




Research Article

# Abnormal Behavior in Online Exam: Distance Learning Assessments Dataset

<sup>1</sup>Muhanad Abdul Elah Abbas Alkhalisy 

University of Information Technology and Communications  
Baghdad, Iraq

[Muhanad\\_alkhalisy@uoitc.edu.iq](mailto:Muhanad_alkhalisy@uoitc.edu.iq)

## ARTICLE INFO

### Article History

Received: 09/05/2025

Accepted: 21/06/2025

Published: 07/08/2025

This is an open-access article under the CC BY 4.0 license:

<http://creativecommons.org/licenses/by/4.0/>



## ABSTRACT

This paper presents a newly collected and highly relevant dataset on students' abnormal behavior in online exams. This dataset focuses on assisting research in building machine-learning models that allow for maintaining academic integrity during the era of online exams. Properly, more than 8,500 annotated images of normal and abnormal behaviors of students during remote examination are held in the dataset hosted at the Harvard Dataverse repository. The dataset has two versions: the original and the augmented. We utilize semantic segmentation and deep learning techniques in the applied data augmentation; this dataset provides a crucial foundation for developing and benchmarking intelligent proctoring systems. We evaluate the dataset using YOLO5 and our improved SPL-YOLO5 model, and the resulting mean average precision (mAP) is close to 1.0.

**Keywords:** Computer vision; behavioral analysis; online exam; student behavior; Deep Learning

## 1. INTRODUCTION

Global events such as the pandemic, like COVID-19 in 2020, prompted a swift shift to online education, revealing challenges in maintaining students' academic integrity during remote exams [1]. Many traditional proctoring methods, like human invigilators, cannot identify subtle or complex students' abnormal behaviors during an online exam. According to established academic integrity frameworks like those established by the International Centre for Academic Integrity ICAI, "abnormal behavior" in this context refers to any behavior during online assessments that deviates from standard exam protocols and could indicate academic dishonesty. Examples of such behavior include using a mobile phone, diverting one's attention, or making suspicious hand gestures. Machine learning and computer vision technologies provide excellent alternatives to proctoring, allowing real-time monitoring and automatic detection of suspicious activities [2]. This is, however, severely hindered by the lack of effective, high-quality datasets available for this particular task [3]. Although existing datasets capture affective states and levels of involvement, such as DAiSEE [4], they do not provide accurate annotations of anomalous test-taking behaviors. We present the "Students' Abnormal Behavior in Online Exam" dataset, a meticulously annotated image collection created to aid in developing and evaluating intelligent, real-time proctoring systems to fill this crucial gap.

## 2. RELATED WORK

However, affective state data for existing datasets, such as DAiSEE [4], is an example of an affective dataset that helps recognize engaged students, even though it cannot detect cases of mischievous behavior with proper labeling. Many prior datasets are restricted in their application to proctoring because they lack enough behavior, have errors in the labelling process, have low annotation accuracy, or have inadequate sample diversity. The high-resolution annotations for behavior types most closely associated with cheating annotations in our data cover behavior types, such as lateral gaze, mobile usage, and actions with the hands, which relate to cheating tackles these limitations. Therefore, it gives academics and researchers new tools for studying academic integrity online. Recent studies have used YOLOv5 [5] and the improved SPL-YOLOv5 model that successfully trained them to detect human behaviors such as mobile use, hand movements and looking away [6]. Semantic Segmentation and Pixel-Based transformation were used as data augmentation strategies that solved the class imbalance problem [7].

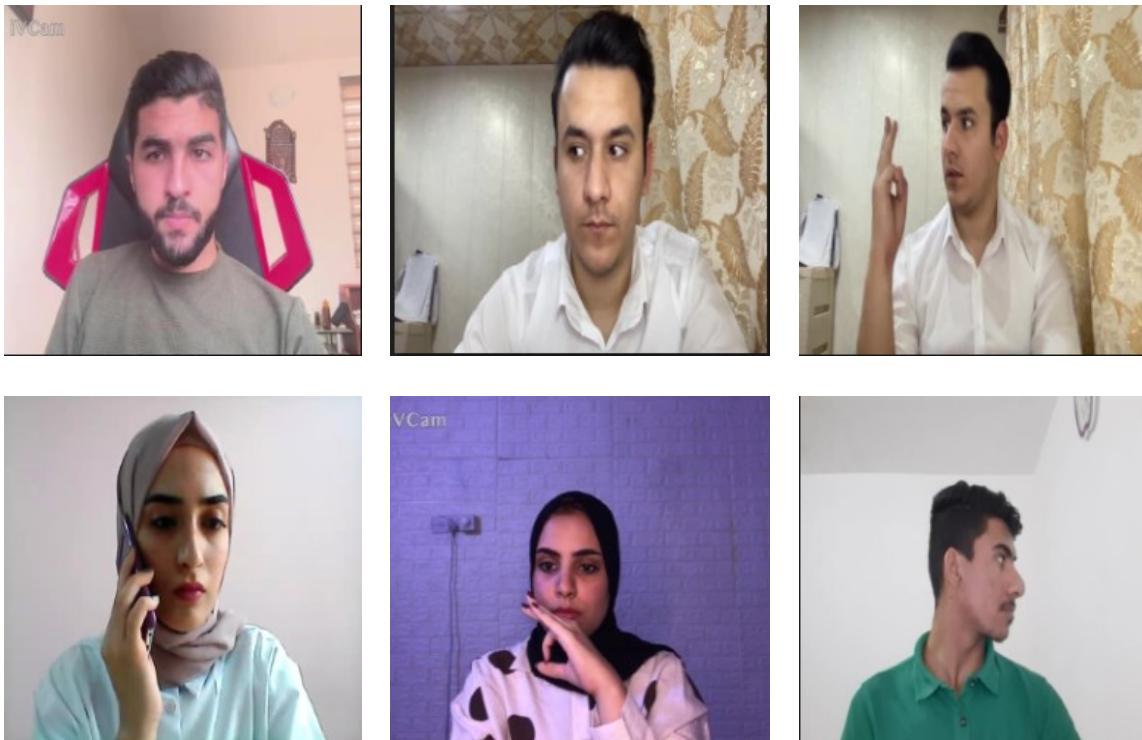
### 3. METHODS APPLIED

The most significant issue in jobs involving human action recognition is the collection of video information. Specifically, obtaining enough data, both in terms of number and quality.

#### 3.1 Data Collection

To develop the dataset, a webcam recorded forty-five videos at a resolution of  $1280 \times 720$ . A web application was designed to simulate an online examination environment, featuring video-capture functionalities for data collection.

The application evaluated students' understanding through tens of questions drawn from a database. Participants from various locations participated in the experiment, utilizing different lighting conditions, positions, places, and camera setups. They all belonged to the Informatics Management Systems ISM department at the University of Information Technology and Communications. Although this promotes uniformity, it also presents generalizability restrictions that should be addressed in future research by collecting data from several institutions. The captured data featured a participant completing an online test using a webcam, with each scenario lasting approximately 15 minutes, totaling nearly 6 hours of footage. The chosen duration strikes a balance between preserving a manageable volume of data for processing and annotation and capturing variances in natural behavior. Snapshots from one of the captured images are shown in Figure 1.



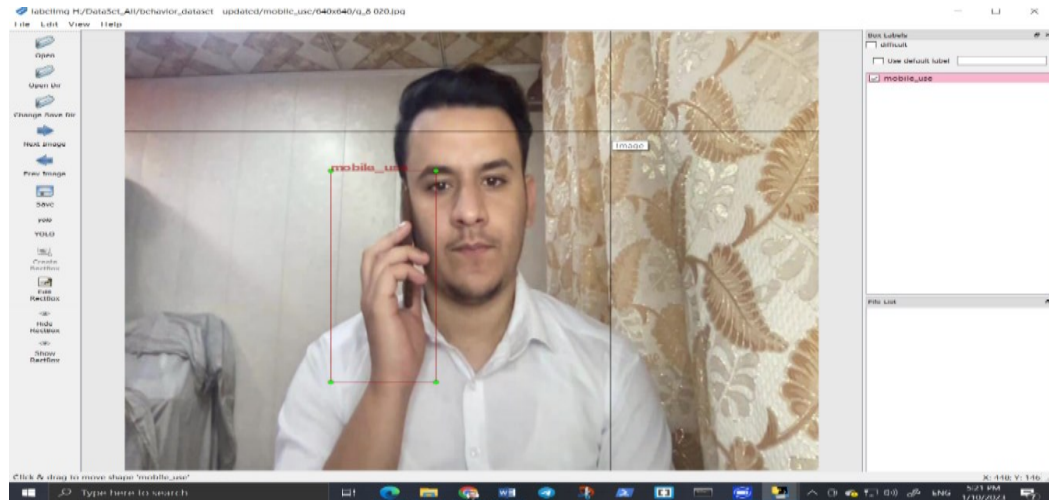
**Fig. 1:** Sample Dataset's Image Snaps

Images were extracted at fixed intervals from each captured video. One frame was selected from every group of 12 consecutive frames to provide a range of scenes. In total, 49500 images were obtained across five distinct classes. After manually filtering the collected images, 8,846 photos were retained. Images that were not clear or didn't show any distinguishable behavior were not included by a human quality filter. The inclusion criteria required clear visibility of at least one target behavior. Because of filtering, your data will be high-quality and include different actions valuable for your training and tests.

#### 3.2 Data Labelling

Manually assigning a class label to each selected frame proved to be quite challenging. Image annotation involves adding descriptive labels to images within a dataset, which helps in training models by conveying information about the content, location, and shape of the images. Labelling and MakeSense are popular tools in computer vision for

conducting annotation tasks. Accurate ground truth annotations are essential for training deep learning algorithms focused on object detection. Both annotation tools were utilized to manually label each instance of cheating. The outcomes of the ground truth data labelling are illustrated in Figure 2.



**Fig. 2.** Ground Truth Labelling

The dataset encompasses five distinct classes categorized according to specific behaviors: mobile usage, hand movement, eye movement, mouth open, and lateral gaze. The annotation process was conducted in accordance with the established YOLO Darknet guidelines format. To minimize subjectivity, multiple annotators were employed. Cohen's Kappa coefficient was used to track inter-annotator agreement, and an average score above 0.85 indicated strong agreement. High annotation reliability was ensured by using a consensus review to settle disagreements.

#### 4. DATASET DESCRIPTION

- **Source:** Harvard Dataverse ([DOI: 10.7910/DVN/WUWRAB](https://doi.org/10.7910/DVN/WUWRAB)).
- **Content:** 8,520 annotated images from 24 recorded online exam videos.
- **Categories:** Mobile phone usage, hand movement, eye movement, looking sideways.
- **Annotations:** Bounding boxes and behavior labels verified via manual annotation using LabelImg and MakeSense.ai.
- **Format:** JPEG images with JSON annotations, compatible with TensorFlow and PyTorch.
- **Resolution:** 640x480 pixels.

#### 5. DATA AUGMENTATION

By using semantic segmentation, objects were separated from the background. Random background replacement and pixel-based color augmentation were used to generate synthetic scenes, which were orders of magnitude more diverse. These augmentation techniques boosted the robustness of the trained models by more than 0.3% in detection precision; these methods also improved model robustness under varying lighting and background conditions. A class-specific performance analysis showed a 5–8% increase in recall for underrepresented classes like “MouthOpen” and “SideWatching,” validating the augmentation’s efficacy. Figure 3 shows some augmented image results.



**Fig. 3.** Results of augmented images

Table 1 presents a detailed comparison between the augmented and original datasets, focusing on the number of images in each class and the total number of images.

TABLE I: Dataset Details after Augmentation

Class Name	No. of Images (Original)	No. of Images (Augmented)
Mobile Using	1042	3450
Hand Move	1215	3600
Eye Movement	1150	3615
Side Watching	1132	3600
Mouth Open	900	3535
<b>Total</b>	<b>8846</b>	<b>18000</b>

Table 2 presents the precision, mean average precision (mAP), and training time of the system when trained on the original dataset for 50 epochs versus when trained on the augmented dataset.

TABLE II Performance of model training on original and augmented data

Dataset	mAP at 0.5	mAP at 0.95	Training Time (Min) for 50 Epochs
Original Dataset	0.55	0.57	58
Augmented Dataset	0.82	0.80	106

Table 3 compares the work based on the data-augmented methods used.

TABLE III COMPARISON OF THE WORK BASED ON THE AUGMENTED METHOD USED

Experiments	Model	Argumentation methods	mAP at 0.5	mAP at 0.95
Exp 1	YOLOv5	No data augmentation used	0.55	0.54
Exp 2	YOLOv5	Horizontal or vertical flip, hsv-hue, rescale, and hsv-saturation, mosaic	0.57	0.56
Exp 3	YOLOv5	Proposed augmentation method	0.82	0.83

## 6. APPLICATIONS

The dataset supports:

- Development of real-time behavior detection systems.
- Training the lightweight deep learning models for the academic integrity solutions.
- Evaluation benchmarks for AI-powered proctoring.
- Research on behaviors in remote assessments.



## 7. LIMITATIONS AND ETHICAL CONSIDERATIONS

While the dataset is diverse, many cultural variations of suspicious behaviour may not be captured. The data still contains only participants from a single university, which can affect the diversity when it comes to age and culture. In order to make the research more inclusive and useful to many, future participants should be collected from different institutions, geographic areas, and types of schooling. To use it ethically, we must adhere to data protection standards, such as GDPR, and have a transparent and fair model for deployment. The ethical aspect is important whenever implementing automatic surveillance tools. The following are the main issues:

- **User Privacy:** Before recording, people need to agree and be aware of how their data will be managed, following rules like those of GDPR.
- **Misidentification Risks:** False identification in noting unusual activity can cause harm to the students, as, for instance, they may be unfairly penalized in school. Ways to mitigate bias rely on confidence scores, using people to review the outcomes, and reviewing what can cause bias.
- **Model Explainability:** Many question whether trust and accountability are ensured using black-box deep learning models. If XAI is integrated, anyone looking at the information will likely find explanations for why specific actions are highlighted as unusual.

It is essential that users clearly understand the systems, there is an institutional review, and frequent monitoring is done to ensure they act reasonably.

## 8. EXPERIMENTAL RESULTS

The results achieved are remarkable when models are trained on this dataset. With the original YOLOv5 and our improved model, SPL-YOLOv5, the  $\text{map}@0.5$  and  $\text{map}@0.95$  reported high accuracy detection rates of approximately 0.83 to 0.97, respectively. These results demonstrate the effectiveness and efficiency of building a system using this dataset. Table 4 presents the detection results obtained by training the original YOLOv5 model on this dataset.

TABLE IV: YOLOV5 MODEL'S OUTCOMES

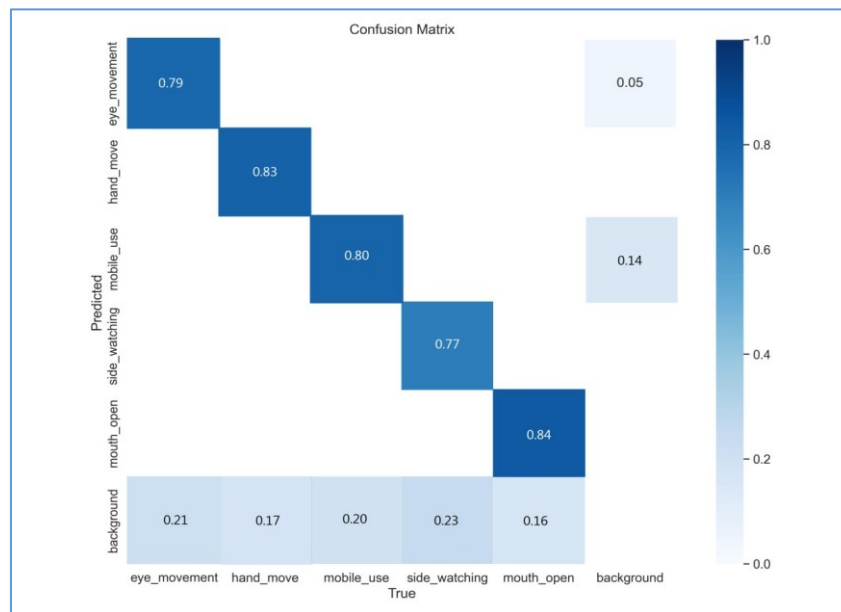
Class	Precision	mAP at 0.5	mAP at 0.95
Mobile Use	0.82	0.83	0.80
Hand Move	0.84	0.85	0.83
Eye Movement	0.84	0.85	0.79
Side Watching	0.78	0.79	0.77
Mouth Open	0.85	0.86	0.84
<b>Average for all classes</b>	<b>0.82</b>	<b>0.83</b>	<b>0.80</b>

Table 5 presents the detection results of training the improved SPL-YOLOv5 model on this dataset.

TABLE V: IMPROVED YOLOV5 MODEL'S OUTCOMES

Class	Precision	mAP at 0.5	mAP at 0.95
Mobile Use	0.97	0.96	0.93
Hand Move	0.96	0.97	0.94
Eye Movement	0.95	0.98	0.92
Side Watching	0.98	0.97	0.95
MouthOpen	0.97	0.98	0.94
<b>Average for all classes</b>	<b>0.96</b>	<b>0.97</b>	<b>0.93</b>

Figure 4 depicts the confusion matrix of the testing dataset after applying the data augmentation method, indicating a noticeable improvement in the results.



**Fig. 4** Validation Confusion Matrix After Applying the Data Augmentation Method

Figure 5 shows some visual scene detection results for the test parts of the dataset.



**Fig. 5.** Some visual scene detection results for the test parts of the dataset

## 9. CONCLUSION AND FUTURE WORK

The Students' Abnormal Behavior in Online Exams dataset represents a vital step towards ensuring the integrity of remote exams. This dataset comprises an extensive collection of annotated images that capture a wide variety of irregular behaviors, including mobile use, hand movements, and eye movement shifts, which are essential for training



ML models designed for the early detection of suspicious conduct in online examinations. The dataset's value lies in its high-quality, manually annotated labels, semantic segmentation-based augmentations, and its ready-to-use compatibility with all major frameworks. Experimental evaluations demonstrate that the dataset exhibits high detection accuracy, particularly when combined with an enhanced SPL-YOLOv5 model, achieving a mean average precision (MAP) of up to 0.97. Future work includes multimodal data integration (i.e., video sequences and audio), extending class labels to include other abnormal behaviors, improving inclusivity by expanding on the demographics of participants, and multi-institutional data gathering, and incorporation of model explainability tools to enhance trust and decrease the possibility of unfair decisions.

### Acknowledgment

We thank all data contributors, volunteers, the University of Information Technology and Communications, and the Harvard Dataverse team for their support and hosting services.

### References

- [1] F. F. Kharbat and A. S. Abu Daabes, "E-proctored exams during the COVID-19 pandemic: A close understanding," *Educ. Inf. Technol.*, vol. 26, no. 6, pp. 6589–6605, 2021, doi: 10.1007/s10639-021-10458-7.
- [2] K. Lee and M. Fanguy, "Online exam proctoring technologies: Educational innovation or deterioration?," *Br. J. Educ. Technol.*, vol. 53, no. 3, pp. 475–490, 2022, doi: 10.1111/bjet.13182.
- [3] L. Tang, T. Xie, Y. Yang, and H. Wang, "Classroom Behavior Detection Based on Improved YOLOv5 Algorithm Combining Multi-Scale Feature Fusion and Attention Mechanism," *Appl. Sci.*, vol. 12, no. 13, 2022, doi: 10.3390/app12136790.
- [4] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards User Engagement Recognition in the Wild," vol. 14, no. 8, pp. 1–12, 2016, [Online]. Available: <http://arxiv.org/abs/1609.01885>
- [5] M. A. E. Alkhalisy and S. H. Abid, "Abnormal Behavior Detection in Online Exams Using Deep Learning and Data Augmentation Techniques," *Int. J. online Biomed. Eng.*, vol. 19, no. 10, pp. 33–48, 2023, doi: 10.3991/ijoe.v19i10.39583.
- [6] M. A. E. Alkhalisy and S. H. Abid, "Students Behavior Detection Based on Improved YOLOv5 Algorithm Combining with CBAM Attention Mechanism," *Int. J. Intell. Eng. Syst.*, vol. 16, no. 5, pp. 473–487, 2023, doi: 10.22266/ijies2023.1031.41.
- [7] M. Abdul Elah Alkhalisy and S. Hameed Abid, "The Detection of Students' Abnormal Behavior in Online Exams Using Facial Landmarks in Conjunction with the YOLOv5 Models," *Iraqi J. Comput. Informatics*, vol. 49, no. 1, pp. 22–29, 2023, doi: 10.25195/ijci.v49i1.380.