Discover Computing

RESEARCH Open Access

Real-time classroom behavior and drawing pattern analysis using YOLOv11 for multi-class student activity detection

Hao Fang[1], Shanghai Mu[2], Xingyu Chen[3*], Javeria Naz[4] and Seng Yue Wong[5*]

*Correspondence:
Xingyu Chen
clairetwo978@gmail.com
Seng Yue Wong
wongsengyue@um.edu.my
[1]Institute for Advanced Studies, University of Malaya, 50603 Kuala Lumpur, Malaysia
[2]International College, Rattana Bundit University, 10240 Bangkok, Thailand
[3]Faculty of Education, University of Malaya, 50603 Kuala Lumpur, Malaysia
[4]Department of Information and Technology, The University of Haripur, 22621 Haripur, Pakistan
[5]Academy of Malay Studies, University of Malaya, 50603 Kuala Lumpur, Malaysia

Abstract

Students are important for the growth of a country. Student attention detection plays an important role in ensuring effective learning in the classroom. It is an important area of research as it has the potential to improve the quality of education and help maintain integrity during exams. Further, the students' monitoring is not only helpful in exams, but it is also useful for educators in maintaining discipline in class, such as it helps in the identification of misconduct and allows timely intervention. With the advancement of technology, researchers proposed several methods for students' classroom behavior classification that rely on traditional feature extraction algorithms and machine learning classifiers. These methods faced several challenges, such as the complexity of features, and the presence of more than one class in the same image, which may lead to false classification results. With the emergence of deep learning, recent research studies have focused on the CNN models, but there is still a gap in the domain. In this research work, we have proposed a method based on the advanced deep learning model YOLOv11 for students' behavior detection. The proposed method consists of several major phases, such as preprocessing, model fine-tuning, hyper-parameter setting, Model training, and evaluation of results. The study utilized three publicly available datasets to train the model for analyzing the student class behavior, class attention, and observing the behavioral patterns from the drawing of students. The datasets are preprocessed to improve the contrast, which is later fed to the fine-tuned YOLOv11 model for training. The proposed method achieved the highest mAP of 96.8%. The results of the model show that it has the potential to be implemented in a real-time classroom monitoring environment. Code: https://www.kaggle.com/code/jianaz786/v5-students-behavior-yolo57811

Keywords Student behavior detection, YOLOv11 model, Deep learning, Object detection, Student classroom behavior

1 Introduction

Student behavior affects academic performance significantly. Recently attitude and behavior detection approaches have emerged significantly [1]. The traditional student monitoring methods involve the analysis of a few students' behavior and struggle

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

to accurately monitor the behavior of all students. Teachers can monitor the students' behavior in an interactive class environment such as a question-answer session, but as the number of students is usually large in class it is a challenging task to interact one-by-one with all students daily [2]. Some students have behavioral disorders or emotional disorders that can lead to further issues such as suspension, and low attendance percentage, and could have several other negative impacts on student health and personality [3, 4]. The manual monitoring of students makes it difficult for the staff and parents to precisely assess the quality of education [5]. The recent trend in the usefulness of deep learning and its emergence in various domains have affected student behavior analysis and detection positively [6]. Several methods are proposed for student classroom behavior detection [7]. These methods are useful for parents, teachers, and students to understand the learning patterns of students [8]. The application of computer-aided methods for students performance monitoring is helpful for both parents and teachers to accurately monitor the student's performance during class time [5]. The recent methods are divided into three major classes such as video-based action recognition [9], object detection methods, and pose estimation methods [10]. The video-based detection methods are good for continuous monitoring of student's behavior. The video-based methods are computationally expensive in terms of hardware and dataset as they need large amounts of annotated samples [11]. There are several datasets available online for student behavior monitoring such as Slow-Fast based detection using the AVA dataset that comprises around 1.58 million annotations. The existing studies have utilized pose estimation methods for classroom posture detection, student attention monitoring, student activity detection, and so on. These existing methods still have some limitations such as these methods are not working precisely for behavior monitoring such as pose estimation methods focusing on motion monitoring not the behavioral properties of students. Motivated by the importance of students' behavioral monitoring in this study we have proposed a method that is based on the You Look Only Once (YOLO) model. The study utilized the Yolov11 pretrained model and using transformer learning, we have trained the model on student's behavior dataset.

1.1 Contributions

Existing approaches primarily target the recognition of a single action or activity. In contrast, this study utilizes three datasets encompassing a large number of classes and proposes a multi-class object detection framework based on the YOLOv11 architecture, enabling simultaneous detection and classification of diverse activities. The proposed study focuses on the integration of deep learning powerful model YOLOv11 for the student's attention tracking and behavior detection. The proposed study has the following contributions:

1. A comprehensive preprocessing image processing pipeline is proposed for enhancing the dataset quality.
2. Three publicly available datasets are preprocessed and augmented to increase sample size. The Yolov11 model is trained and performance is validated using standard performance measures.
3. We have designed a flask based full-stack web application, and seamlessly integrated the trained YOLOv11 model to enable real-time inference and evaluation on unseen test data.

4. We conducted extensive qualitative and quantitative performance evaluation, including rigorous experimental analysis to validate model accuracy, generalization capability, and deployment effectiveness.

The remaining article is organized as: Sect. 2 discusses the recent studies in the domain, Sect. 3 discusses the proposed methodology, Sect. 4 presents the experiments setup details, Sect. 5 gives insights into the experiments performed, Sect. 6 concludes the findings and gives future directions.

2 Related works

The rapid development of deep learning models has revolutionized various domains such as agriculture, medicine, surveillance, and several types of object detection [12]. Student activity monitoring during exams is a concerning area, as academic dishonesty is crucial. Several methods have been proposed in the literature for monitoring student's activities during exams. In a study [13], authors proposed a method that focuses on the iris, head, and hand movement tracking to detect irregular student behavior detection during exams. This study utilized the labeled dataset of physical movement activities of students such as abnormal student behavior or normal behavior. They have encountered the detection of abnormal student behavior during exams.Several studies proposed methods based on YOLOv7 and YOLOv8 for student classroom activity detection [14–16]. Another approach [17] based on remote server technologies and RFID was proposed for face detection. This method was computationally costly, and it focuses on face detection in exams. The dataset was collected from the surrogate examinees during examination set-ups. AdaBoost is utilized for monitoring the students' presence using the images captured from several angles. The model struggled to achieve reliable performance and achieved the highest accuracy of below 75%. The statistical features are extracted in the study. They have designed a model for the impaired invigilators. The authors have not specified the dataset used in the study [?]. Zaffar et al. [18], utilized DeepSort and YOLOv7 models for student detection and tracking from the video data. A real-time alert is generated for suspicious student behavior detection, and an email will be sent to the exam invigilator through SMTP. They have achieved an accuracy of 90.67% for real-time student detection and cheating detection. An improved YOLOv8 model is utilized in the study for student classroom behavior detection. The study embedded the Res2Net block with the original YOLOv8 C2f module, which resulted in improved detection performance [5]. Several researchers have proposed methods for student classroom behavior detection. The unavailability of a suitable number of datasets is a serious concern for researchers.

A comparison of existing models is given in the Table 1. Other researchers dealt with the single-action recognition or fewer behaviors, which could not be used in a dynamic classroom setting because of the multiple activities happening at the same time. Also, a number of them were based on small or controlled data, which is restrictive in terms of extrapolation to practical conditions. Several studies are proposed to analyze the student's behavior and psychological pattern analysis [19, 19–21]. Although a one-stage detector increased the inference speed, not every study discussed the issues of multi-class scalability or efficiency in a real-life monitoring environment. Also, the lack of incorporation into behavioral analytics (e.g. activity frequency or duration analysis) limits the practical usefulness of the activity in higher education. The existing literature

Table 1 Comparison of traditional detectors and YOLO versions

Model	Architecture Type	mAP (%)	FPS	Real-Time Suitable
R-CNN	Two-Stage CNN	53–58	∼5	No
Fast R-CNN	Two-Stage CNN	60–65	∼7	No
Faster R-CNN	Two-Stage CNN	70–75	∼10	Limited
SSD	One-Stage CNN	72–77	∼22	Moderate
YOLOv3	One-Stage CNN	78–80	∼30	Yes
YOLOv5	One-Stage CNN	85–88	∼45	Yes
YOLOv8	One-Stage CNN	88–90	∼55	Yes
YOLOv11	Optimized One-Stage (Enhanced Backbone + Neck)	90+	>60	Highly Suitable

points to the obvious gap in research: the necessity of a powerful, real-time multi-class framework to be efficient in the classroom during the lessons, and to address various behavioral patterns of students. The reason behind this gap is that the current approach seeks to integrate the capability of inference in real-time with the ability to detect all forms of behavior to facilitate practical classroom monitoring applications.

Previous CNN-based object detection algorithms like R-CNN, Fast R-CNN, Faster R-CNN, and SSD had restrictions when applying to real-time monitoring because they have multi-stage detection, are computationally complex, and have relatively slow inference. These models had a tendency to need region proposal steps and powerful hardware, and were not ideal in live video-based systems. These gaps are, instead, overcome by extending YOLOv11 in the single-stage trainable end-to-end architecture, which localizes and classifies by location, thereby reducing latency drastically. This design is optimized and lightweight, hence allowing more frames per second (FPS), enhanced efficiency, and greater adaptability to a dynamic environment, and thus better suited to real-time inference when used in live monitoring applications.

3 Methodology

In this work, we use a modified YOLOv11 model for student behavior detection. The model is fine-tuned, and using the transfer learning approach, we have trained the model after fine-tuning. This research work comprises several phases, including dataset loading, preprocessing, model fine-tuning, model training, and evaluation of results. In this section, we have discussed the dataset used in this research along with the preprocessing employed on the dataset before the model training phase. Performance measures, model architecture, hyperparameter settings, and experimental setup are also discussed in detail.

3.1 Dataset description

In this section, the dataset used in this research project is briefly discussed. In this study, students' class behavior and action detection tasks are performed. In this project, we have utilized three datasets STEAM dataset, the SCB dataset, and the student drawing behavior dataset. Some sample images of these datasets are shown in the Fig. 1.

1. The STEAM dataset [22], consists of seven classes, such as Look at your phone, Playing with Pen - Taking a pen, Playing with fire, Playing with watch, Reading the book or notes, Sleeping during exam, and reading. consisting of 799 training images, 228 validation images and 114 images are reserved for test.

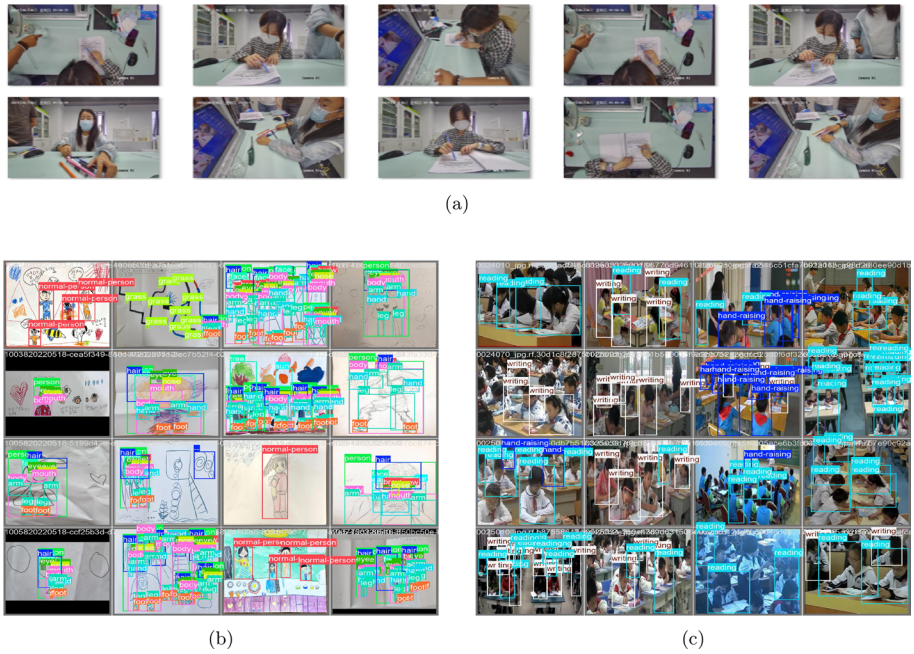


Fig. 1 Comparison of dataset sample: **a** Sample image from the STEAM drawing dataset, **b** Validation image of Child Drawing Dataset, and **c** Validation image of SCB dataset

2. The second dataset that we have utilized for students' action recognition is the SCB dataset [23], which consists of three classes: hand-raising, reading, and writing. The dataset is split in 70:20:10 ratio for train, val, and test sets. It consists of 3520 training images, 996 validation images, and 503 test images.
3. We have also utilized the children's drawing dataset [24] to analyze and detect the children's psychological behavior patterns. This dataset consists of 2078 images, each drawing have multiple objects. The dataset includes the following classes: car, face, smoke, bird, tree, door, hand, window, flower, food, house, eye, animal, grass, nose, mouth, butterfly, tooth, person, body, cloud, roof, fence, chimney, foot, mountain, arm, brow, ear, hair, leg, normal-person, rainbow, and sun.

3.2 Preprocessing

In this phase, we have performed dataset preprocessing to improve the image quality. The first step of this phase is to resize the dataset images to 640×640 dimensions. The dataset contrast is improved using the histogram equalization algorithm. The mathematical formulation of histogram equalization is given in equation (1).

$$H_{EQ} = F(r) = (U-1)\sum_{i=0}^{r} p_r(r_i) \tag{1}$$

Where: H_EQ is the equalized output image, U represents the total gray levels which are 256, $p_r(r_i)$ is the probability of intensity level (r_i). The $p_r(r_i)$is computed using p_i/p where, p_i is the total number of pixels with intensity level r_i and p is the total number of pixels in the image. After preprocessing the dataset contrast, we performed data augmentation. Two methods are used to augment the dataset for increasing the size which helped us in reducing the potential risk of model overfitting. In this research work, we have employed horizontal flip and clockwise rotation at 90 degrees. The final

preprocessed dataset is then utilized for the model training. Sample output images of the preprocessing phase are shown in Fig. 2.

3.3 Performance measures

In this research work, performance evaluation will be performed using several performance measures because in some cases accuracy is not sufficient to assess the model's efficiency. A weighted combination of CIoU loss (box regression), Binary Cross Entropy loss (classification), and Distribution Focal Loss (bounding box refinement).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

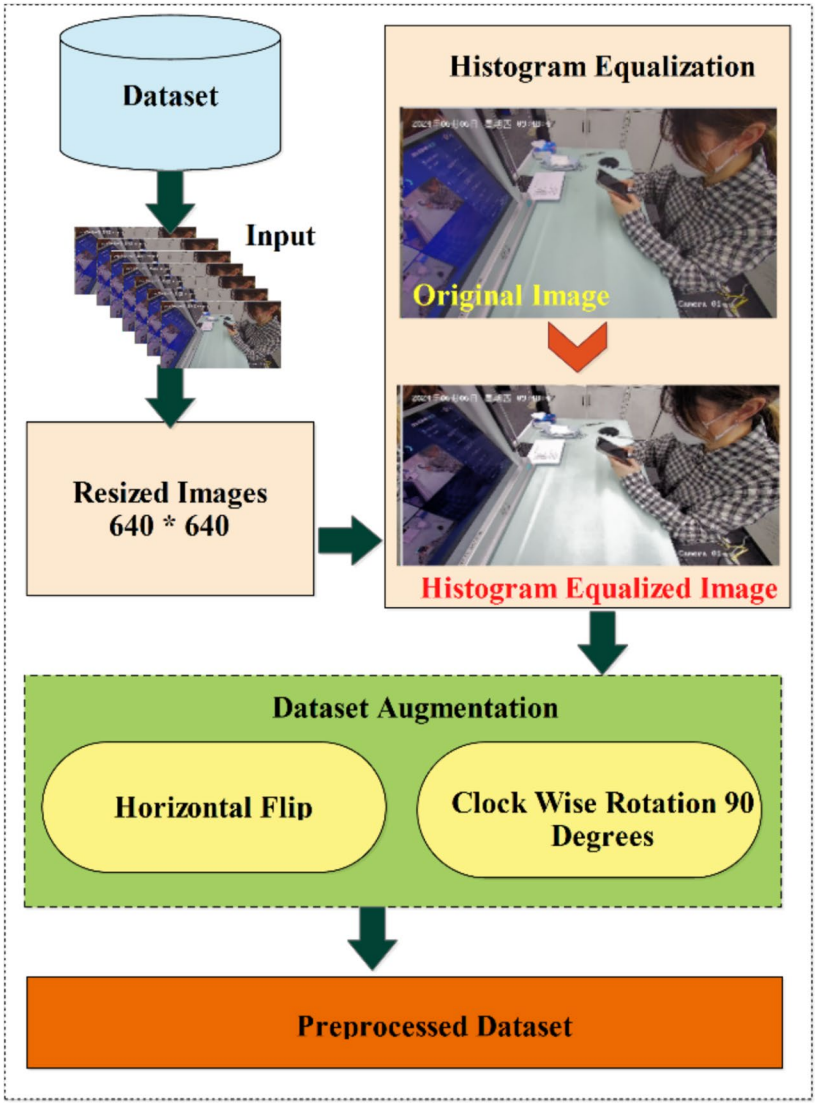


Fig. 2 Proposed method preprocessing technique

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{AP} = \sum_{n} (\text{Recall}_n - \text{Recall}_{n-1}) \, \text{Precision}_n \tag{5}$$

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \tag{6}$$

Whereas TP, FP, TN, FN, Acc, Recall, Precision, AP, and mAP are the short terminologies utilized for true positives, false positives, true negatives, false negatives, accuracy, recall, precision, average precision, and mean average precision, respectively. Here, $Recall_n$ and $Recall_(n-1)$ are the adjacent neighboring values, n is the discretizations index used for the identification of sampling point over the precision-recall curve, $Precision_n$ is the highest value of the precision at the recall level, and N represents the total number of classes.

3.4 Model architecture

In this study, we have utilized the YOLOv11 model which is a real-time object detection algorithm used to precisely classify or detect the region of interest. This model has advanced features as compared to previous versions of YOLO models. Key features include the enhanced feature pyramid network (FPN), anchor-free heads, refined loss function, more advanced transformer-oriented backbone, and modular designs for custom tasks. The model's backbone extracts the wide-range dependencies, whereas the FPN utilizes the dynamic head for the accurate detection of small and large objects. The attention mechanism focuses on the channel features and spatial features. The model uses refined loss functions for the bounding box regression and to handle class imbalance. The architecture of the YOLOv11 model is given in Fig. 3.

3.5 Hyperparameters

In this research work, we have fine-tuned the YOLOv11 model. The list of hyperparameters used in this study is given in Table 2. This training environment is aimed at effective and stable YOLO-based object detection. A learning rate of 0.001 guarantees the gradual changes in weights throughout the backpropagation process and encourages the gradual convergence with AdamW optimizer that also uses weight decay to reduce overfitting and enhance generalization. The images are downscaled to 640 x 640 which is an excellent trade-off between detecting small objects and performance. A batch size of 32 aspect offers steadfast gradient estimations and it is attainable by medium-frame GPUs, and applying 2 data-loading workers eliminates CPU bottlenecks. The training lasts 300 epochs to enable comprehensive feature learning and includes a patience of 25 epochs to early terminate training should no improvement be found to avoid overfitting and waste the computational resources. This design is the right balance of accuracy, efficiency and resource management.

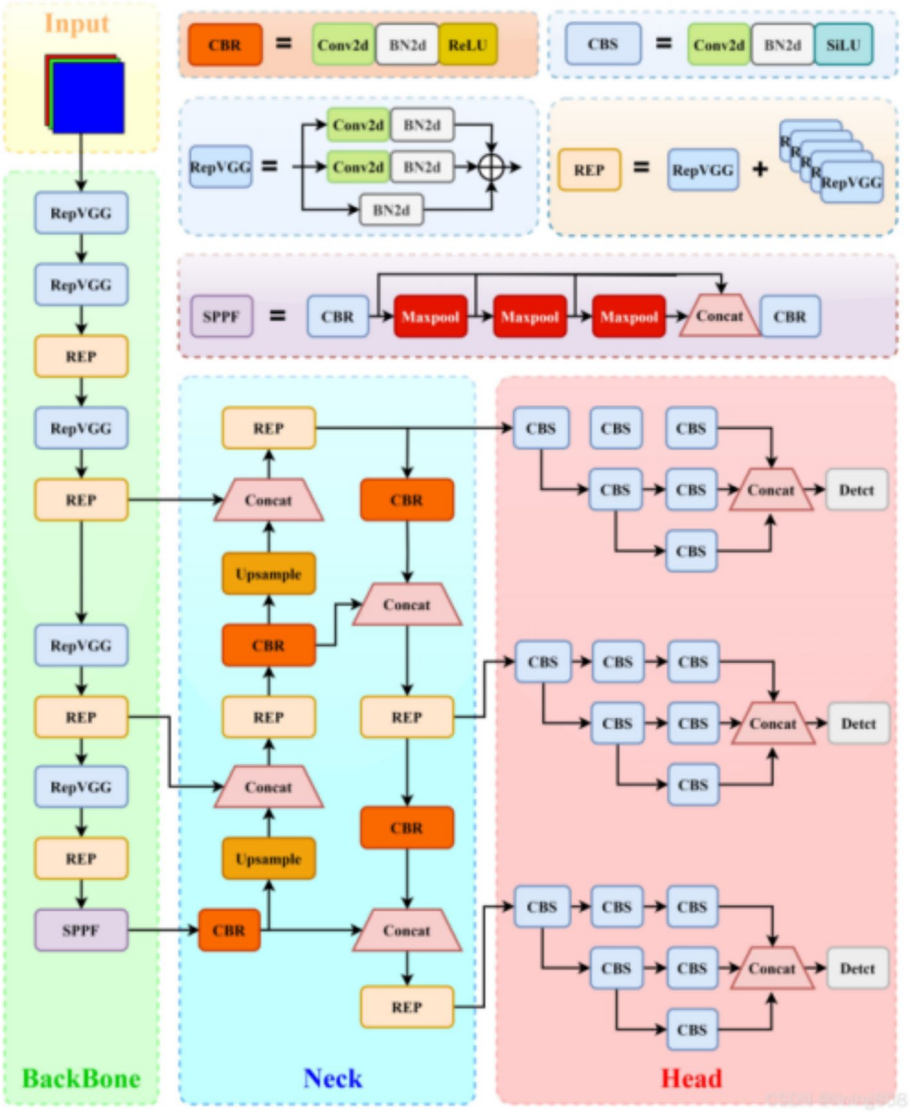


Fig. 3 The structure of the YOLOv11 model [25]

Table 2 Training hyperparameters

Sr.#	Name of parameter	Value
1	Learning Rate (LR)	0.001
2	Image Size	640×640
3	Optimizer	AdamW
4	Batch Size	32
5	Workers	2
6	Epochs	300
7	Patience	25 epochs
8	Early Stopping	True

4 Experimental setup

In this experiment, we have utilized the Kaggle Integrated Development Environment for the implementation of the proposed modified YOLOv11 model architecture for classroom behavior detection. The deep learning model is trained using the GPU-P100

Table 3 Test-1 results for students behavior detection using YOLOv11-n model

Classes	Box(P)	Recall	mAP@50	mAP@50–95
Class 1	0.935	0.883	0.954	0.809
Class 2	0.962	0.879	0.965	0.692
Class 3	0.949	0.869	0.945	0.806
Class 4	0.886	0.863	0.966	0.770
Class 5	0.948	1.000	0.991	0.957
Class 6	0.990	1.000	0.995	0.892
Class 7	0.983	0.900	0.957	0.901

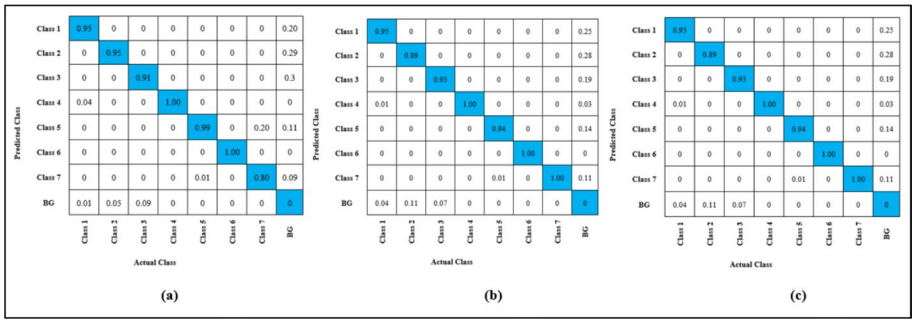


Fig. 4 Test 1 confusion matrix: **a** Test 1 results using YOLOv11-n model, **b** Test 2 results using YOLOv11-s model, and **c** Test 3 results using YOLOv11-m

and a system having Core-i7 with 8 GB RAM is utilized for the implementation of the proposed framework. Moreover, Python programming language is utilized in this study.

4.1 Experiments 1: student behavior and action detection using STEAM dataset

In this section, a brief description of experiments performed in this research work for student behavior detection is provided. The study is reliant on the YOLOv11 model. We have trained the model for 300 epochs with an early stopping criteria and patience of 25 epochs. The study evaluates the performance of the YOLOv11 model using three model architectures such as YOLOv11-n, YOLOv11-s, and YOLOv11-m (whereas n, m, and s stand for nano, small, and medium YOLOv11) for the detection task. These experiments are detailed in the following subsections:

4.1.1 Test-1: students behavior detection using YOLOv11-n model

In this Test, the YOLOv11-n model is fine-tuned and utilized for the detection of students' behavior using the preprocessed STEAM dataset. The dataset is resized to the dimensions 640×640. We have fine-tuned the model from 80 classes to 7 classes. The task is set to the detection, and model training is performed for 300 epochs with enabled early stopping criteria. A patience of 25 is utilized in case model performance does not improve for 25 epochs model training stops automatically. In this experiment, the model training stopped at the 211th epoch and the best results were saved at epoch 186. The best weight file is later utilized for making the prediction on the test set to evaluate the model performance on YOLOv11-n architecture. The quantitative results of this experiment are shown in Table 3. The results can be rectified from the confusion matrix given in Fig. 4a.

The results of this experiment show that the model achieved relatively high detection performance for class 6 in terms of Box(P), Recall, and mAP50. The class in which

we achieved the highest results and the overall performance for all classes are bold in the Table 3. Overall, we have achieved 0.950, 0.913, 0.968, and 0.833 of average Box(P), Recall, mAP50, and mAP50-95, respectively, for all the classes.

4.1.2 Test 2: student behavior detection using the YOLOv11-s model

In this experiment, we have used the fine-tuned YOLOv11-s model for the student's classroom behavior detection. The resized dataset (640×640) is used to train the YOLOv11-s model with a batch size of 32 and early stopping criteria for 300 epochs. The model training stopped early at epoch 155. A patience of 25 is utilized in case model performance does not improve for 25 epochs model training stops automatically. The model's best weights are saved at epoch 130. The best weight file is later utilized for making the prediction on the test set to evaluate the model performance on YOLOv11-s architecture. The quantitative results of this experiment are shown in Table 4. It is observed from the results that the model achieved the highest value of Box(P), Recall, and mAP50 for class 6 the results are bold in Table 4. The overall results for all classes are 0.916, 0.934, 0.962, and 0.810 for average Box(P), Recall, mAP50, and mAP50-95 respectively. The confusion matrix of this experiment is shown in Fig. 4b and precision recall curve is given in the Fig. 6-(a).

4.1.3 Test-3: students behavior detection using YOLOv11-m model

The YOLOv11-m model is fine-tuned and utilized for the detection of students' behavior using the preprocessed STEAM dataset. The dataset is resized to the dimensions 640×640. We have fine-tuned the model from 80 classes to 7 classes. The task is set to the detection and model training is performed for 300 epochs with enabled early stopping criteria. A patience of 25 is utilized in case model performance does not improve for 25 epochs model training stops automatically. In this experiment, the model training stopped at the 99th epoch and the best results were saved at epoch 74. The best weight file is later utilized for making the prediction on the test set to evaluate the model performance on YOLOv11-m architecture. The quantitative results of this experiment are shown in Table 5, whereas the confusion matrix of this experiment is shown in Fig. 4c. The results of this experiment show that the model achieved relatively high detection performance for class 6 in terms of Box(P), Recall, and mAP50. The class in which we achieved the highest results and the overall performance for all classes are bold in Table 5. Overall, we have achieved 0.932, 0.928, 0.964, and 0.833 of average Box(P), Recall, mAP50, and mAP50-95 respectively for all the classes. Quantitative results shows that Test 1 achieved promising results on STEAM dataset.

Table 4 Test-2 results for students behavior detection using YOLOv11-s model

Classes	Box(P)	Recall	mAP50	mAP50-95
Class 1	0.921	0.867	0.923	0.704
Class 2	0.937	0.846	0.930	0.641
Class 3	0.954	0.907	0.960	0.816
Class 4	0.946	0.970	0.989	0.788
Class 5	0.987	0.947	0.993	0.940
Class 6	0.988	1.000	0.995	0.892
Class 7	0.679	1.000	0.946	0.886
All	0.916	0.934	0.962	0.810

Table 5 Test-3 results for students behavior detection using YOLOv11-m model

Classes	Box(P)	Recall	mAP50	mAP50-95
Class 1	0.862	0.847	0.907	0.764
Class 2	0.960	0.868	0.972	0.688
Class 3	0.947	0.897	0.954	0.816
Class 4	0.861	0.944	0.943	0.788
Class 5	0.988	0.981	0.994	0.968
Class 6	1.000	0.988	0.995	0.895
Class 7	0.906	0.967	0.986	0.915
All	0.932	0.928	0.964	0.834



(a) (b)

Fig. 5 Visualization of YOLOv11-n model curves: **a** Using STEAM dataset, **b** Using SCB Dataset

4.1.4 Experiments 2: student action detection using SCB dataset

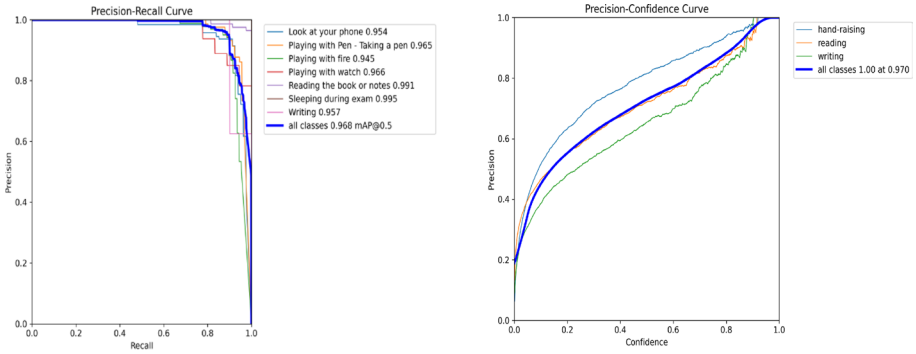
In this experiment, we have utilized the preprocessed SCB dataset for the detection of students' school behavior detection. The model is evaluated using 100 epochs, a batch size of 32, image size of 640×640, and a learning rate of 0.001. The model is trained and performance is evaluated using unseen images to validate model efficacy. This dataset consists of three classes as described in the Section 3.1. Proposed model achieved significantly better results for multi class detection of student action. YOLOv11-n model achieved above 90% accurate class prediction. A few predicted sample images are shown in the Fig. 5a.

4.1.5 Experiment 3: student behavioral pattern detection using drawing dataset

In this experiment, we have utilized children's drawings to analyze the psychological patterns of children. The model is evaluated over 100 epochs with a batch size of 32, an image size of 640×640, and a learning rate of 0.001. Its performance is assessed using unseen images to reserved for testing using learned weights. A few predicted sample images are shown in the Fig. 5b. Precision confidence curve is given in the Fig. 6-(b)

5 Web application development

In this research work, we have employed utilized three architectures of YOLOv11 model for student behavior analysis and attention monitoring. The proposed study uses the STEAM dataset for experimentation. The overall highest mAP of 0.968 is achieved using YOLOv11-n model architecture.



(a) (b)

Fig. 6 Results of YOLOv11-n model prediction: **a** STEAM dataset Precision-Recall curve of YOLOv11-n model for students' behavior detection, **b** SCB dataset Precision Confidense Curve

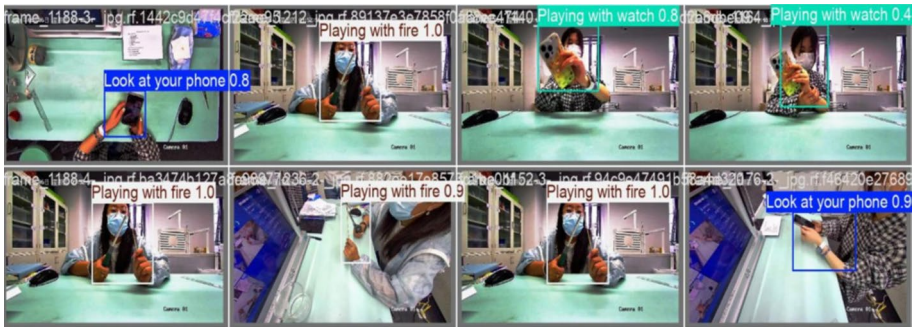


Fig. 7 Sample predicted results on validation images

The experiment precision-recall curve is shown in the Fig. 6, and validation results using STEAM dataset are given in the Fig. 7. In this study, we have proposed a web application named as Student Behavior Detection (SBD-App). The proposed web application is a complete Flash-based web application that is integrated with the YOLOv11 model. A simple interactive interface is designed using HTML, CSS, and JavaScript. There is a button that is used to upload the image that is saved securely by the backend using a unique filename. The uploaded image is passed to the YOLOv11 model (best.pt) weights are loaded using Ultralyics library.

Once the detection is complete, the results with bounding boxes are drawn on the detected object and the predicted image is saved at the output directory and displayed on the web application. The application uses the cv2 library (OpenCV) to save the output image, and Flask's render template function to switch between the upload and result pages. The front end is straightforward, consisting of a file input field and a submit button that sends the image to the server. This setup is ideal for tasks such as detecting student behavior or object presence in classroom images, providing an interactive and user-friendly interface for real-time object detection. The proposed web application interface is shown in the Fig. 8.

6 Conclusion

Student attention detection is an active research area. It play an important role in ensuring improved learning in the classroom. The maintenance of integrity and quality of education cannot be neglected as it helps to ensure the maintenance of discipline in

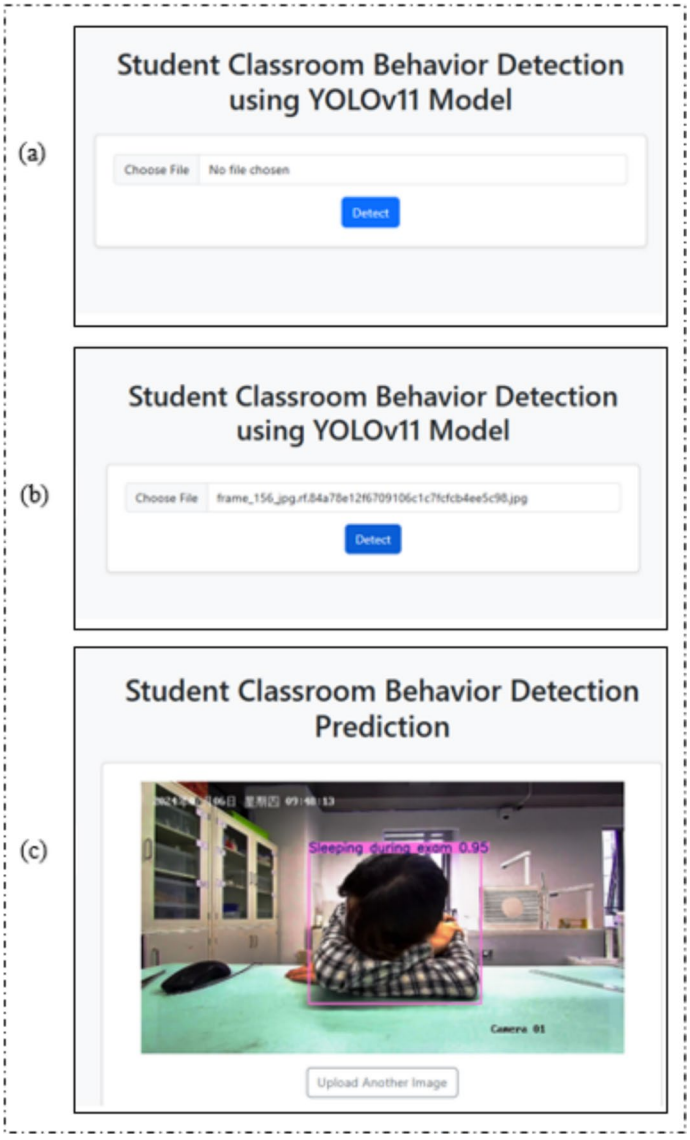


Fig. 8 Proposed SBD-App interface **a** First interface to choose the image, **b** After image upload, and **c** Output image

the class also it is helpful in exam monitoring. Previous CNN models often struggled with real-time inference, making them unsuitable for live monitoring due to slow detection speeds and high computational costs. YOLOv11 addresses these gaps by offering a lightweight, optimized architecture that delivers fast, accurate object detection, enabling efficient real-time monitoring without sacrificing performance. In this research work, we have proposed an automated method for classroom monitoring based on fine-tuned YOLOv11 model architectures. The study focuses on the student's attention detection; we have utilized the STEAM dataset that is publicly available. After preprocessing the dataset, we have trained YOLOv11 model on the preprocessed dataset. The study experimented with three architectures of YOLOv11 model, and the results of these experiments suggest that the model can be adopted in real-world classroom monitoring. The study utilized three publicly available datasets for analyzing student action, activity, and drawing behavior pattern analysis. The model achieved mAP50 of 96.8%. The detection

performance with the model achieved a significantly good detection performance with 2.58 M parameters, 1.2 ms of inference per image, and the model is trained for 300 epochs. In future work the model could be optimized, and a more advanced technique may improve the results. Moreover, in future work, large diverse datasets will improve the model's generalization. Despite the model demonstrating strong class detection performance, there are limitations in our work. The drawing dataset contains a large number of classes, with several classes having very few samples, which negatively impacted the overall detection performance. Future work could leverage larger and more balanced datasets to improve the model's learning capacity, enabling more accurate and robust predictions of a child's behavior.

Authors contributions
HF, XYC, SSM, NJ, SYW: Planning, data collection, and data analysis– original draft review & editing. All the authors read and approved the final manuscript.

Funding
This research received no funding.

Data availability
The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare no competing interest.

Received: 11 December 2025 / Accepted: 23 April 2026
Published online: 04 May 2026

References

1. Zhu Y, et al. A comprehensive study of deep video action recognition. arXiv 2020.
2. Murtiningrum S. Classroom interaction in English learning. Universitas Sanata Dharma Yogyakarta; 2009.
3. Anderson DH, Munk JAH, Young KR, Conley L, Caldarella P. Teaching organizational skills to promote academic achievement in behaviorally challenged students. Teach Except Child. 2008;40:6–13.
4. Caldarella P, Larsen RA, Williams L, Lulla SR, Wills HP. Behavior monitoring in the middle grades: evaluation of the classroom performance survey. RMLE Online. 2022;45:1–15.
5. Chen H, Zhou G, Jiang H. Student behavior detection in the classroom based on improved yolov8. Sensors. 2023;23:8385.
6. Lu M, Li D, Xu F. Recognition of students' abnormal behaviors in English learning and analysis of psychological stress based on deep learning. Front Psychol. 2022;13:1025304.
7. Khan A, Ghosh SK. Student performance analysis and prediction in classroom learning: a review of educational data mining studies. Educ Inf Technol. 2021;26:205–40.
8. Otto HE, Mandorli F. Parametric feature-based solid model deficiency identification to support learning outcomes assessment in cad education. Comput Aided Des Appl 2021;18.
9. Yongkang H, Meiyu L, Xiaoxiao W, Zheng C, Xiaowen C. Multi-person classroom behavior recognition in classroom teaching videos based on deep spatio-temporal residual convolutional neural network. Comput Appl. 2022;42:736–42.
10. Lin F-C, Ngo H-H, Dow C-R, Lam K-H, Le HL. Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection. Sensors. 2021;21:5314.
11. Pawar K, Attar V. Deep learning approaches for video-based anomalous activity detection. World Wide Web. 2019;22:571–601.
12. Saleem MH, Potgieter J, Arif KM. Automation in agriculture by machine and deep learning techniques: a review of recent developments. Precis Agric. 2021;22:2053–91.
13. Roa'a M, Aljazaery IA, Alaidi A. Automated cheating detection based on video surveillance in the examination classes. IJIM. 2022;16:125.
14. Ma L, et al. Improving yolov7 for large target classroom behavior recognition of teachers in smart classroom scenarios. Electronics. 2024;13:3726.
15. Aglaia AN, et al. Facial expression detection system for students in classroom learning process using Yolov7. Jurnal Nasional Pendidikan Teknik Informatika JANAPATI. 2024;13:546–60.
16. Sheng X, Li S, Chan S. Real-time classroom student behavior detection based on improved yolov8s. Sci Rep. 2025;15:14470.

17. Yu XG, Sun JY, He B, Zhuang JJ, Dai ZC. Design and implementation of automatic invigilation functions using the embedded technology. Proc Comput Sci. 2020;166:41–5.
18. Zaffar A, Jawad M, Shabbir M. A novel CNN-RNN model for e-cheating detection based on video surveillance. Univ Wah J Comput Sci. 2023;5:1–13.
19. Alshahrani A, Almatrafi MM, Mustafa JI, Albaqami LS, Aljabri RA. A children's psychological and mental health detection model by drawing analysis based on computer vision and deep learning. Eng Technol Appl Sci Res. 2024;14:15533–40.
20. Cao Q, Doma BT. Yolov8-based student behavior detection in the classroom: Introducing bar attention and shape IOU improvements to enhance detection capabilities. J Artif Intell Technol. 2026;6:42–51.
21. Pan H, Li Y, Zhao D. Recognizing human behaviors from surveillance videos using the SSD algorithm: H. Pan et al. J Supercomput. 2021;77:6852–70.
22. ABC. Steam-dataset dataset. Roboflow Universe. 2024. https://universe.roboflow.com/abc-jgsod/steam-dataset. Accessed 28 Feb 2026.
23. studentactivity. Scb-dataset-3 dataset. Roboflow Universe. 2024. https://universe.roboflow.com/studentactivity/scb-dataset-3. Accessed 27 Feb 2026
24. Drawing ODC. Children drawing object dataset. Roboflow Universe. 2025. https://universe.roboflow.com/object-detection-children-drawing/children-drawing-object. Accessed 27 Feb 2026
25. Xu J, Chen H, Xiao X, Zhao M, Liu B. Gesture object detection and recognition based on YOLOv11. Appl Comput Eng. 2025;133:81–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.