

[2] L. Ni, J. Shi, B. Han, N. Zhang, Q. Lan, and Z. Su, "Classroom Roll Call System Based on Face Detection Technology," 2022 10th International Conference on Information and Education Technology (ICIET), pp. 42–46, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:249101012

[3] M. A. M. Ali, T. Aly, A. T. Raslan, M. Gheith, and E. A. Amin, "Advancing Crowd Object Detection: A Review of YOLO, CNN and ViTs Hybrid Approach," Journal of Intelligent Learning Systems and Applications, vol. 16, no. 3, pp. 175–221, 2024.

[4] X. Zhang, X. Dong, Q. Wei, and K. Zhou, "Real-time

object detection algorithm based on improved YOLOv3," J Electron Imaging, vol. 28, no. 5, p. 53022, 2019.

[5] S. Liu, Y. Xu, L. Guo, M. Shao, G. Yue, and D. An, "Multi-scale personnel deep feature detection algorithm based on Extended-YOLOv3," Journal of Intelligent & Fuzzy Systems, vol. 40, no. 1, pp. 773–786, 2021.

[6] N. Głowacka and J. Rumiński, "Face with Mask Detection in Thermal Images Using Deep Neural Networks," Sensors, vol. 21, no. 19, 2021, doi: 10.3390/s21196387.

[7] L. Tan, T. Huangfu, L. Wu, and W. Chen, "Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identification," BMC Med Inform Decis Mak, vol. 21, no. 1, pp. 1–11, 2021, doi: 10.1186/s12911-021-01691-8.

[8] R. Shamim and Y. Farhaoui, "An In-depth Comparative Study: YOLOv3 vs. Faster R-CNN for Object Detection in Computer Vision," in Artificial Intelligence, Big Data, IOT and Block Chain in Healthcare: From Concepts to Applications, Y. Farhaoui, Ed., Cham: Springer Nature Switzerland, 2024, pp. 266–277.

[9] K. Wang, M. Liu, and Z. Ye, "An advanced YOLOv3 method for small-scale road object detection," Appl Soft Comput, vol. 112, p. 107846, 2021, doi: https://doi.org/10.1016/j.asoc.2021.107846.

[10] D. Wanyonyi and T. Celik, "Open-source face recognition frameworks: A review of the landscape," IEEE Access, vol. 10, pp. 50601–50623, 2022.

[11] N. Sadman, K. A. Hasan, E. Rashno, F. Alaca, Y. Tian, and F. Zulkernine, "Vulnerability of Open-Source Face Recognition Systems to Blackbox Attacks: A Case Study with InsightFace," in 2023 IEEE Symposium Series on Computational Intelligence (SSCI), 2023, pp. 1164–1169.

[12] O. Yakovleva, A. Kovtunenko, V. Liubchenko, V. Honcharenko, and O. Kobylin, "Face Detection for Video Surveillance-based Security System," in CEUR Workshop Proceedings, 2023, pp. 69–86.

[13] J. Guo, J. Deng, and A. Lattas, "Sample and Computation Redistribution for Efficient Face Detection".

[14] K. Gkrispanis, N. Gkalelis, and V. Mezaris, "Filter-Pruning of Lightweight Face Detectors Using a Geometric Median Criterion," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 280–289.

[15] Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo, and L. Ali, "Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition," Big Data and Cognitive Computing, vol. 7, no. 1, 2023, doi: 10.3390/bdcc7010048.

[16] J.-H. Won, D. Lee, K.-M. Lee, and C.-H. Lin, "An Improved YOLOv3-based Neural Network for De-identification Technology," 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), pp. 1–2, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:199540948

[17] S. Mukherjee, T. Sharma, A. Singh, S. S. Gayathri, and S. Dhanalakshmi, "Multi-Pedestrian Detection using Hybrid ML Algorithms for Autonomous Vehicles," 2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET), pp. 1–4, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:267575819

[18] M. Widiasri, A. Z. Arifin, N. Suciati, E. R. Astuti, and R. Indraswari, "Alveolar Bone Detection from Dental Cone Beam Computed Tomography using YOLOv3-tiny," 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), pp. 1–6, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:236191711