Lighter Student Engagement Recognition in a Classroom Environment Using Skeletal Keypoints

Gabriel Asael Tarigan, Gregorius Natanael Elwirehardja, Kuncahyo Setyo Nugroho, Bens Pardamean

***Abstract*—Student retention is crucial for educational institutions, influencing reputation, finances, and ranking metrics. Engagement, reflecting a student's connection, interest, and effort, plays a vital role in learning, fostering critical thinking, and supporting retention. Recent advancements use students' poses to predict engagement, providing valuable insights without disrupting the teaching-learning dynamic. The prevailing research trend leans toward employing multi-modal approaches, such as a combination of pose detection with object detection. However, current methods use out-of-date object detection methods and manual dataset creation, which is cumbersome, requiring thousands of manually annotated data. This problem is then addressed by proposing a novel method of student engagement detection, using a combination of You Only Look Once Version 8 Mini (YOLOv8m) and MediaPipe as state-of-the-art alternatives to improve both object detection and human pose estimation. The (YOLOv8m + MediaPipe) method surpasses the baseline (YOLOv4 + OpenPose) with higher accuracy (0.70 vs. 0.41) and lower cross-entropy loss (0.40 vs. 0.60) on the test set, confirmed by a statistically significant paired t-test. It also exhibits a remarkable speed advantage, around 16 times faster than the baseline in pose detection data collection rates. Despite not being designed for it, the proposed method achieves multiple keypoint detection, matching the baseline's amount.**

***Index Terms*—computer vision, object detection, pose detection, student engagement**

I. Introduction

THE ever-expanding application of artificial intelligence (AI) impacts people's daily lives. This includes the field of education, where AI is being used to develop methods to aid learning. Learning is a critical measure of the world's civilization and evolution, with enormous implications for individuals and societies [1], [2]. Furthermore, university or school-related variables like ranking, revenue, and reputation are linked to student retention. Therefore, learning and engagement in such activities should be monitored and evaluated, and as such, machine learning applications in the field of education are necessary [3].

The level of engagement that students demonstrate with the information they are taught is an important aspect of learning that must be evaluated. Their conscious or subconscious behaviors will be imitated as posture states, also referred to as poses, and knowing this can assist teachers in providing valuable feedback, enhancing their instruction, and gauging the cognitive load of their students—especially during complicated learning [4], [5]. Applying AI techniques such as machine learning (ML) techniques can provide significant insights and highlight important trends in students' learning practices [6].

Non-verbal clues obtained from the video picture frames of the classroom data may be used to identify the non-intrusive students' participation [7] efficiently. Computer vision is particularly good at posture estimation to determine human behavior [2], [8]. However, datasets that contain engagement levels in a teaching-learning setting are extraordinarily rare, especially when the learning is done onsite in a classroom setting. Therefore, a primary dataset was used for this research, and the process of collecting and using the primary dataset combined with the object detection mechanism will also be explained.

To ensure learners understand the learning materials, the teacher-learner interaction must be invested in making appropriate educational decisions that can continue and even intervene when learners are not engaging. As discussed, the current trend mostly uses varied methods to detect non-intrusive student engagement levels. However, two main weaknesses in such methods arise. The first major issue is current methods that use out-of-date object detection methods, such as YOLOv3 and YOLOv5, which have been beaten in accuracy and performance time with current state-of-the-art object detection methods such as YOLOv8 [9], [10], [11], [12], [13]. Data collection presents a second challenge, as the process often requires substantial time and effort to gather, organize, and validate information, especially for large or complex datasets. This intensive demand for resources slows down the workflow and limits the method's responsiveness to changes, as adjustments may require additional data gathering and verification rounds. Consequently, the method's flexibility and adaptability become constrained, making it more difficult to promptly respond to evolving requirements or integrate new insights.

Manuscript received February 17, 2024; revised April 7, 2025.

Gabriel Asael Tarigan is a postgraduate student of the Computer Science Department, Binus Graduate Program – Master of Computer Science Program, Bina Nusantara University, Jakarta 11480, Indonesia (corresponding author, email: gabriel.tarigan@binus.ac.id).

Gregorius Natanael Elwirehardja is a lecturer of the Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 1140, Indonesia (email: gregorius.e@binus.edu).

Kuncahyo Setyo Nugroho is a researcher at Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia (email: kuncahyo.nugroho@binus.edu).

Bens Pardamean is a Professor of Computer Science, BINUS Graduate Program – Master of Computer Science Program, Bina Nusantara University, Jakarta 11480, Indonesia (email: bpardamean@binus.edu).

In this study, a novel pipeline utilizes a combination of You Only Look Once Version 8 Mini (YOLOv8m) and MediaPipe to solve the weaknesses of previous methods, particularly in the computational cost and data collection speed. This pipeline will now be referred to as the proposed method. YOLOv8m will be used as the object detector mechanism to detect the students in the classroom, as it is a version of the most advanced real-time object detectors available today [14]. MediaPipe will detect key points from each student's body and infer the action in poses as it can only detect one body. The object detector is used to detect multiple bodies. To compare the effectiveness and improvements of the proposed method, a similar method by [9] using YOLOv4 and OpenPose is chosen as the baseline. They used macro-average precision, macro-average recall, and accuracy as their evaluation metrics. Therefore, these evaluation metrics will also be used in this study, and the metrics results will be compared with the proposed method. A paired t-test is performed on the cross-entropy loss of the test set to assess the statistical significance of the results compared to the baseline. Cross-entropy loss is chosen for the paired t-test because it reflects better how accurately the method classifies specific actions when compared to the ground truth [15], [16], [17].

II. Related Works

Several studies have identified student behavior in a classroom learning setting by detecting students' poses. A study suggested in research utilizing person detection and skeleton position estimation will be used as the baseline [9]. Their method uses a deep neural network to classify behaviors and outperforms previous skeleton-based approaches. While OpenPose detection and manual annotation were used in another study, body pose classification used mathematical formulas derived from joint angles. Using this technique, they could identify individual and group student behavior and assess how it affected classroom participation [2].

Other methods use a multi-object pose estimation model that incorporates spatiotemporal semantics for various sizes and poses of video multi-objects. It uses temporal clues between video frames to improve the location of important human body parts. It creates modular parts to enhance the pose data, improving the pose estimation process using YOLOv3 and Lite-HRNet to infer poses [10]. A similar technique for identifying and locating student behaviors from CCTV images in computer labs of a smart campus was proposed in the study. The system uses YOLOv3 for object detection and deep neural network-based methods for recognizing human activity [11].

Furthermore, a technique utilizing the upgraded Faster R-CNN model was shown for recognizing student postures. Due to the low-resolution imaging settings and students' concentrated attention in the classroom, tiny, low-quality targets are difficult to see. The suggested technique uses locality-preserving loss functions to enhance the classifier's performance using low-level convolution features, which are frequent in high-resolution data [12].

Several studies identify the students' stances using transfer learning techniques like VGG-19. Images from high-density recordings taken in the classroom using fixed-angle cameras were gathered into a dataset. Nine hundred forty-two files total from eight classes—interest (engagement) and non-interest—are included in the collection (disengagement) [18]. A single-stage object identification technique was also investigated to address pose detection issues such as object size fluctuations, imbalanced categories, and similarity between categories. To enhance the detection performance, the method incorporates an adaptive fusion mechanism and a multi-scale feature detection branch [5]. Another similar method uses a deep learning methodology based on spatiotemporal representation learning to identify abnormal behavior in the classrooms of college students. The study finds that the algorithm performs 5% better than the benchmark three-dimensional CNN, making it an invaluable tool for ensuring efficient classroom education. (C3D) [19].

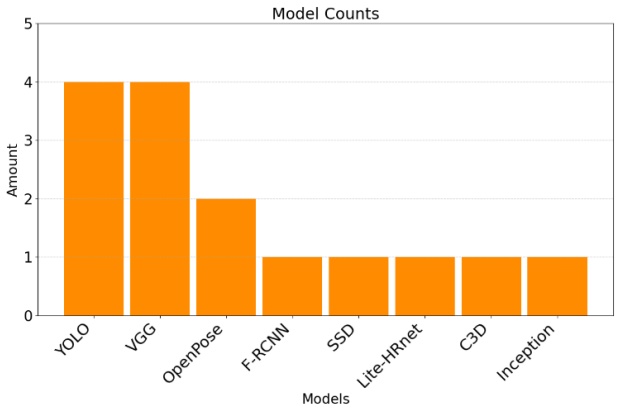


Fig. 1. Utilized methods found in previous studies

The strategy for covertly analyzing student participation in a classroom setting utilizing non-verbal signs, including body language, hand gestures, and facial expressions, is suggested in this study. The proposed technique classifies student involvement levels with 71% accuracy and uses convolutional neural network architecture [20]. The Squeeze-and-Excitation Networks (SENet) attention detection mechanism recommends a YOLOv5s network structure based on the YOLO algorithm to recognize and evaluate students' classroom behavior, and such a method makes scenarios with complicated backgrounds have more accurate predictions [13]. For assessing the emotional states of students in a classroom setting, it is suggested that a unique hybrid convolutional neural network (CNN) design. The proposed architecture consists of two models: CNN-1, which examines a single student's emotional states in a single picture frame, and CNN-2, which employs a few students in a single image frame to forecast the overall affective state of the class [21]. Studies reviewed showed multiple students' pose prediction methods using base CNN architecture. However, it is not included in Fig. 1 as most base CNN methods are combined with other methods.

III. Research Methodology

In most past research, the methods required hundreds to thousands of manually annotated data to be accurate, and if the data were small, then the accuracy would be terrible [2]. These weaknesses are then addressed by proposing a novel method of student engagement detection, using a combination of YOLOv8m and MediaPipe as state-of-the-art alternatives.

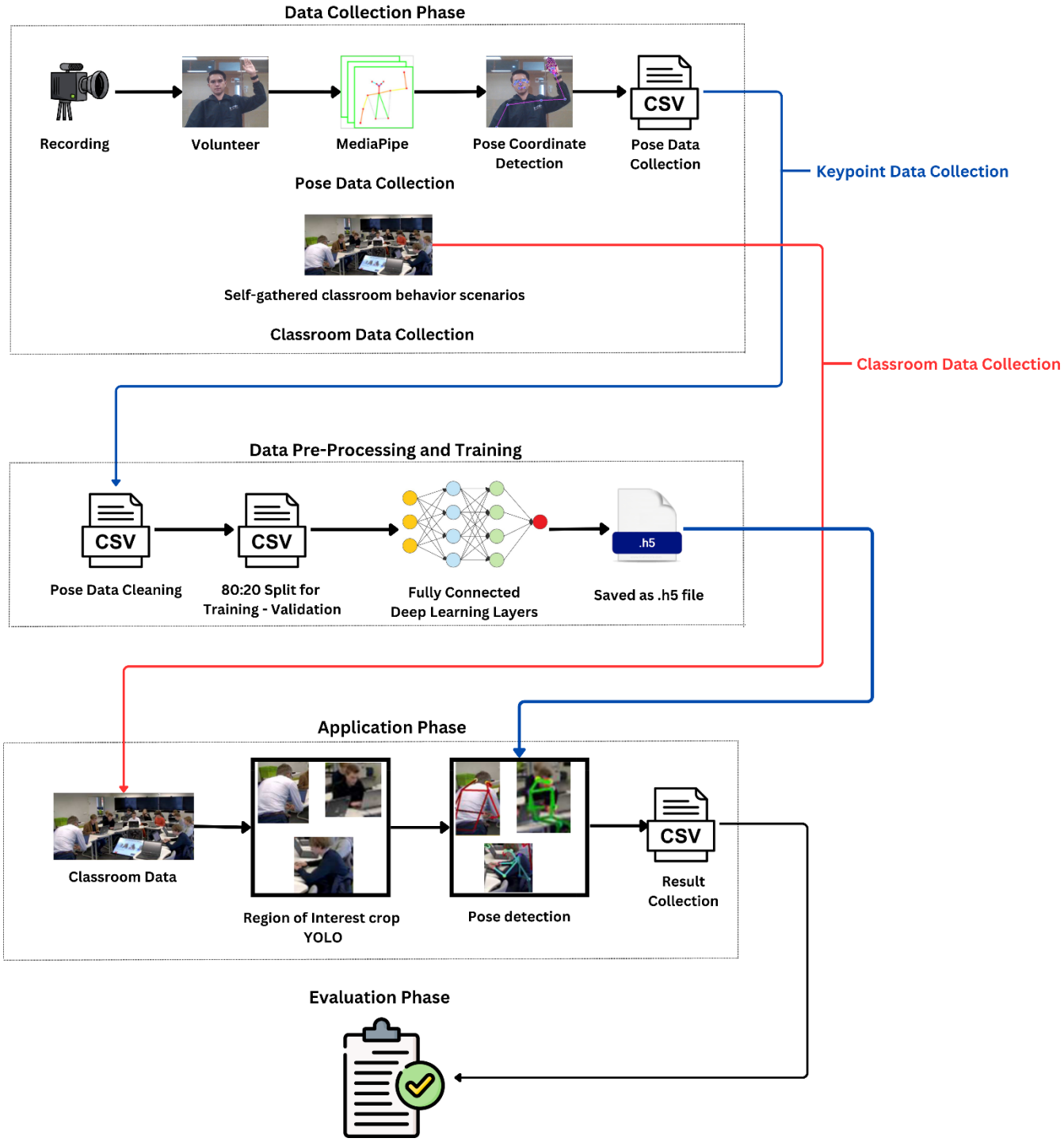


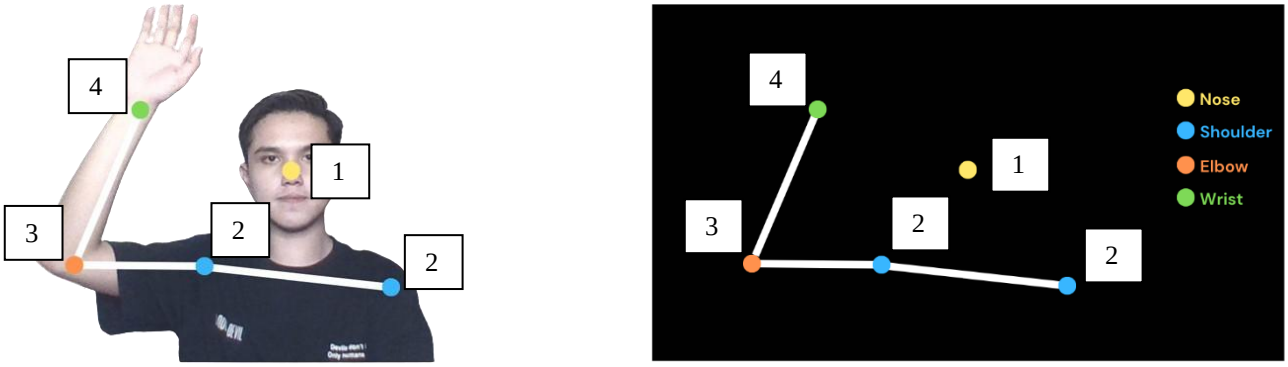
Fig. 2. Conceptual framework

This methodology allows for more efficient data collection, reduces the reliance on manual annotation, and improves scalability. The overall pipeline is depicted in Fig. 2 to better illustrate the research workflow.

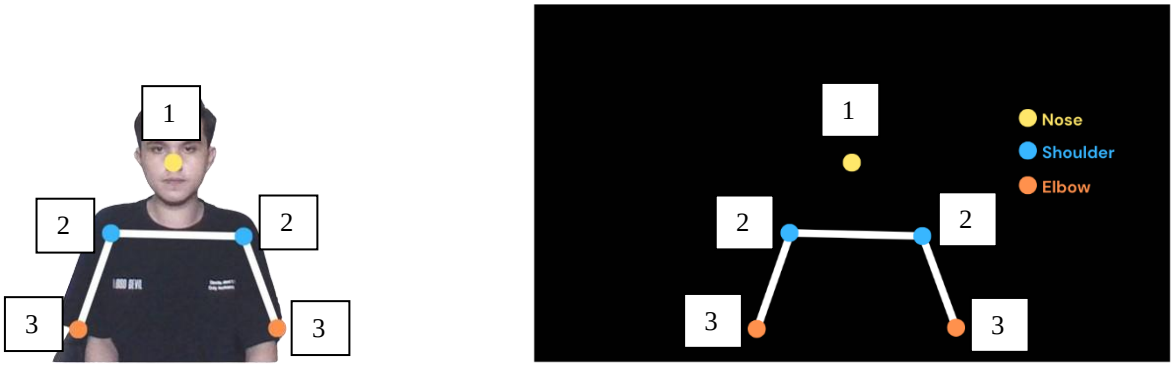
A. Data collection

The initial data collection phase involves capturing poses, which will be recorded in video format. These poses will be captured using a laptop webcam, with keypoint coordinates and pose classifications determined using MediaPipe. The poses to be classified include "on_the_phone," "raise_hand," "bored," "sleeping," and "engaged," as illustrated in Fig. 3.

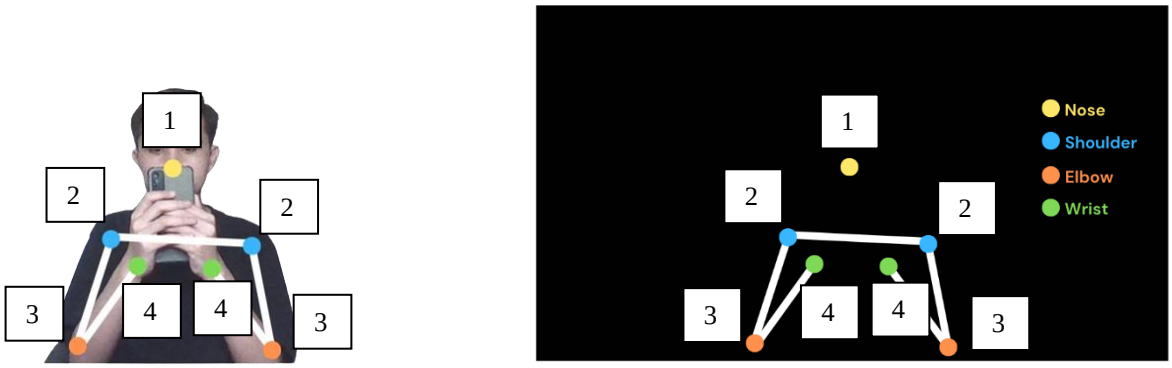
This process involves filtering to identify keypoints that significantly impact detection accuracy. The figure illustrates the poses with color-coded keypoints numbered as follows: “1” for nose, “2” for shoulder, “3” for elbow, and “4” for wrist. For instance, in Fig. 3a, five keypoints are captured: the nose, shoulders, the elbow, and the wrist. This result is captured in coordinates, with an example in Table I. Each pose detection includes 14 features, comprising seven keypoints' x-axis- and y-axis coordinates. Dataset is available at: https://github.com/Zayphen/pose_coordinate.



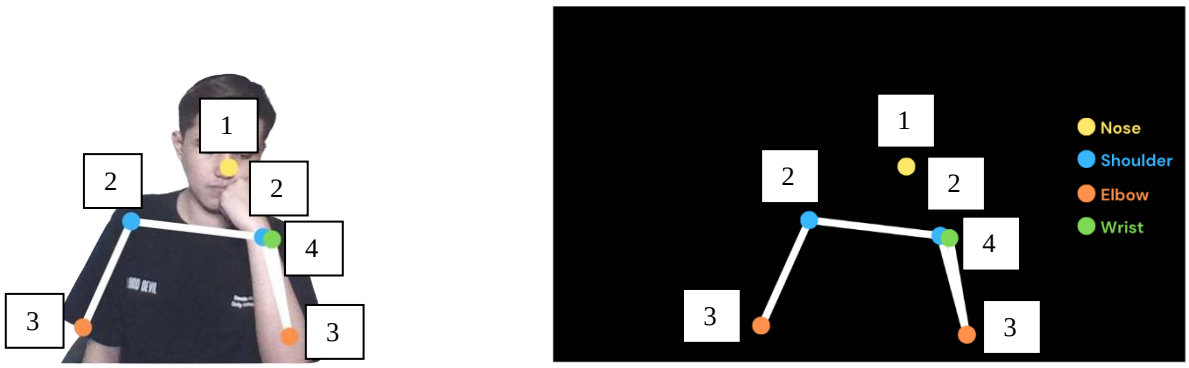
(a) Raise hand



(b) Engaged



(c) On the phone



(d) Bored

Fig. 3. Samples of captured pose keypoints

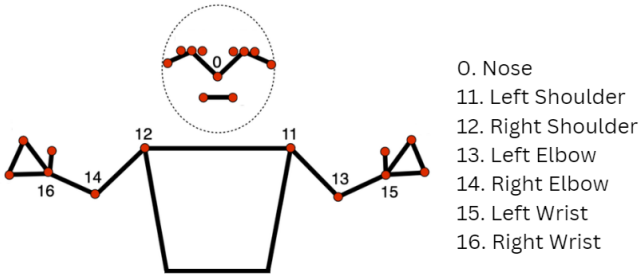


Fig. 4. Keypoints used for pose detection

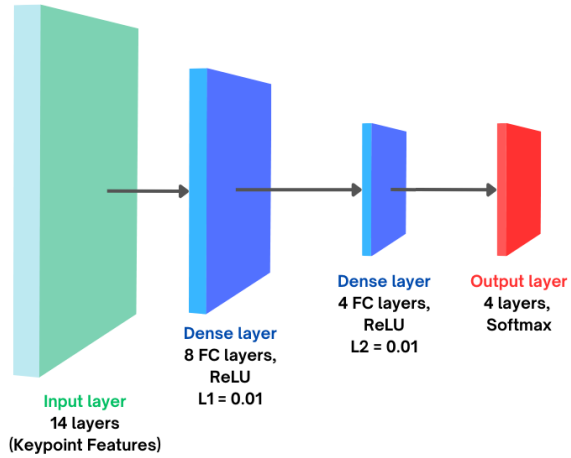


Fig. 5. Fully-connected network classifier model

TABLE I
POSE DATA REPRESENTATION

class	x11	y11	x12	…	y16
on_phone	0.4687	0.478	-0.688	…	0.999
on_phone	0.471	0.467	-0.715	…	0.999
on_phone	0.469	0.466	-0.748	…	0.999
on_phone	0.458	0.466	-0.800	…	0.999
…	…	…	…	…	…

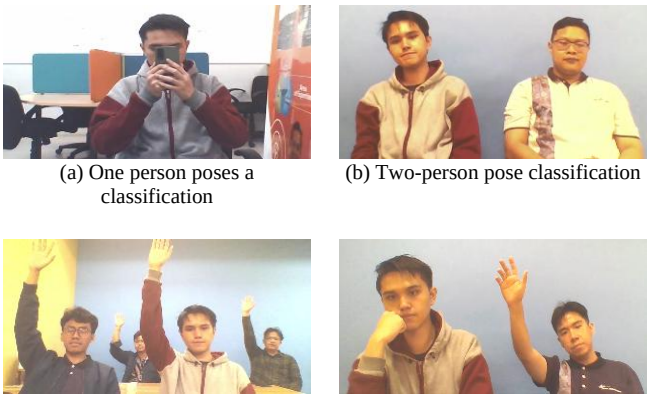
The data collection process comprises six sessions, each gathering an average of 500 frames. The "raise_hand" and "bored" poses were collected twice, accounting for variations in right—and left-hand dominant movements. Fig. 4 provides examples of each pose and its key points, with each key point color-coded for clarity.

B. Pre-processing

Data pre-processing is done to ensure the optimal performance of the created dataset. Afterward, shuffling is done on the dataset, given that dataset shuffles are known to enhance the learning's statistical performance [24]. The dataset was split to train and validation split, as it had a more substantial influence, notably on modeling at larger dataset sizes, and the overall performance of the models rose with the size of the dataset [25]. Therefore, the obtained data was divided into training and validation sets in an 80:20 ratio, with 80% used for training and 20% for validation.

C. Classifier Model

The pre-processed data was trained, the purpose of which was to classify the poses from the coordinates collected. The model layers can be seen in Fig. 5. The model uses a simple, fully connected layer with the input layer having the same amount of number as the feature, which is 14, and then put



(a) One person poses a classification

(b) Two-person pose classification

(c) Four-person pose classification

(d) Two-person pose classification with different poses per person

Fig. 6. Scenarios of test set

To eight and four dense layers, respectively, until the output layer of four dense layers, as there are four classes, is present.

D. Inference

To better see the pose detection driven by the proposed method, it encapsulates the diverse conditions and situations curated and examined as part of the experimental framework. YOLOv8m is used for the first stage of inference to detect students sitting in a classroom setting and separate each student into bounding boxes. Afterward, MediaPipe inferred each pose of the detected students and collected them for evaluation. There are 3049 data for the test set. The various pose scenarios used to test the model are shown in Fig 6. The first pose scenario is single-person pose detection, followed by two-person pose detection, and afterward, four-person detection to see the proposed method's capability to determine the poses when depth is involved. Apart from that, a scenario of two people with different poses was also carried out to test the robustness of the model. The scenario conditions were similar in terms of the classroom environment. All the test scenarios used were in 480p.

E. Evaluation Schema

The baseline method based on [9] research that uses the combination of OpenPose and YOLOv4 and the proposed method were compared on the test set to produce comparable evaluation metrics on macro-average precision (Equation 1), macro-average recall (Equation 2), accuracy (Equation 3), softmax probabilities (Equation 4), and cross-entropy loss (Equation 5). Finally, the paired t-test was performed to establish the result's significance and to confirm that the result was not acquired by chance. The paired t-test is a widely used statistical hypothesis test in pain studies, evaluating the probability of a difference between two groups without relying on an absolute standard; therefore, it will be used as the inferential statistic metric to see the magnitude of significance [16]. The paired t-test formula is represented by Equation 6.

$$Macro - Average\ Precision = \frac{\sum_{k=1}^{K} Precision_k}{K} \quad (1)$$

$$Macro - Average\ Recall = \frac{\sum_{k=1}^{K} Recall_k}{K} \quad (2)$$

$$Multiclass\ Accuracy(z_i, y_i) = \frac{1}{n}\sum_{1}^{n}[[z_i == y_i]] \quad (3)$$

$$Softmax(p_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \quad (4)$$

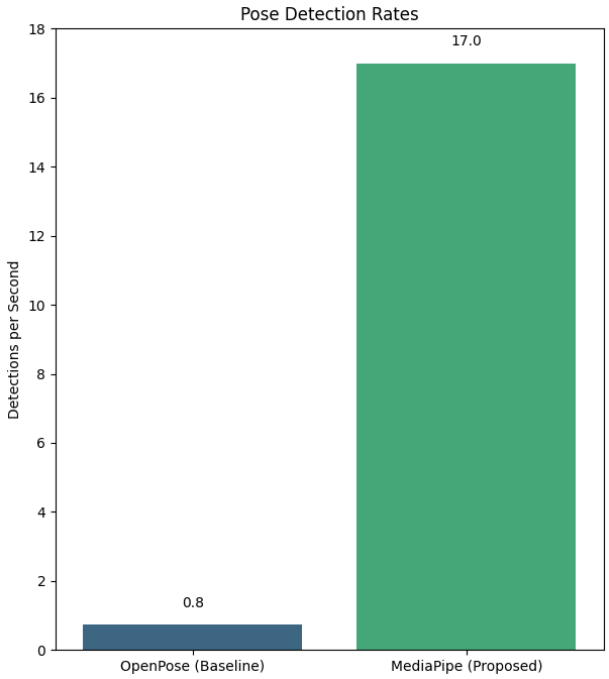


Fig. 7. Comparison of total data collection rate per second

$$H(p,q) = \sum_i p_i \log(q_i) \quad (5)$$

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \quad (6)$$

To visualize the results gathered better, illustrations of the misclassification rate of the models by confusion matrix are also used. The y-axis shows the actual label (ground truth), and the x-axis shows the predicted label. Therefore, only the diagonal line of the matrix indicates its correct prediction or true positive (TP).

IV. Result and Discussion

A. Pose Extraction Time

The proposed method extracts pose much faster than the baseline, lasting about 2 minutes and 30 seconds every session. The proposed approach took 15 minutes in total. In contrast, the baseline method took 255 minutes (4 hours and 15 minutes). Therefore, the proposed method's data gathering is 17 times faster than the baseline, as shown in Fig. 7. Efficiency in data gathering is crucial for assessing model robustness across multiple settings or populations, underlining the need for wide-ranging data collecting [26]. Moreover, model performances are increased with the database size scaling, and the power law in the context of more data equates to better results [27].

B. Exploratory Data Analysis

A total of 7351 keypoints data were collected for training, and 3049 frames of video data were collected for testing. In Fig. 8, a representation of the data distribution is shown; the collected data consists of 1350 "engaged" class keypoints, which represent 18.36 % of the data, 2700 "bored," which represents 36.72 % of the data, 2700 "raise hand," which represents 36.72 % of the data, and 600 "on_the_phone," which represents 8.16% of the data. The color coding reveals the coordinates of each keypoint according to the

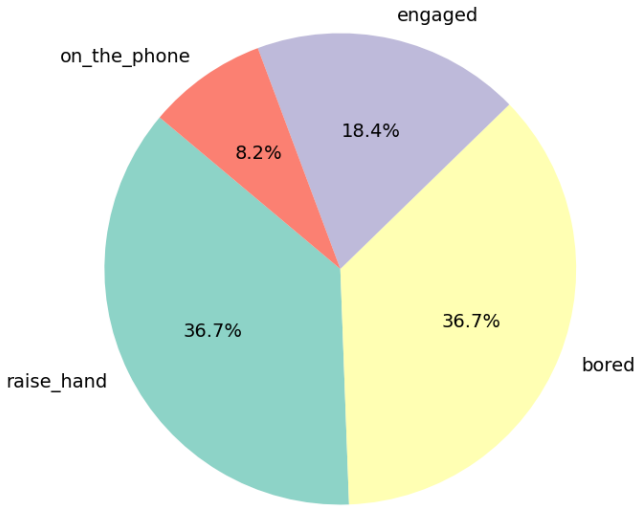


Fig. 8. Distribution of classes in the dataset

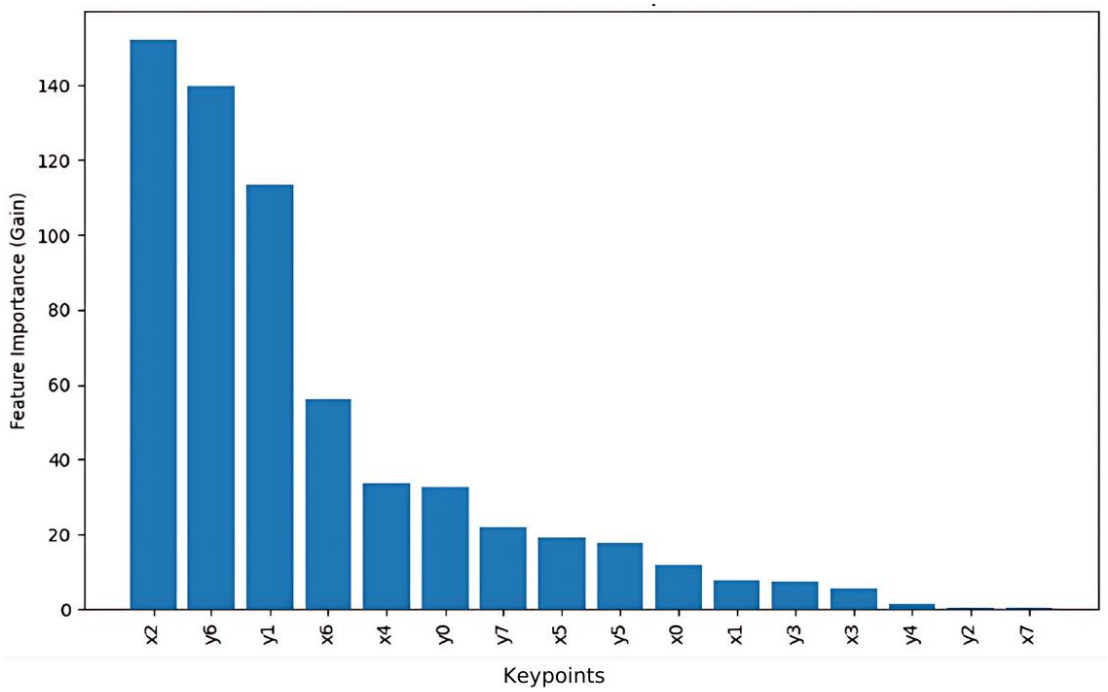
classes. It is also clear that "raise hand" and "bored" have substantially more keypoints collected, accounting for 72% of the information because both classes are differentiated by the right or left hand.

Focusing on hands and upper body joints is critical for observing student behavior in the classroom. Hand movement changes significance as a feature increases the model's categorization abilities using XGBoost's gain-based feature significance score [28]. As shown in Fig. 9, the baseline's (Fig. 9a) most important keypoints are $x2$, $y6$, $y1$, $x6$, $x4$ and are related to the general arm area, with a decline from the neck ($y1$) to the left elbow ($x6$). The proposed method (Fig. 9b) identifies. $x13$, $x12$, $x14$, $y13$, and $y15$ The most critical key points correspond to the general arm area.

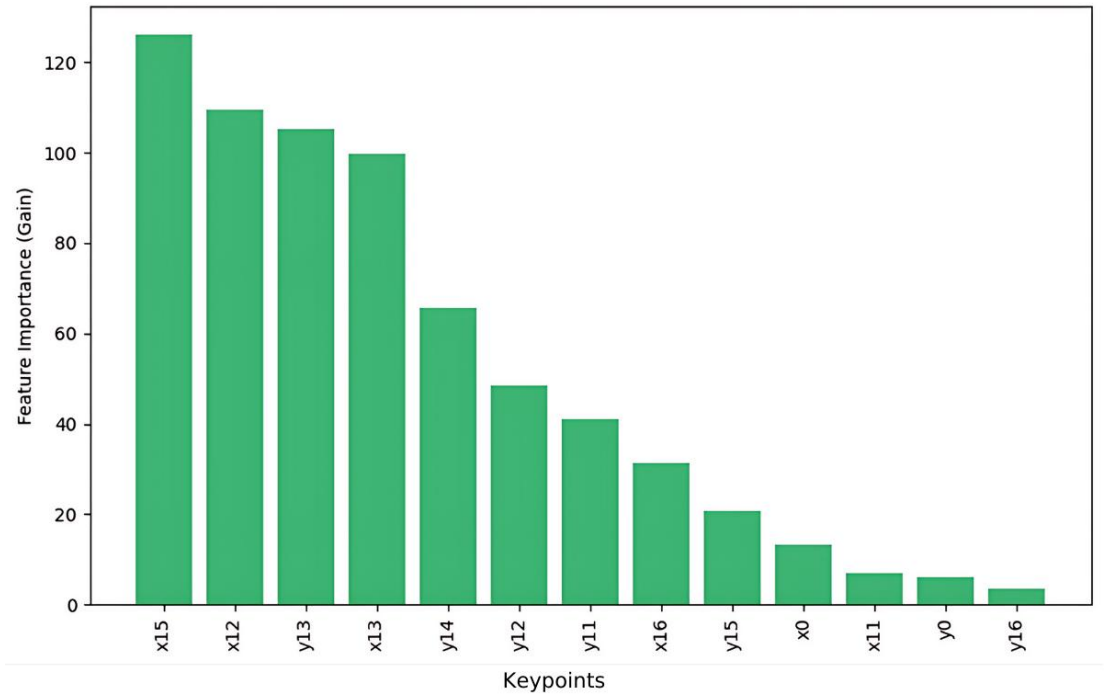
To conclude, the proposed method significantly improves the efficiency of pose extraction, reducing the session time to 2 minutes and 30 seconds per session, totaling 15 minutes overall. In contrast, the baseline method required 255 minutes (4 hours and 15 minutes), making the proposed method 17 times faster, as demonstrated in Fig. 7. This enhanced efficiency is vital for evaluating model robustness across diverse populations and settings, supporting comprehensive data collection. Additionally, model performance improves with larger datasets, following the power-law relationship where more data yields better results.

7,351 key points were gathered for training, with 3,048 frames used for testing. The dataset distribution, shown in Fig. 8, includes 1,350 key points for the "engaged" class (18.36%), 2,700 for both "bored" (36.72%) and "raise hand" (36.72%), and 600 for "on_the_phone" (8.16%). The majority of key points belong to the "bored" and "raise hand" classes (72%), which are distinguished by hand movements.

To conclude, hand movements (shoulders to palm area) enhance the model's classification ability using XGBoost's gain-based feature significance score. Fig. 9 highlights the most important keypoints for the baseline (Fig. 9a) and proposed (Fig. 9b) methods. While both models prioritize keypoints in the general arm area, the baseline shows a decline from the neck to the elbow, while the proposed method identifies keypoints in the lower arm and hand regions as more critical.



(a) Baseline method



(b) Proposed method

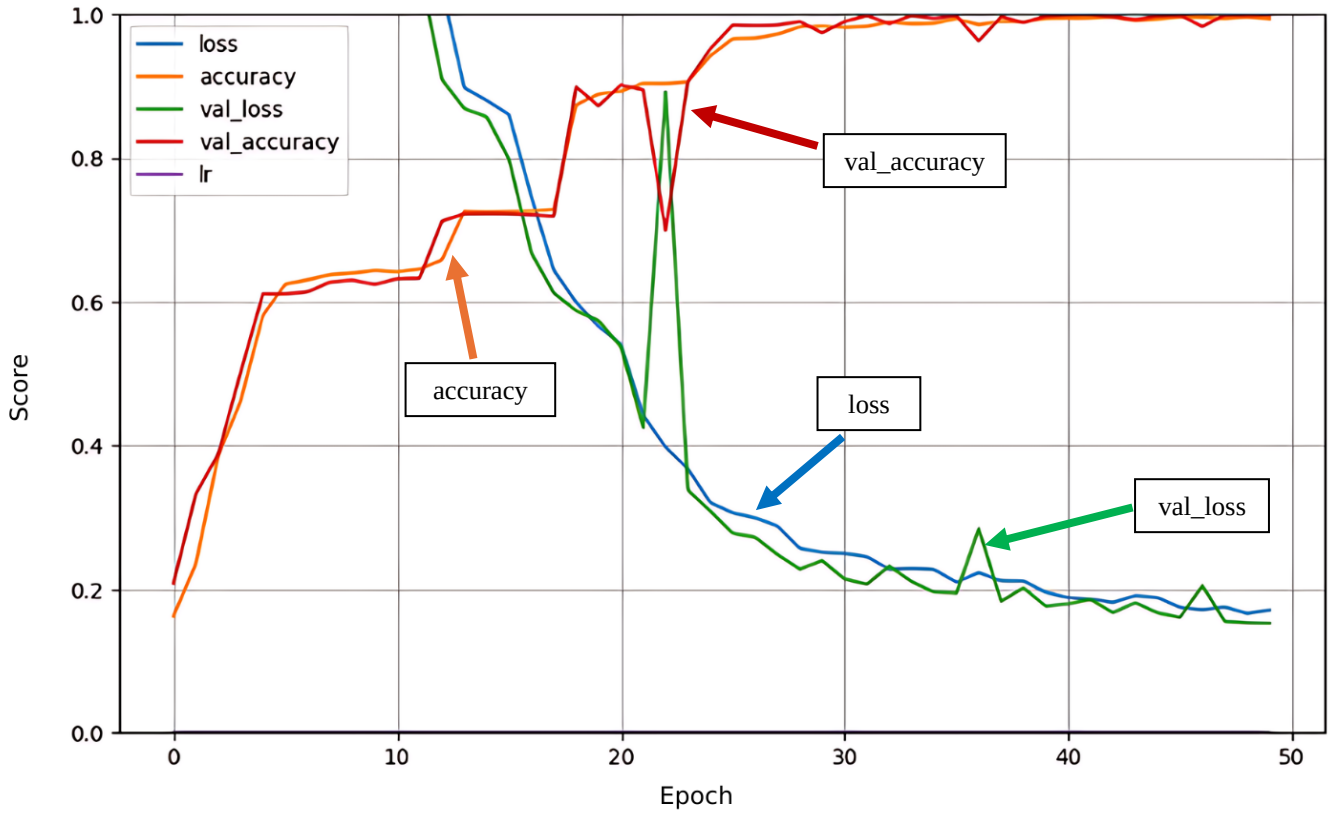
Fig. 9. Feature importance of each method

Both methods recognize elbow and wrist keypoints as essential but differ in the neck as they are absent from the proposed method. Both methods also exhibit satisfactory training validation results, indicating effective model generalization. However, a good fit is indicated by stable points in both training and validation losses, minimizing the generalization gap [29], where smoother learning curves are generally preferred for better generalization [30]. The proposed method displays a smoother curve compared to the baseline, as shown in Fig. 10, which helps to identify overfitting, where training loss decreases continuously while validation loss starts to increase. The result of this process is shown in Table II.

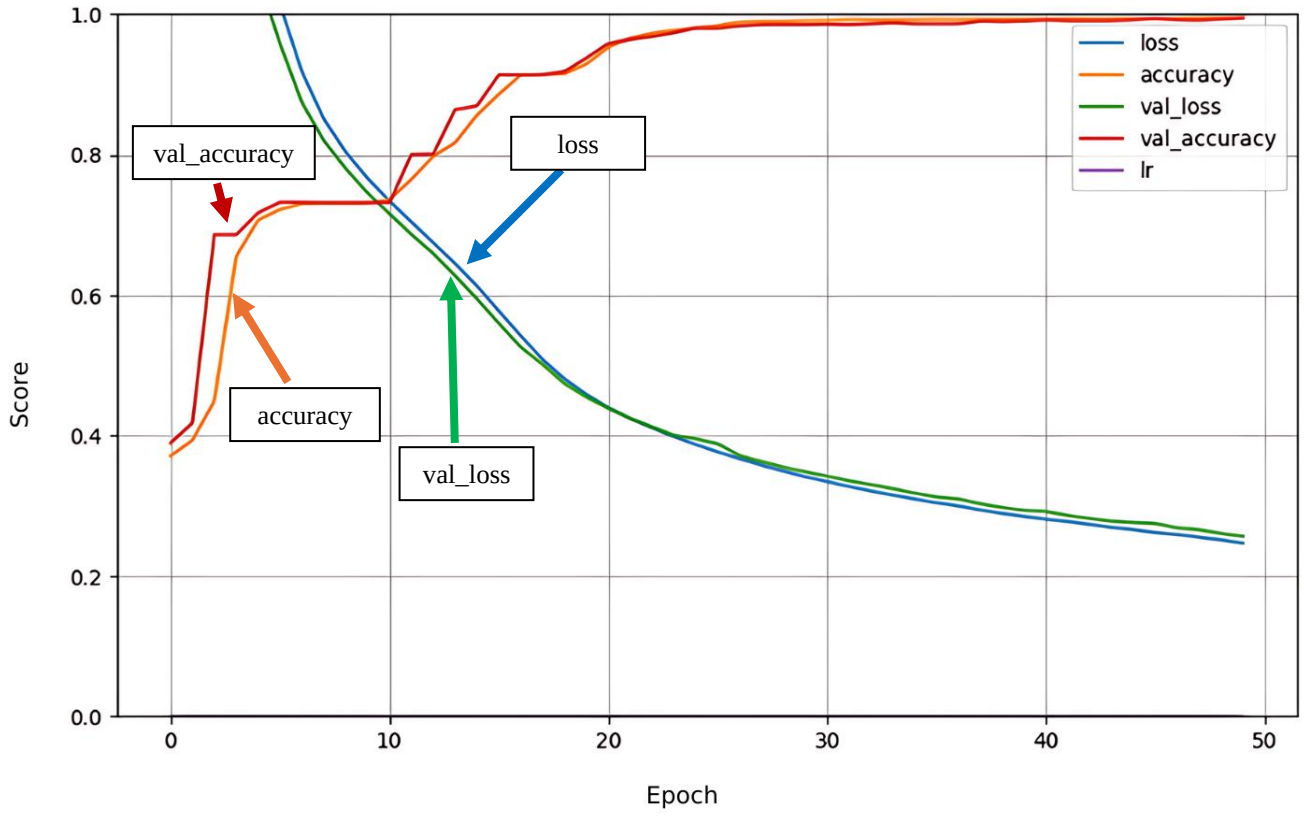
TABLE II
TRAINING VALIDATION RESULT

Metric		Baseline	Proposed
Training	Accuracy	0.9942	0.9951
	Loss	0.1705	0.2469
Validation	Accuracy	0.9993	0.9946
	Loss	0.1525	0.2566

The proposed method's smoother learning curve (Fig. 10b), compared to the baseline method (Fig. 10a), suggests it may achieve better generalization performance on unseen data. A smoother curve indicates that the model is effectively learning the underlying patterns of the data without overfitting.



(a) Baseline method



(b) Proposed method

Fig. 10. Feature importance of each method

s

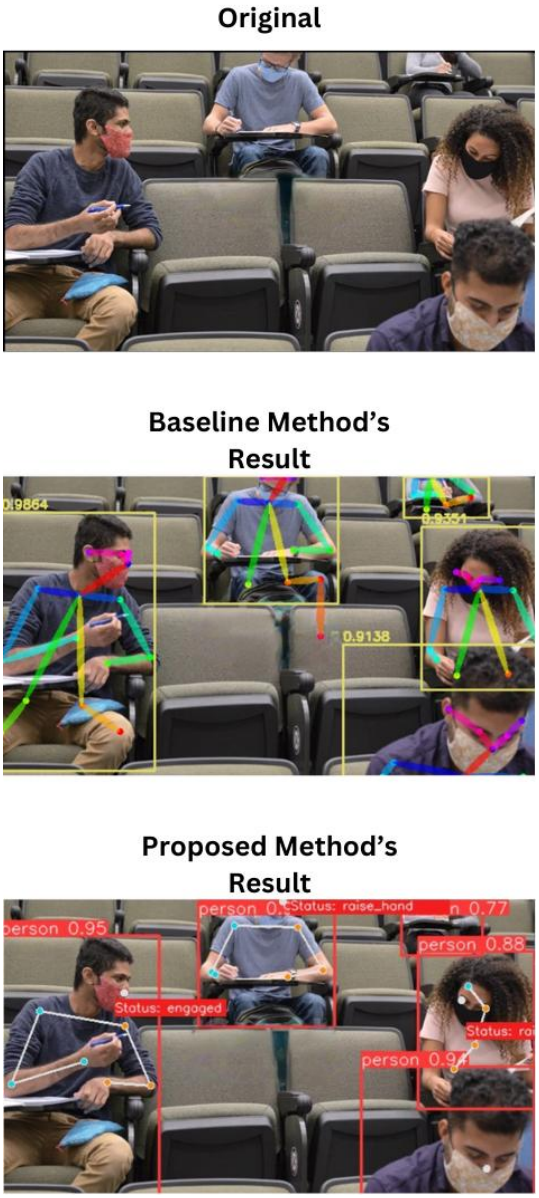


Fig. 11. Illustration of the expected result

C. Scenarios (Test Sets) Result

This sub-chapter details the results achieved by integrating object detection with pose estimation techniques of the baseline method, YOLOv4, and OpenPose, with the proposed YOLOv8 and MediaPipe based on given scenarios, by highlighting key differences in performance, accuracy, and clarity between the two approaches. The comparative analysis emphasizes the proposed method's faster inference and superior detection capabilities, demonstrating its potential advantages in real-world applications and improving multi-modal human pose analysis. To better visualize the expected outcomes from the methods, an illustrative example of the proposed method's results is presented in Fig. 11. This figure compares the outputs inferred by the baseline method above the illustration inferred by the proposed method, which will also be the case for the subsequent comparisons.

In addition to the visual comparison of inferences between the proposed and baseline methods, a detailed analysis uses quantitative

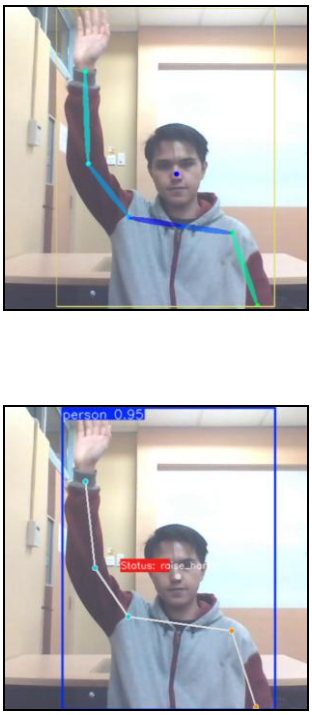


Fig. 12. One-person pose recognition

metrics such as correct prediction rates, cross-entropy loss, macro-average precision, recall, and accuracy. A spider chart will highlight each model's strengths and weaknesses, with axes representing each class. The further the performance extends along an axis, the stronger the model for that class, offering a clear and intuitive comparison. To complement this, confusion matrices will also be color-coded: the baseline method in blue and the proposed method in green, with lighter shades indicating fewer predictions and darker shades indicating higher predictions, making it easy to distinguish the performance differences between the two methods.

1) One-person Pose Recognition

The first illustrated inference comparison focuses on one-person detection. The data for this scenario consists of 501 frames, distributed across four distinct poses: 78 frames for "raise_hand," 121 for "engaged," 241 for "bored," and 61 for "on_the_phone." This distribution reflects the likelihood of these actions occurring in real-life situations, allowing for a comprehensive evaluation of the model's performance in a realistic context. Fig. 12 provides an example of the "on_the_phone" pose demonstrated by a volunteer.

In this scenario, the baseline method shows misclassification, particularly for the "on_the_phone" class, which is often confused with "engaged" and "bored." The baseline method failed to correctly predict any "on_the_phone" instances, misclassifying all 62 frames as "bored." The "raise_hand" class was sometimes misclassified as "bored." The confusion matrix highlights these issues, with the baseline achieving an overall accuracy of 0.80, while the macro-average precision, recall, and F1-score were 0.66, 0.63, and 0.62, respectively.

In comparison, the proposed method performed better, though the "on_the_phone" class still faced some misclassification as "engaged" and "bored." Despite this, the proposed method achieved more correct predictions overall, with occasional misclassifying of "raise_hand" as "bored." The proposed method's overall accuracy was 0.90,

TABLE III
ONE-PERSON RECOGNITION RESULT

Evaluation Metric	Baseline	Proposed
Total data	501	
Cross-entropy Loss	0.23	0.28
Macro-Average Precision	0.66	0.87
Macro-Average Recall	0.63	0.93
Accuracy	0.80	0.90

And its macro-average precision, recall, and F1-scores were 0.87, 0.93, and 0.89, respectively.

Table III presents a detailed comparison between the two methods. While the baseline method showed a lower cross-entropy loss (0.23 vs. 0.28), the proposed method outperformed other metrics. Macro-average precision was higher for the proposed method (0.87 vs. 0.66), as was macro-average recall (0.93 vs. 0.90). Finally, the proposed method achieved greater accuracy (0.90 vs. 0.80).

2) Two-people Pose Recognition

The second inference comparison focuses on the two-person detection scenario to see each model's capability in detecting multiple people in the frame; the data for this scenario has a total of 822 frames. These frames are categorized into four actions: 210 for "raise_hand," 212 for "engaged," 214 for "bored," and 186 for "on_the_phone." This diverse set evaluates the model's ability to detect interactions between two subjects performing different actions. Fig. 13 presents an example where both volunteers perform the "engaged" pose, illustrating the model's capacity to accurately recognize simultaneous actions within the same frame.

The baseline method's correct prediction rate dropped significantly for two-person detection, especially in the "bored" and "engaged" classes. Out of 619 frames labeled "bored," only 183 were classified correctly, while "engaged" had only 26 correct classifications out of 164. The "raise_hand" class performed better, achieving a higher correct classification rate. The confusion matrix (Fig. 14) shows that most misclassifications were labeled as "bored," with 619 out of 821 predictions classified incorrectly, leaving only 202 correct. Additionally, the baseline method struggled again with "on_the_phone" classifications. As a result, the macro-average precision was 0.32, the macro-average recall was 0.28, and the overall accuracy was 0.29.

The proposed method, in contrast, showed flawless prediction accuracy for the "on_the_phone" class in this scenario, which previously faced issues. However, misclassifications were more frequent in the "engaged" class, with 93 out of 186 frames misclassified. The confusion matrix for the proposed method (Fig. 18) also highlighted that "bored" was often misclassified as "raise_hand," with 107 out of 214 frames misclassified. Despite these challenges, the proposed method delivered better performance, achieving a macro-average precision of 0.83, a macro-average recall of 0.75, and an overall accuracy of 0.75.

Fig. 13. Two-people pose recognition

TABLE IV
TWO-PEOPLE RECOGNITION RESULT

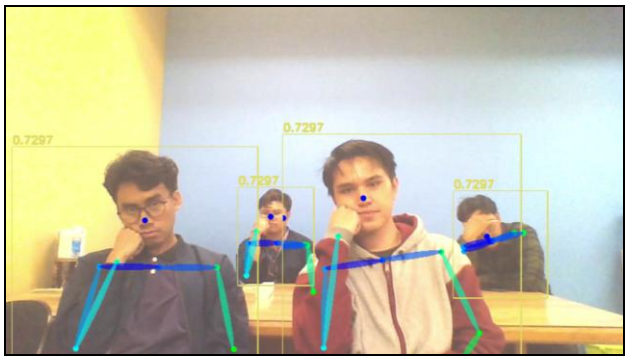
Evaluation Metric	Baseline	Proposed
Total data	822	
Average Cross-entropy Loss	0.71	0.34
Macro-average Precision	0.32	0.83
Macro-average Recall	0.28	0.75
Accuracy	0.29	0.75

Table IV provides a summary of the two-person detection results. The baseline method exhibited a higher cross-entropy loss (0.71 vs. 0.34), while the proposed method outperformed all other metrics. The macro-average precision was significantly higher for the proposed method (0.83 vs. 0.32), as was the macro-average recall (0.83 vs. 0.28). Overall accuracy for the test set was also higher for the proposed method (0.75 vs. 0.29).

3) Four-people Pose Recognition

The four-person detection scenario includes a total of 1298 frames, divided into 488 for "raise_hand," 262 for "engaged," 503 for "bored," and 45 for "on_the_phone." This dataset comprehensively tests the model's ability to track and differentiate between multiple subjects performing various actions simultaneously. An example from the test set is shown in Fig. 14, where all volunteers perform the "bored" pose with their right hand. This example highlights the model's capability to detect even more complex coordinated actions across multiple individuals within the same frame.

The baseline method's correct prediction rate dropped even more than the one-person and two-person detections. Although the baseline method detected "on_the_phone" poses, every classification for this class was incorrect. The highest correct classification rate was for the "bored" pose, frequently misclassified as "raise_hand," with 253 out of 597 instances being wrongly labeled.



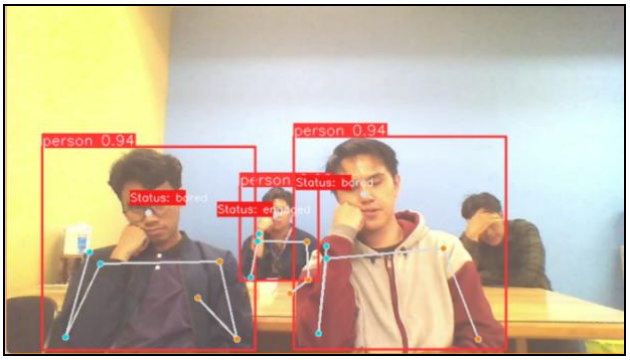


Fig. 14. Four-people pose recognition

Fig. 15. Two-people pose recognition

TABLE V
FOUR-PEOPLE RECOGNITION RESULT

Evaluation Metric	Baseline	Proposed
Total data	1298	
Average Cross-entropy Loss	0.61	0.53
Macro-average Precision	0.27	0.57
Macro-average Recall	0.29	0.55
Accuracy	0.41	0.53

TABLE VI
DIFFERING POSES RECOGNITION RESULT

Evaluation Metric	Baseline	Proposed
Total data	428	
Average Cross-entropy Loss	0.77	0.27
Macro-average Precision	0.22	0.88
Macro-average Recall	0.23	0.55
Accuracy	0.19	0.89

While the proposed method improved overall performance, it also struggled with misclassifications, especially for the "on_the_phone" and "engaged" poses. Misclassifications were similar to the one-person and two-person scenarios. The "on_the_phone" class had the highest rate of error, with 200 out of 393 misclassified as "bored," while the "engaged" class saw 54 misclassifications as "bored" and 65 as "on_the_phone," resulting in 119 incorrect labels from 262 frames.

Regarding metrics, the baseline method had a higher cross-entropy loss (0.61 vs. 0.53). In contrast, the proposed method showed superior macro-average precision (0.56 vs. 0.27), macro-average recall (0.50 vs. 0.29), and overall accuracy (0.53 vs. 0.41). These comparisons demonstrate that while both methods struggled with certain classes, the proposed method consistently outperformed the baseline across all the evaluation metrics.

The four-people pose recognition result summary can be seen in Table V. The four-people pose detection is the final identical pose recognition trial. It can be assumed that the cause for both methods, especially the baseline struggle to classify "on_the_phone" and misclassifications, are mainly classified as the "bored" class. The dataset imbalance most likely causes this, as the "on_the_phone" class only represents the lowest class, representing 8.17% of the total dataset. Also, its keypoint locations are similar to other poses, specifically, "raise_hand" and "bored."

4) Different Poses Recognition

The differing poses detection scenario includes a total of 428 frames, divided into 213 for "raise_hand," 91 for "engaged," 63 for "bored," and 61 for "on_the_phone." Using two volunteers, this dataset evaluates the model's ability to detect different actions performed by multiple subjects within the same frame. Fig. 15 provides an example where one volunteer is shown in the "bored" pose with their, while the other performs the with their left hand "raise_hand" pose, demonstrating the model's effectiveness in recognizing varied actions simultaneously.

The results of the different poses detection test set, summarized in Table VI, reveal a clear contrast in performance between the baseline and proposed methods. One key difference is the cross-entropy loss, where the baseline method showed a significantly higher value (0.77 vs. 0.27), indicating that the proposed method was far more effective in minimizing prediction errors and aligning the predictions closer to the actual labels.

Regarding macro-average precision, the proposed method outperformed the baseline by a large margin (0.88 vs. 0.22). This metric highlights how the proposed method is more precise in its predictions and better at accurately classifying positive samples across the various pose categories. The same pattern is evident in macro-average recall, where the proposed method achieved a considerably higher score (0.55 vs. 0.23). This improvement in recall shows that the proposed method was more capable of identifying true positive cases, effectively minimizing the misclassifications for each pose.

TABLE VII
EVALUATION RESULT

Evaluation		
Metrics	**Baseline**	**Proposed**
Average Test-set Cross-entropy Loss	0.60	0.40
Macro-average Precision	0.22	0.88
Macro-average Recall	0.23	0.55
Accuracy	0.19	0.89

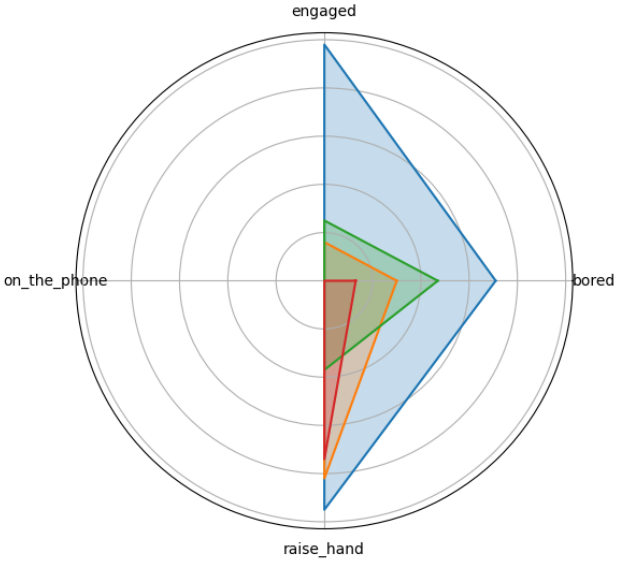
Finally, the overall accuracy of the models provides a clear indication of their performance, with the proposed method demonstrating a significantly better accuracy (0.89 vs. 0.19). This stark difference shows that the proposed method could correctly classify most frames in the test set, whereas the baseline struggled to achieve reliable predictions.

D. Overall result

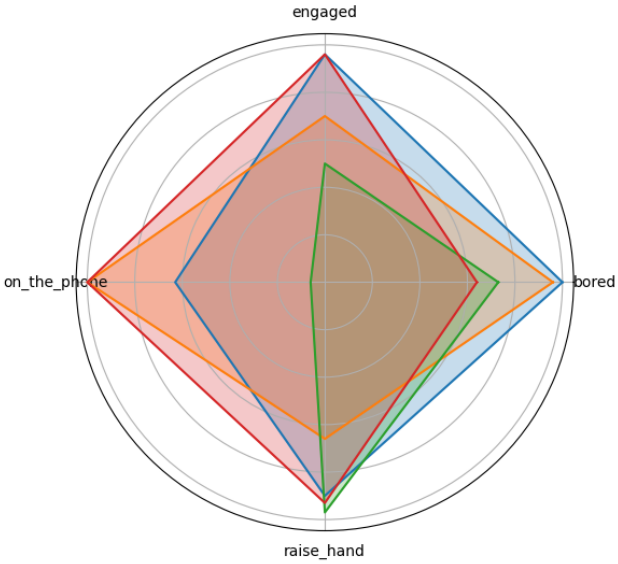
From all the comparisons that have been made to measure the performance of the baseline and proposed. Starting from the data extraction time. The proposed pose extraction approach outperforms the baseline, with an average processing time of 2 minutes and 30 seconds each round, against the 42 minutes and 30 seconds for the baseline. The proposed method takes around 15 minutes, but the baseline method takes roughly 255 minutes, comparable to 4 hours and 15 minutes. This implies that the proposed method's data-collecting procedure is 17 times quicker than the baseline. Regarding scalability, when new data collection is necessary or a continuation of currently collected data must be done, the proposed method will require much less time than the baseline. Determining how robust a model is to change in setting or population is a crucial part of this, and it typically requires applying the model to several independent datasets. The proposed method outperforms the baseline across all key metrics, demonstrating improved accuracy, precision, recall, and a lower cross-entropy loss, as shown in Table VII.

The spider charts shown in Fig. 16 illustrate the prediction performance of different pose recognition models. The legend in Fig. 16 represents the color-coded classifications used in the spider chart to distinguish performance across different pose recognition scenarios: "One-Person Pose Recognition" (blue), "Two-People Pose Recognition" (orange), "Four-People Pose Recognition" (green), and "Differing Poses" (red). Each color aligns with its respective model's performance line, clearly comparing strengths and weaknesses across these classifications. Both models, particularly for the class "on_the_phone," demonstrate weaknesses in prediction accuracy. This class is consistently predicted with lower correctness across the charts. Moreover, the models face further challenges as the number of objects increases, such as moving from one-person to four-people pose recognition. This decrease in performance can be seen across the radar chart, where the overlap between the different models becomes more significant, indicating reduced precision in recognizing specific poses as complexity increases.

The confusion matrix results in Fig. 17 help identify the classification result holistically based on the 3049 data on the test set. The baseline (Fig. 17a) misclassified all the “on_the_phone” classifications. Only the “raise_hand” classification resulted in a higher correct classification rate,



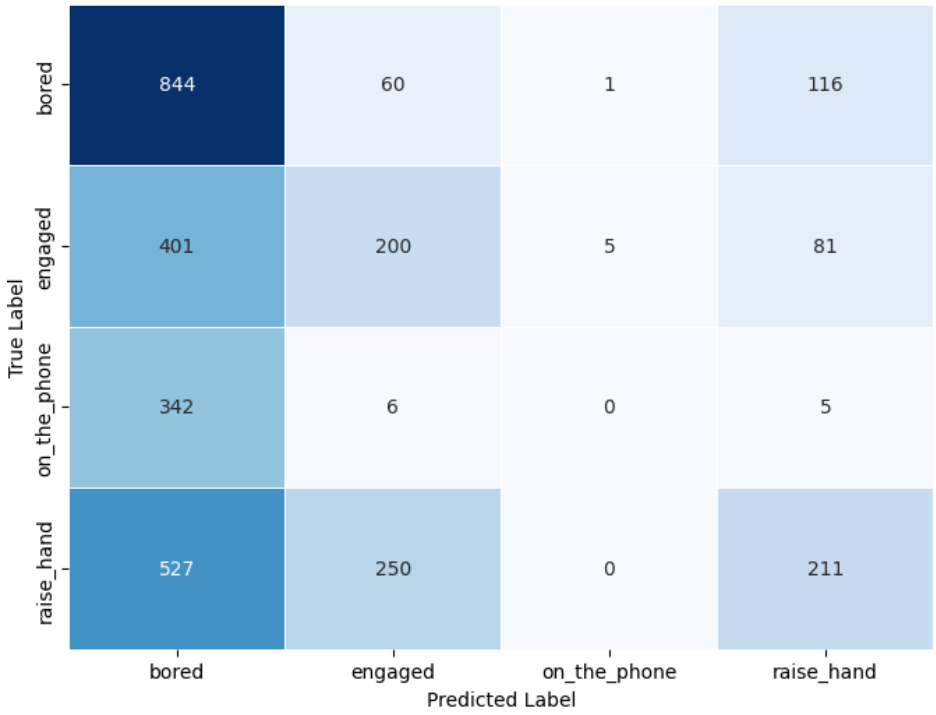
(a) Baseline method



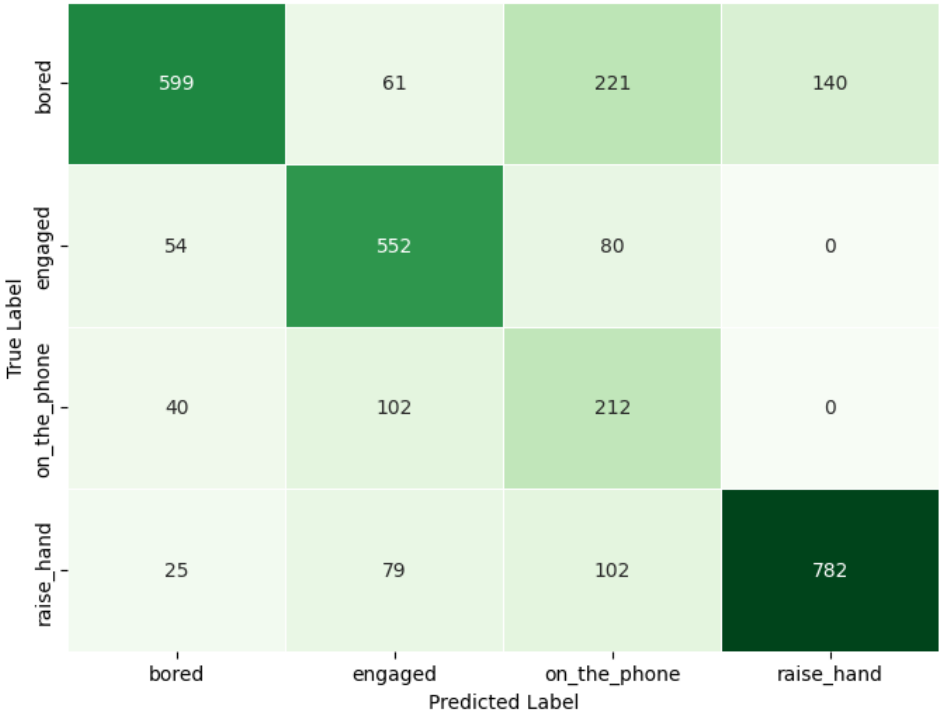
(b) Proposed method

Fig. 16. Comparison of spidercharts

although very slightly. The proposed method’s correct classification result (Fig. 17b) shows that it only struggled on “on_the_phone” classifications, showing a higher incorrect classification rate. All the other classes, however, have a higher correct classification rate. It is worth investigating the challenges both methods faced in classifying "on_the_phone" instances as further improvements for similar scenarios, as the observed disparity in classification performance between the two methods underlines the importance of identifying and addressing the specific challenges encountered by both methods in other similar scenarios.



(a) Baseline method

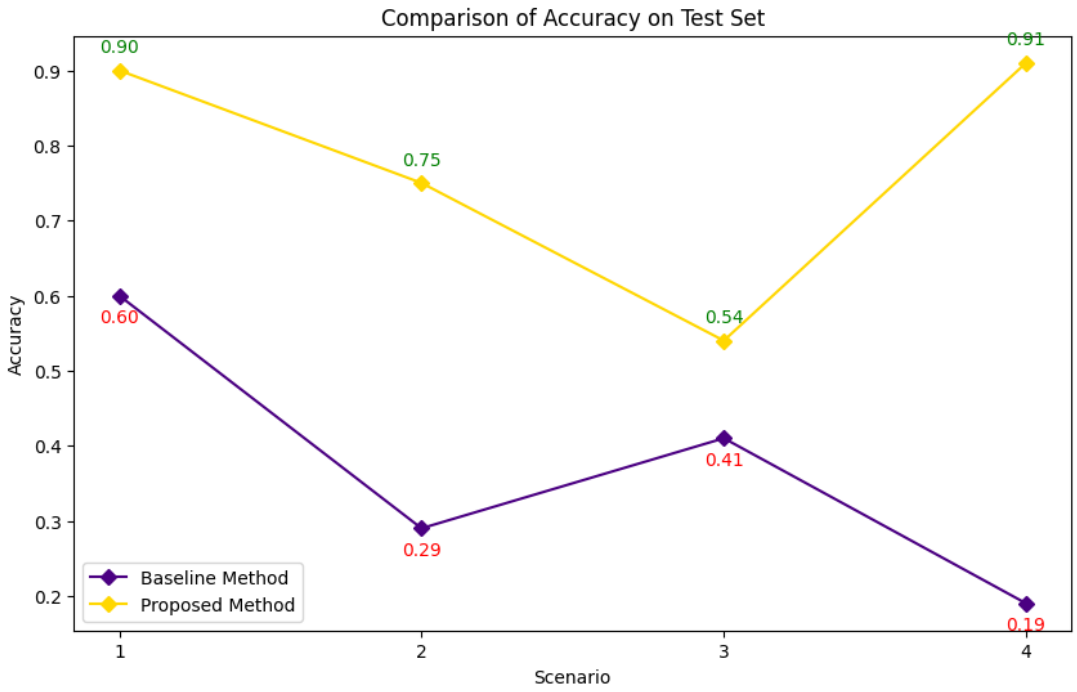


(b) Proposed method

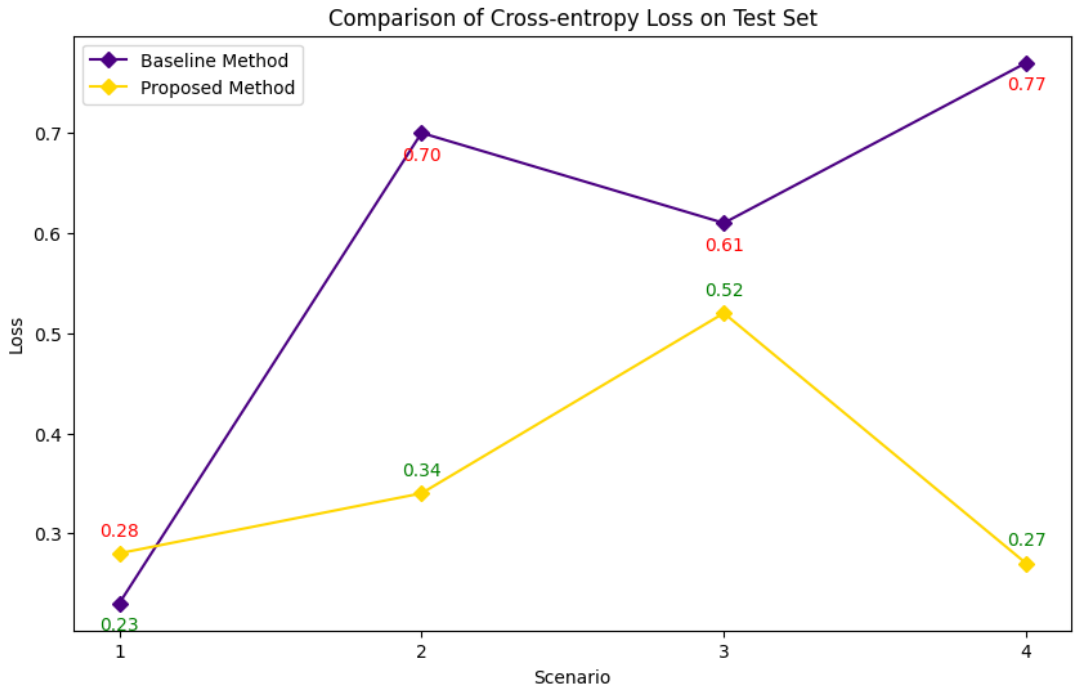
Fig. 17. Comparison of each method's confusion matrix

Evaluation based on accuracy values, where higher values indicate superior performance, reveals that the proposed method consistently outperforms the baseline. Fig. 18 illustrates the results gathered from the testing on each scenario, with green indicating the better result and red indicating the worst result. The baseline method achieved results of 0.60, 0.29, 0.41, and 0.19 across different scenarios, resulting in an overall accuracy of 0.40. In contrast, the proposed method achieved better results with higher accuracy values of 0.90, 0.75, 0.54, and 0.91, giving an overall accuracy of 0.78.

In terms of the cross-entropy loss on the test set, the baseline method records test-set cross-entropy losses of 0.23, 0.70, 0.61, and 0.77, while the proposed method achieves better results with lower cross-entropy test-set losses of 0.28, 0.34, 0.52, and 0.27. The proposed model demonstrates superior performance with lower loss values in three scenarios. It is crucial to highlight that the baseline method's loss increases with more people in the frame, reaching its worst result in the scenario involving differing poses, where the proposed method excels.



(a) Baseline method

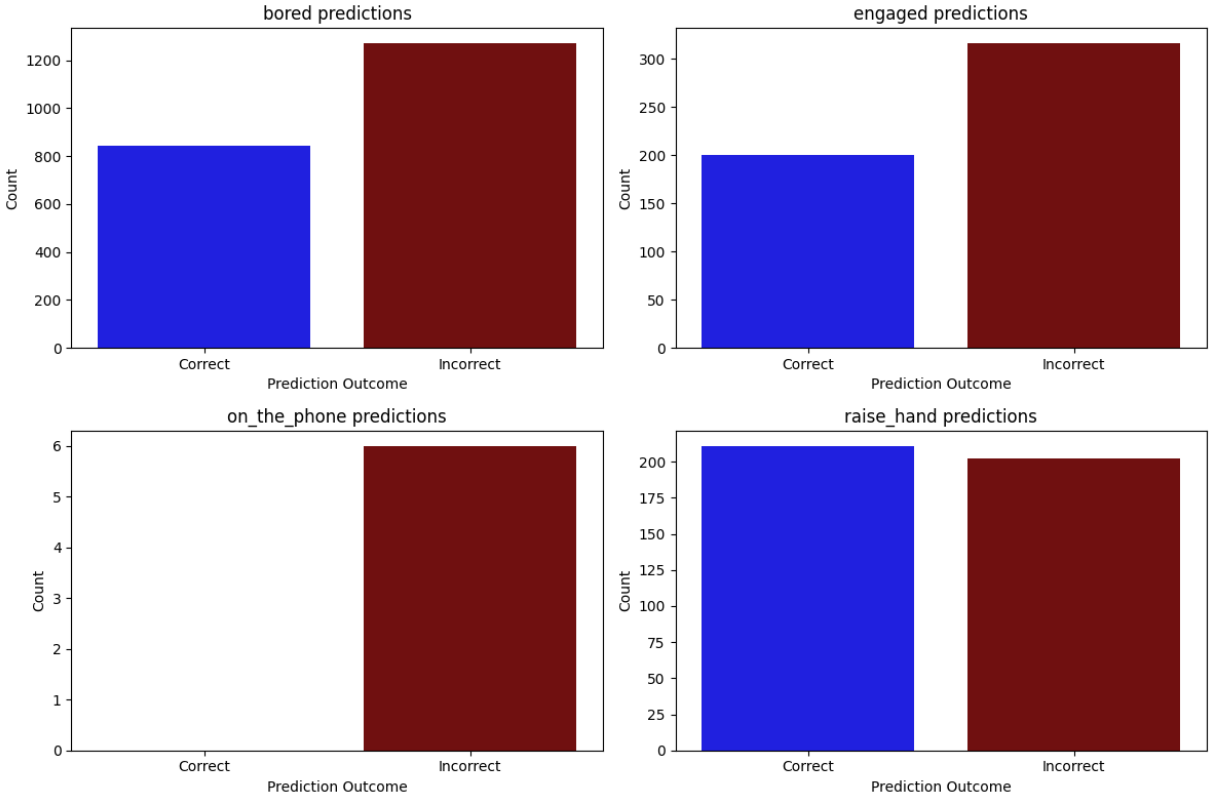


(b) Proposed method

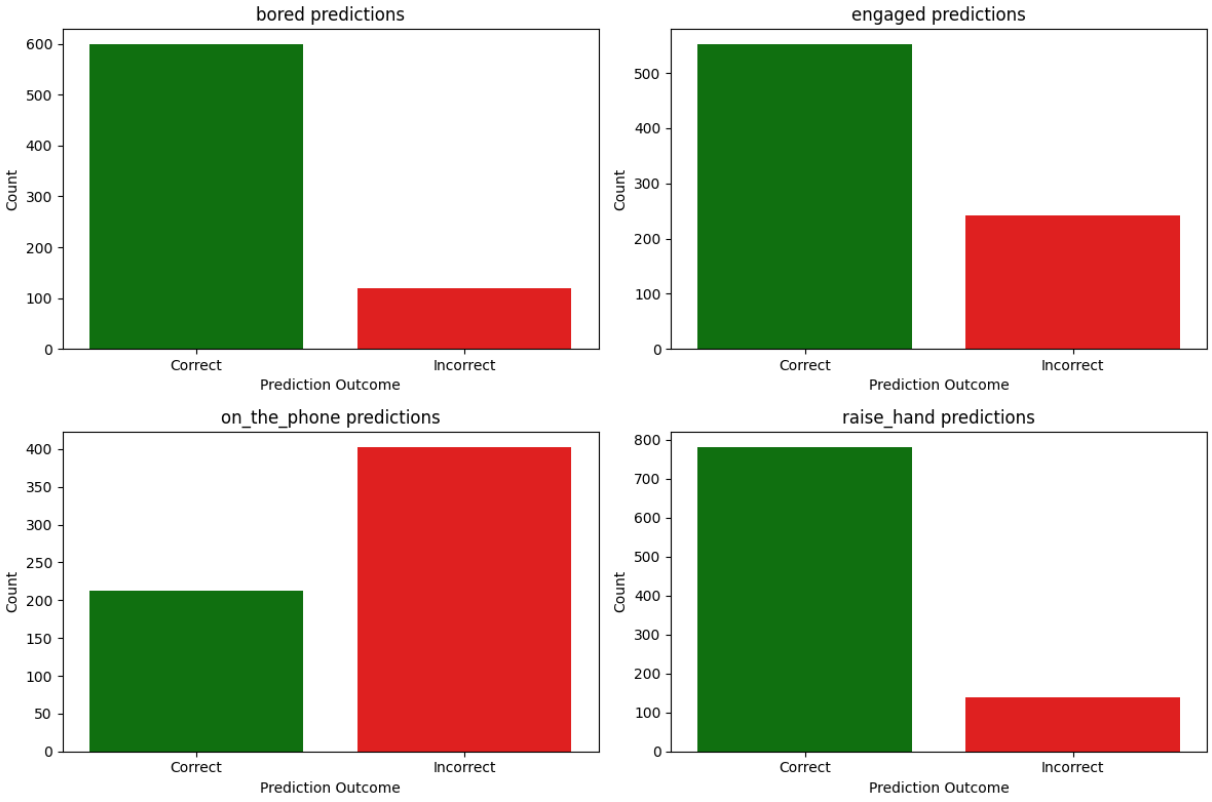
Fig. 18. Comparison of accuracy and test-set loss of the methods

In summary, in both metrics regarding the accuracy of the method and the cross-entropy loss, the proposed method consistently outperforms the baseline method. This consistent higher result in accuracy and lower cross-entropy loss highlights the proposed method's efficacy and potential for improved performance in diverse environments, emphasizing its robustness and reliability in accurate classifications. The proposed model's demonstrably superior accuracy across various conditions.

The cause of differing inference speeds is how the baseline method handles keypoint detection. Baseline uses Part Affinity Fields, and it can capture the relationships between different body parts, helping the network understand how limbs connect and how bodies are structured which includes the ability to capture rich spatial information about key point locations, which facilitates better handling of complex poses and occlusion, as well as offering flexibility in post-processing, enabling techniques like non-maximum suppression for refining predictions [31]. However, it entails the limitation of accuracy by the resolution of the heatmaps, potentially resulting in imprecise localization, and peak detection on heatmaps might not always align with the actual keypoint position, thereby introducing errors, and the error rate of such can be shown in Fig. 19.



(a) Baseline method



(b) Proposed method

Fig. 19. Correct classification of each method

The proposed method uses heatmap regression that can directly predict keypoint coordinates with higher accuracy than peak finding on heatmaps and less sensitivity to heatmap resolution, potentially maintaining accuracy even with smaller feature maps. When heatmap is combined with YOLOv4, the computational expenses make it much slower than the proposed method, and it introduces the problem of double detection when region-of-interest intersects on the same person, which only happens with the proposed method. A clearer side-by-side comparison is shown in Fig. 20.

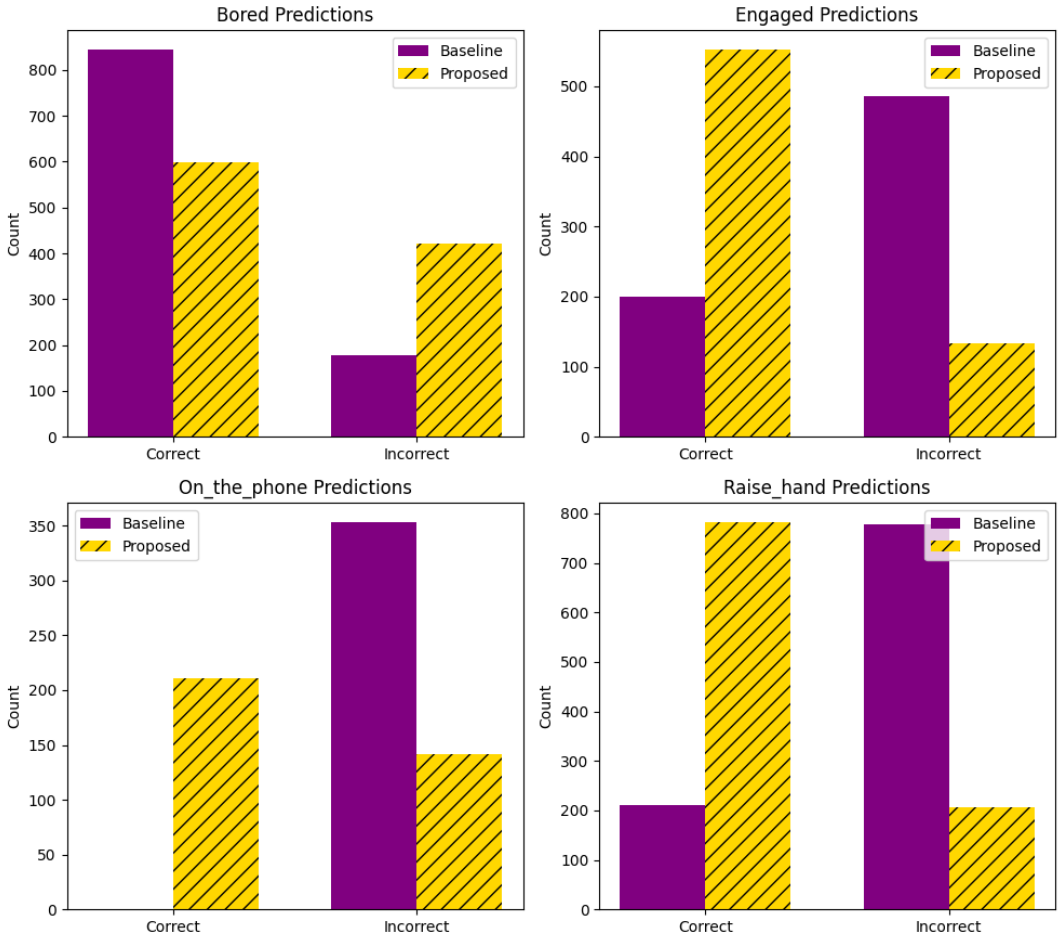


Fig. 20. Overall correct classification chart comparison

To summarize, most of the scenarios resulted in the lower cross-entropy loss of the proposed method, with the correct classification prediction rate shown in Fig. 20, where the baseline method shows a higher misclassification rate when compared to the proposed method. The baseline misclassified all the “on_the_phone” classifications, and only the “raise_hand” classification resulted in a higher correct classification rate, even though it was very slight. Although incorrect “on_the_phone” classification also happened in the proposed method, it showed that more data for the “on_the_phone” class was necessary as the pose was similar to “raise_hand.”

Regarding data extraction time, the proposed pose extraction approach outperformed the baseline, with an average collection time of 2 minutes and 30 seconds each round, against the 42 minutes and 30 seconds for the baseline. The proposed method took around 15 minutes, but the baseline method took roughly 255 minutes, comparable to 4 hours and 15 minutes. This implies that the proposed method's data-collecting procedure is 17 times quicker than the baseline. Regarding scalability, when new data collection is necessary or a continuation of currently collected data must be done, the proposed method will require much less time than the baseline. Determining how robust a model is to change in setting or population was a crucial part of this, and it typically requires applying the model to several independent datasets.

For this reason, the significance of data gathering cannot be overstated [26]. Furthermore, the power law in the setting of additional data yields better findings, and model performances grow as database size scales [27].

While both methods show satisfactory training validation results and the model can generalize well, smoother curves are usually considered good learning curve behavior in terms of better generalization [30]. The proposed method has a softer curve than the baseline, as shown in Fig. 9. When the validation loss lowers up to a particular point before climbing again. At the same time, the training loss keeps decreasing with experience. A plot of learning curves suggests overfitting. A "generalization gap" occurs when a model performs worse on the training dataset than on the validation dataset. When the training loss drops to a stable point, and the validation loss likewise reaches a stable point with a small gap relative to the training loss, a learning curve plot shows a successful match [29].

The proposed method resulted in a lower cross-entropy loss for most scenarios, with a higher correct classification rate than the baseline, which struggled more with misclassification. The baseline misclassified all "on_the_phone" instances, while the proposed method performed better, though both methods faced challenges with this class due to its similarity to "raise_hand." The confusion matrix shows that the baseline mostly misclassified "on_the_phone" as "bored," while the proposed method had higher correct classification rates across other classes but still struggled with "on_the_phone.". Descriptive and inferential statistics are shown in Table VIII and Table IX, respectively.

TABLE VIII
OVERALL RESULTS STATISTIC

Descriptive Statistics		
Metrics	**Baseline**	**Proposed**
Number of Samples (Total)	3049	
Standard Deviation	0.41	0.36
Variance	0.17	0.13
Evaluation		
Metrics	**Baseline**	**Proposed**
Average Test-set Cross-entropy Loss	0.60	**0.40**
Macro-average Precision	0.22	**0.88**
Macro-average Recall	0.23	**0.55**
Accuracy	0.19	**0.89**

TABLE IX
PAIRED T-TEST RESULT

T-statistics	**P-value (α = 0.05)**
-19.3801	5.3538e-80

The result's statistical significance is proven with a paired t-test done on the cross-entropy loss, with $\alpha = 0.05$; therefore, if the p-value < 0.05, the result is deemed statistically significant. In contrast, if the result is statistically significant, H(0) is rejected, and H(1) is accepted; otherwise, if the result is deemed statistically insignificant, H(0) is accepted. The proposed method's mean value is 0.402, with a standard deviation of 0.356 and a variance of 0.127. In contrast, the baseline method exhibits a higher mean of 0.597, accompanied by a standard deviation of 0.410 and a variance of 0.168.

The proposed method substantially improves efficiency, accuracy, and overall performance compared to the baseline approach. With a data extraction process that is 17 times faster, the proposed method significantly reduces processing time, making it more suitable for scalable applications. Moreover, the proposed model consistently outperforms the baseline regarding cross-entropy loss, accuracy, macro-average precision, and recall, achieving superior results across multiple evaluation metrics.

The proposed method exhibits a smoother learning curve, better generalization capabilities, and a lower misclassification rate, especially in challenging pose recognition scenarios. While both methods struggled with the "on_the_phone" classification, the proposed approach still showed a higher correct classification rate. The statistical analysis, reinforced by a paired t-test, reveals a significant difference between the two methods, with a p-value < 0.05, confirming the robustness and reliability of the proposed method. These findings underline the method's potential for diverse real-world applications, offering both computational efficiency and improved classification accuracy.

V. CONCLUSION

The proposed YOLOv8 and MediaPipe combination method is proven better in accuracy and test-set cross-entropy loss. The baseline method, consisting of YOLOv4 and OpenPose combination accuracy, achieved only 0.41 on average. The proposed method managed to outperform the baseline method by scoring 0.70. The same happens with the average cross-entropy loss, which also shows that the baseline is beaten by the proposed method, as the baseline method scored only 0.60, while the proposed method scored 0.40. The paired t-test result is proven to be statistically significant. The comparison in terms of data collection rates for pose detection reveals a considerable difference, with the proposed method showing a remarkable advantage in speed, achieving a rate 17 times faster than the baseline method. The proposed method is not primarily intended for multiple pose detection. However, it can give the same number of person predictions, indicating that the proposed method can do multiple keypoint detection with the help of an object detector [32], [33].

This research suggests several directions for future exploration regarding engagement detection in a classroom environment. First, there is potential to optimize the accuracy of the proposed pose detection method while maintaining its impressive speed, especially in situations where blur and occlusion can be handled [34], [35]. Additionally, adapting the approach to detect multiple poses efficiently could broaden its applicability. The last is the development of hybrid methods that combine the strengths of different techniques that may offer a balanced solution, aiming for versatility and practicality in real-world scenarios [36].

OPEN DATA RESOURCE

Open ssssData is available at: https://github.com/Zayphen/pose_coordinate.

CONTRIBUTORSHIP

Gabriel Asael Tarigan: Conceptualization, Methodology, Writing – Original draft preparation, Writing – Review and editing. **Gregorius Natanael**: Writing – Original draft preparation, Methodology, Writing – Review and editing. **Kuncahyo Setyo Nugroho**: Writing – Original draft preparation, Methodology, Writing – Review and editing. **Bens Pardamean**: Data curation, Writing – Review and editing, Supervision.

REFERENCES

[1] Z. M. Machardy, K. Syharath, and P. Dewan, "Engagement analysis through computer vision," *CollaborateCom 2012 - Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing*, no. November, pp. 535–539, 2012, doi: 10.4108/icst.collaboratecom.2012.250429.

[2] P. Vanneste *et al.*, "Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement," *Mathematics*, vol. 9, no. 3, pp. 1–20, 2021, doi: 10.3390/math9030287.

[3] D. Kučak, V. Juričić, and G. Đambić, "Machine learning in education - A survey of current research trends," *Annals of DAAAM and Proceedings of the International DAAAM Symposium*, vol. 29, no. 1, pp. 0406–0410, 2018, doi: 10.2507/29th.daaam.proceedings.059.

[4] C. Larmuseau, P. Vanneste, J. Cornelis, P. Desmet, and F. Depaepe, "Combining physiological data and subjective measurements to investigate cognitive load during complex learning," *Frontline Learn Res*, vol. 7, no. 2, pp. 57–74, 2019, doi: 10.14786/FLR.V7I2.403.

[5] C. Gao, S. Ye, H. Tian, and Y. Yan, "Multi-scale single-stage pose detection with adaptive sample training in the classroom scene," *Knowl Based Syst*, vol. 222, p. 107008, 2021, doi: 10.1016/j.knosys.2021.107008.

[6] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, 2019, doi: 10.1016/j.tele.2019.01.007.

[7] A. T. S and R. M. R. Guddeti, "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks," *Educ Inf Technol (Dordr)*, vol. 25, no. 2, pp. 1387–1415, 2020, doi: 10.1007/s10639-019-10004-6.

[8] G. A. Tarigan, E. G. Natanael, and B. Pardamean, "A Review of Body Poses Detection in a Classroom Environment for Engagement Assessment," *International Conference on Sustainable and Smart Engineering*, vol. 1, 2023.

[9] F.-C. Lin, H.-H. Ngo, C.-R. Dow, K.-H. Lam, and H. L. Le, "Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection," *Sensors*, vol. 21, no. 16, p. 5314, Aug. 2021, doi: 10.3390/s21165314.

[10] J. Liu, X. Mu, Z. Liu, and H. Li, "Human skeleton behavior recognition model based on multi-object pose estimation with spatiotemporal semantics," *Mach Vis Appl*, vol. 34, no. 3, pp. 1–13, 2023, doi: 10.1007/s00138-023-01396-0.

[11] M. Rashmi, T. S. Ashwin, and R. M. R. Guddeti, "Surveillance video analysis for student action recognition and localization inside computer laboratories of a smart campus," *Multimed Tools Appl*, vol. 80, no. 2, pp. 2907–2929, 2021, doi: 10.1007/s11042-020-09741-5.

[12] L. Tang, C. Gao, X. Chen, and Y. Zhao, "Pose detection in complex classroom environment based on improved faster R-CNN," *IET Image Process*, vol. 13, no. 3, pp. 451–457, 2019, doi: 10.1049/iet-ipr.2018.5905.

[13] Wang, J. Yao, C. Zeng, W. Wu, H. Xu, and Y. Yang, "Learning Behavior Recognition in Smart Classroom with Multiple Students Based on YOLOv5," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.10916

[14] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," pp. 1–15, 2022, [Online]. Available: http://arxiv.org/abs/2207.02696

[15] M. Xu, D. Fralick, J. Z. Zheng, B. Wang, X. M. Tu, and C. Feng, "The differences and similarities between two-sample t-test and paired t-test," *Shanghai Arch Psychiatry*, vol. 29, no. 3, pp. 184–188, 2017, doi: 10.11919/j.issn.1002-0829.217070.

[16] T. K. Kim, "T test as a parametric statistic," *Korean J Anesthesiol*, vol. 68, no. 6, p. 540, 2015, doi: 10.4097/kjae.2015.68.6.540.

[17] H. K. Hamarashid, "Utilizing Statistical Tests for Comparing Machine Learning Algorithms," *Kurdistan Journal of Applied Research*, no. July, pp. 69–74, Jul. 2021, doi: 10.24017/science.2021.1.8.

[18] H. T. Binh, N. Q. Trung, H. A. T. Nguyen, and B. T. Duy, "Detecting Student Engagement in Classrooms for Intelligent Tutoring Systems," *ICSEC 2019 - 23rd International Computer Science and Engineering Conference*, pp. 145–149, 2019, doi: 10.1109/ICSEC47112.2019.8974739.

[19] Y. Xie, S. Zhang, and Y. Liu, "Abnormal behavior recognition in classroom pose estimation of college students based on spatiotemporal representation learning," *Traitement du Signal*, vol. 38, no. 1, pp. 89–95, 2021, doi: 10.18280/TS.380109.

[20] T. S. Ashwin and R. M. R. Guddeti, "Unobtrusive Behavioral Analysis of Students in Classroom Environment Using Non-Verbal Cues," *IEEE Access*, vol. 7, pp. 150693–150709, 2019, doi: 10.1109/ACCESS.2019.2947519.

[21] T. S. Ashwin and R. M. R. Guddeti, "Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures," *Future Generation Computer Systems*, vol. 108, pp. 334–348, 2020, doi: 10.1016/j.future.2020.02.075.

[22] J. Terven and D. Cordova-Esparza, "A Comprehensive Review of YOLO: From YOLOv1 to YOLOv8 and Beyond," pp. 1–33, 2023, [Online]. Available: http://arxiv.org/abs/2304.00501

[23] J. L. Chung, L. Y. Ong, and M. C. Leow, "Comparative Analysis of Skeleton-Based Human Pose Estimation," *Future Internet*, vol. 14, no. 12, 2022, doi: 10.3390/fi14120380.

[24] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding Up Distributed Machine Learning Using Codes," *IEEE Trans Inf Theory*, vol. 64, no. 3, pp. 1514–1529, 2018, doi: 10.1109/TIT.2017.2736066.

[25] A. Rácz, D. Bajusz, and K. Héberger, "Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification," *Molecules*, vol. 26, no. 4, p. 1111, Feb. 2021, doi: 10.3390/molecules26041111.

[26] A. Subbaswamy, R. Adams, and S. Saria, "Evaluating Model Robustness and Stability to Dataset Shift," *Proc Mach Learn Res*, vol. 130, pp. 2611–2619, 2021.

[27] H. Ruiz, M. Chaumont, M. Yedroudj, A. O. Amara, F. Comby, and G. Subsol, "Analysis of the Scalability of a Deep-Learning Network for Steganography 'Into the Wild,'" *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12666 LNCS, no. December 2020, pp. 439–452, 2021, doi: 10.1007/978-3-030-68780-9_36.

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[29] M. J. Anzanello and F. S. Fogliatto, "Learning curve models and applications: Literature review and research directions," *Int J Ind Ergon*, vol. 41, no. 5, pp. 573–583, 2011, doi: 10.1016/j.ergon.2011.05.001.

[30] T. Viering and M. Loog, "The Shape of Learning Curves: A Review," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 6, pp. 7799–7819, 2023, doi: 10.1109/TPAMI.2022.3220744.

[31] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 1, pp. 172–186, 2021, doi: 10.1109/TPAMI.2019.2929257.

[32] T. W. Cenggoro, F. Tanzil, A. H. Aslamiah, E. K. Karuppiah, and B. Pardamean, "Crowdsourcing annotation system of object counting dataset for deep learning algorithm," *IOP Conf Ser Earth Environ Sci*, vol. 195, no. 1, 2018, doi: 10.1088/1755-1315/195/1/012063.

[33] K. Muchtar, F. Rahman, T. W. Cenggoro, A. Budiarto, and B. Pardamean, "An Improved Version of Texture-based Foreground Segmentation: Block-based Adaptive Segmenter," *Procedia Comput Sci*, vol. 135, no. September, pp. 579–586, 2018, doi: 10.1016/j.procs.2018.08.228.

[34] B. Pardamean, F. Abid, T. W. Cenggoro, G. N. Elwirehardja, and H. H. Muljo, t"Counting people inside a region-of-interest in CCTV footage with deep learning," *PeerJ Comput Sci*, vol. 8, pp. 1–21, 2022, doi: 10.7717/peerj-cs.1067.

[35] N. Dominic, Daniel, T. W. Cenggoro, A. Budiarto, and B. Pardamean, "Transfer learning using inception-resnet-v2 model to the augmented neuroimages data for autism spectrum disorder classification," *Communications in Mathematical Biology and Neuroscience*, vol. 2021, no. April, 2021, doi: 10.28919/cmbn/5565.

[36] M. F. Kacamarga, B. Pardamean, and H. Wijaya, "Lightweight Virtualization in Cloud Computing for Research," in *Communications in Computer and Information Science*, vol. 516, 2015, pp. 439–445. doi: 10.1007/978-3-662-46742-8_40.