Quality Improvement of an AI System for Determining Pass-Fail in the Fundamentals of Laparoscopic Surgery: Accuracy on a Cohort of New Users

Finn Kliewer, Yiran Huang, Advaith Bongu, Yunzhe Xue, Andrew Hu, Usman Roshan

Abstract

Surgical residents must pass the fundamentals of laparoscopy surgery test to proceed with their training. While simulation gives them an environment to practice outside the operating room, it lacks supervision. To fill this gap, we recently proposed an AI system that evaluates pass-fail in the fundamentals of laparoscopic surgery. In this quality improvement study, we sought to evaluate the model's accuracy when detecting a failure. To do this, we performed software testing on a cohort of high school students. The students were asked to conduct the essential peg transfer FLS task under the supervision of our AI system, which evaluates them in real-time as they perform the task. Out of 18 students, the system correctly predicted the student error in 13 cases. The model must catch up on student errors due to underlying model mispredictions in the remaining five. The model was not trained to handle such edge cases where it failed. These results show the potential of AI to make grading more fair, objective, and efficient.

1 Introduction

Surgical trainees preparing for a career in surgery must demonstrate proficiency in basic surgical skills. Evaluating these skills in the operating room requires considerable resources and increases the risk of surgical complications [1], [2]. Simulation has emerged as a cost-effective and accurate alternative [3–6].

Laparoscopic surgery, also known as minimally invasive surgery, involves performing operations through small incisions with the assistance of a camera and specialized instruments. This approach offers benefits such as reduced pain, shorter recovery times, and decreased risk of infection compared to traditional open surgery.

The Fundamentals of Laparoscopic Surgery (FLS) simulation kit is designed to train and evaluate students in laparoscopic surgery [7]. The first FLS task is to transfer six rings from one set of pegs to another set, picking up the ring with the left grasper, transferring it to the right grasper, and then placing it down on the goal peg. Once all rings are transferred to the right-side pegs, they must be moved back to the original left-side pegs. The requirements to pass are to make the transfer within 300 seconds and not drop a ring during the transfer.

The FLS simulation kit replicates the conditions of laparoscopic surgery by placing a webcam within a non-transparent box. This setup mimics the lack of depth perception experienced when operating via a camera and screen, a typical challenge trainees face during laparoscopic procedures. Trainees must adapt to this limitation while using two graspers (Maryland dissectors) to manipulate the objects, forcing them to develop their hand-eye coordination and spatial awareness.

Recently, our team proposed an AI system [8], [9] to evaluate automatically and grade students doing the peg transfer task using the FLS simulation kit. The current study builds upon this prior work, improving the existing model's accuracy and robustness in providing standardized, objective evaluations. Grading these FLS exams are very time-intensive and subject to variation between graders, which can lead to evaluation inconsistencies. Some graders consider a particular performance a failure, while others may not. Using an AI grading system can standardize the process. The proposed AI system functions by analyzing the video

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

feed from the webcam inside the simulation kit. Once completed, the AI system generates a report displaying metrics for each move and an overall pass/fail score. In this quality improvement study, we evaluate our AI system on a cohort of medical high school students. Students were given verbal rudimentary directions on the FLS task beforehand and asked to conduct it while our software ran. While they carried out the task, the system evaluated their performance in real-time with overlayed feedback and looked out for any errors. When an error occurred, the system informed the student and stopped the evaluation.

Of all 18 students who took the test, our system correctly detected errors in 13. In the remaining 5, the system encountered previously unseen scenarios, such as grasping a ring with both graspers, grabbing a ring in such a way that flung it out of bounds, and grabbing two rings simultaneously with one grasper. We expect the system to improve these predictions as we retrain with more data.

2 Methods

Objective

In this study, our object was to test the accuracy of our model in detecting failures and errors during the peg transfer FLS task. Specifically, we sought to determine the accuracy with which our model could accurately flag these failure conditions.

Data

We collected videos of 18 high school students, ages 16-18, performing FLS training on a school field trip at the Robert Wood Johnson Hospital in New Brunswick, New Jersey, USA. The aforementioned high school students had zero previous FLS training. Each student performed the task once, with collected videos ranging between 30 seconds to 2 minutes. Unexperienced high schoolers were perfect candidates for this study because we expected them to make many mistakes, which would test the limits of our system. All students who visited our booth during the field trip were included in the study, as there were no exclusion criteria. Since we were testing the AI's ability to detect mistakes, we needed a group of students likely to make numerous and unpredictable errors. Below are screen snapshots of the FLS videos showing the left grasper picking a ring in (Figure 1(a)), a ring transfer between left and right graspers in (b), and placing a ring on the peg in (c). For each video, we determined a ground truth of whether a failure was present using the rules set by the medical team at Robert Wood Johnson Hospital.

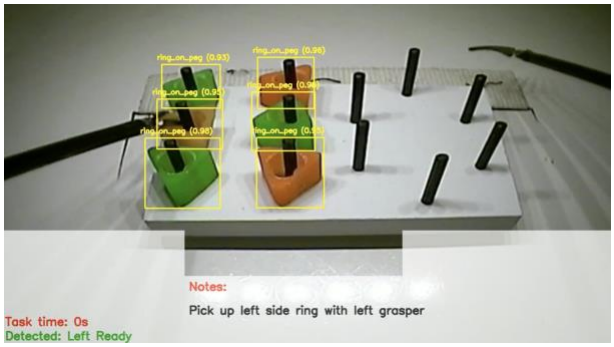
Rules for Determining Pass or Fail

1) Timing starts when the first ring is touched on the peg and the last ring is placed on the peg. Total time should not exceed 300s.
2) The student fails the task if the ring is dropped outside of the FLS box.
3) We look for six successful consecutive transfers from the left set of pegs to the right set. A single completed task is defined as grasping the ring with the left grasper, picking it from the peg, transferring it to the right grasper, grasping the ring with the right grasper, and finally placing it on the peg on the right side of the board. We summarize this as grasp1-pick-transfer-grasp2-place. If six such transfers have occurred without a drop outside the box, continue to the next step.
4) If a ring is dropped inside the box, it must be picked with the same grasper - in this case, grasper1. After picking up the correct grasper, the student can continue. However, if they pick with the wrong grasper, then this is a failure.
5) We look for six successful transfers from the right set of pegs to the left set. We follow the same rules as we did above when transferring from left to right, except that a successful transfer is now given by grasp2-pick-transfer- grasp1-place.
6) As noted above, if a ring drops inside the box, it must be picked with the same grasper - in this case, grasper2. After picking it up with the correct grasper, the student can continue, but picking with the wrong grasper is considered a

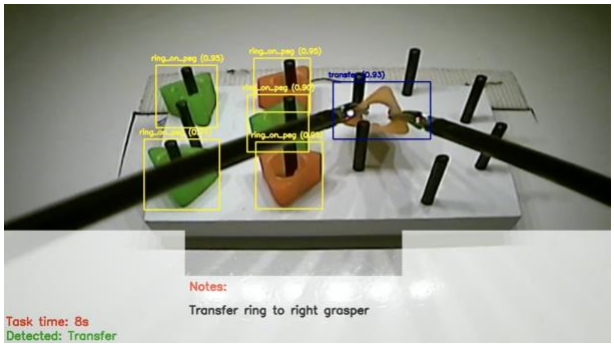
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

failure.

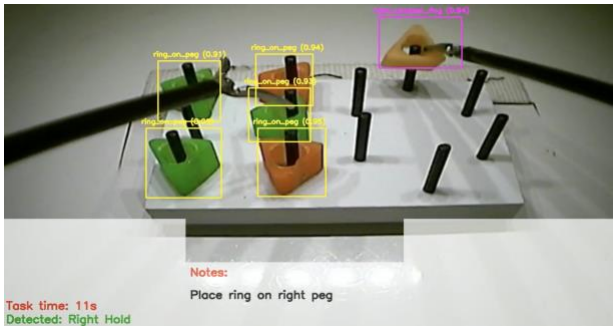
7) If the six transfers have taken place without a drop outside the box and the total time is below 300 seconds, then the student has passed.



(a) Pick a ring



(b) Transfer ring



(c) Place ring on peg

Figure 1: Examples of picking a ring, transferring a ring, and placing one on the peg

Based on the above rules, we determined whether each video had a ground truth error. In Table I, we list the videos that we collected.

AI System

Our overall approach is to pass a video through our trained YOLO (You Only Look Once) model, a machine learning model developed and fine-tuned using a training dataset to efficiently identify objects in images or videos—in this case, rings and pegs. Our training dataset consisted of annotated video frames where the rings and pegs were marked. These annotations teach the YOLO model to recognize objects associated with our training data, such as rings and pegs. The size of our training data set was $n = 60$ videos. The YOLO model recognizes objects by dividing the image into a grid and predicting bounding boxes and confidence scores for each grid section to help determine where our rings and pegs are. We then apply logic, which involves using predefined rules and conditions based on the task requirements, to analyze the model's outputs. Specifically, we check whether the actions in the video follow the rules outlined above in Subsection II-C to determine if a resident passed or failed the FLS peg transfer task. See Figure 2 for an overview of our system.

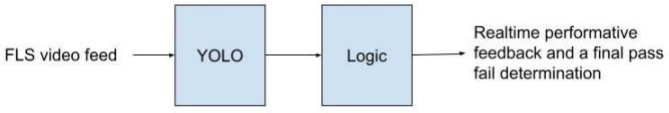


Figure 2: Overview of AI System

3 Results

Overall Metrics

In Table I below, we show our AI system's ground truth and evaluation of each test video. Our model correctly identifies the failures in 13 of the 18 videos recorded from the high school students' field trip - thus giving an accuracy of 72.2%. In the three cases, it gives a wrong determination due to the system encountering an edge case it was not trained to handle (Figure 3) The other two mispredictions are due to incorrect bounding box predictions made by the underlying YOLO model (Figure 4). In Figure 4(a), the model incorrectly identifies two pegs as only one peg,

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

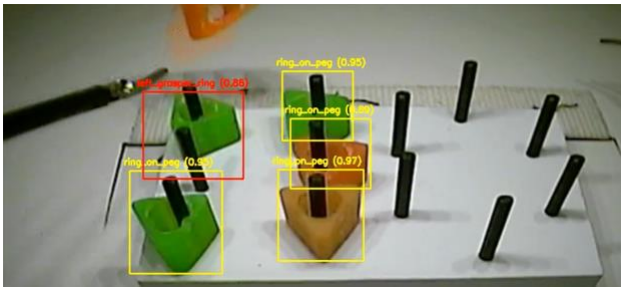
and in Figure 4(b), the ring dropped outside the box labeled as "drop in ring". This procedure is expected to improve as we increase the training set size (the dataset used to train the machine learning model).

4 Discussion

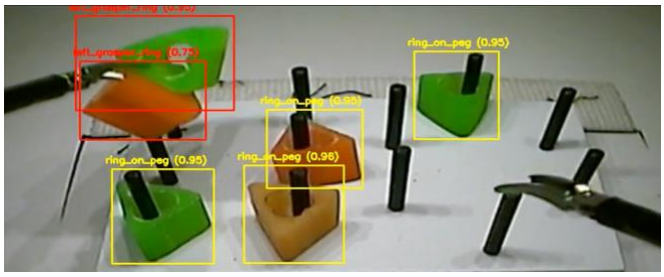
Out of 18 test runs, our model successfully detected errors in 13 instances. The five errors it failed to detect were unconventional and unlikely to be encountered in typical scenarios when surgical trainees perform the task. Unconventional errors refer to less systematic mistakes that go against the rules of the task, which are more likely to occur in individuals with no prior experience. This indicates that our AI model can effectively auto-grad the Fundamentals of Laparoscopic Surgery (FLS) exam.

Video	Failure Present	Model Predicted Correctly
1	Yes	No
2	Yes	Yes
3	Yes	No
4	Yes	Yes
5	Yes	No
6	Yes	Yes
7	Yes	No
8	Yes	Yes
9	Yes	Yes
10	Yes	Yes
11	Yes	Yes
12	Yes	No
13	Yes	Yes
14	Yes	Yes
15	Yes	Yes
16	Yes	Yes
17	Yes	Yes
18	Yes	Yes

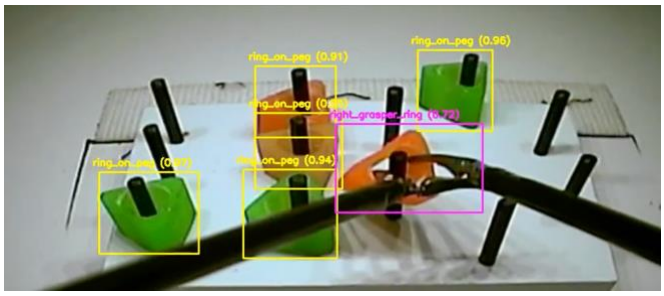
Table 1: Medical resident level, ground truth, and prediction as given by the AI System



(a) Ring being flung out of bounds

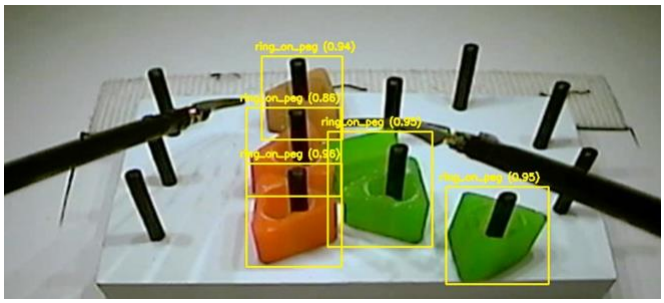


(b) Picking up two rings with one grasper

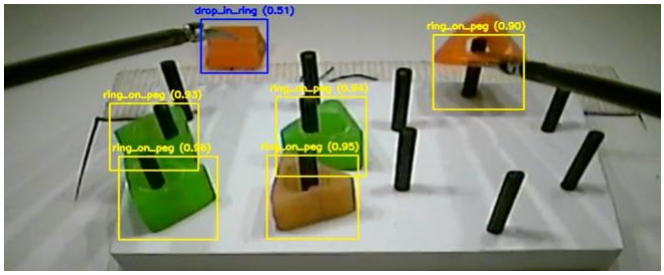


(c) Picking up a ring with two graspers

Figure 3: Examples of the edge cases of the system was not trained to detect



(a) Model draws one bounding box around two pegs



(b) Model predicts a ring outside the box as "drop in ring"

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Figure 4: Examples of the system not marking a failure due to incorrect bounding box predictions

Edge Cases

In our testing, we encountered three unforeseen scenarios, including students grasping a ring with both graspers, pinching a ring leading to it being inadvertently flung out of bounds, or attempting to pick up two rings simultaneously with one grasper. We identified the cause of why the model incorrectly predicted each of these failures. These cases are considered edge cases since these failures typically do not happen when surgical trainees perform the task. In Figure 3(a), a ring flung out of bounds in the top left corner. The model did not catch this failure because the sample rate was too low and did not catch the ring flying out of bounds in time. We can fix this by adding a check to ensure all six rings are always on screen or increasing the sample rate. In Figure 3(b), we can see the grasper picking up two pegs simultaneously, and in Figure 3(c), two graspers picked up one ring. We did not anticipate such behavior when training the model. By retraining the system with additional data that includes these edge cases, as well as adding logic rules, we can improve its performance in identifying and marking such edge cases in future assessments.

Future Improvements

Our system is the first to give a fully automated pass-or-fail outcome of the FLS task with real-time performative feedback. These scenarios are considered edge cases, which refer to rare or unexpected situations outside routine procedures. These cases are edge cases because these failures did not happen when the surgical trainees performed the task. To improve this system in the future, we will use the videos collected from this study to train the model in these specific edge cases. We will also add additional training data from resident surgeons to improve the underlying YOLO model and counter mispredictions. Another significant improvement on our agenda is adding more precise individualized directions and instructions overlayed on the screen to lower the failure rate. While this may not directly improve the accuracy of our model when predicting failures, this will help our system provide more effective training for resident surgeons when studying for the FLS exam.

Ethics

Ethical concerns and potential future consequences must be addressed when considering implementing an AI system to grade FLS exams. The use of AI in this context raises questions about accountability; specifically, who is responsible for errors in grading—the researchers, the institution, or the AI system itself? While the AI system will provide significant time savings and standardize grading to eliminate variations between human graders, it is crucial to maintain transparency in grading decisions. It will be essential to monitor its impact on training and certification processes to ensure it enhances, rather than hinders, the development of skilled surgeons.

5 References

[1] M. Bridges and D. L. Diamond, "The financial impact of teaching surgical residents in the operating room," *The American Journal of Surgery*, vol. 177, no. 1, pp. 28–32, 1999.

[2] R. W. Allen, M. Pruitt, and K. M. Taaffe, "Effect of resident involvement on operative time and operating room staffing costs," *Journal of Surgical Education*, vol. 73, no. 6, pp. 979–985, 2016.

[3] V. N. Palter and T. P. Grantcharov, "Individualized deliberate practice on a virtual reality simulator improves technical performance of surgical novices in the operating room," *Journal of the American College of Surgeons*, vol. 213, no. 3, p. S126, 2011.

[4] S. R. Dawe, G. Pena, J. A. Windsor, J. Broeders, P. C. Cregan, P. J. Hewett, and G. J. Maddern, "Systematic review of skills transfer after surgical simulation-based training," *Journal of British Surgery*, vol. 101, no. 9, pp. 1063–1076, 2014.

[5] B. Zendejas, D. A. Cook, J. Bingener, M. Huebner, W. F. Dunn, M. G. Sarr, and D. R. Farley, "Simulation-based mastery learning improves patient outcomes in laparoscopic inguinal hernia repair: a randomized

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

controlled trial," *Annals of surgery*, vol. 254, no. 3, pp. 502–511, 2011.

[6] T. Cox, N. Seymour, and D. Stefanidis, "Moving the needle: simulation's impact on patient outcomes," *Surgical Clinics*, vol. 95, no. 4, pp. 827– 838, 2015.

[7] N. J. Soper and G. M. Fried, "The fundamentals of laparoscopic surgery: its time has come," *Bull Am Coll Surg*, vol. 93, no. 9, pp. 30–32, 2008.

[8] Y. Xue, A. Hu, R. Muralidhar, J. W. Ady, A. Bongu, and U. Roshan, "An ai system for evaluating pass fail in fundamentals of laparoscopic surgery from live video in realtime with performative feedback," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, pp. 4167–4171.

[9] Y. Xue, O. Eletta, J. W. Ady, N. M. Patel, A. Bongu, and U. Roshan, "A cascaded neural network system for rating student performance in surgical knot tying simulation," in *Proceedings of the 11th IEEE International Conference on Healthcare Informatics (ICHI)*, 2023.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Finn Kliewer graduated from Rutgers University in 2024 with a degree in Computer Science and a minor in Mathematics. During his time at Rutgers, he developed a strong interest in artificial intelligence and its real-world applications, especially in areas that combine technology and human impact. Outside of academics, Finn enjoys hiking, making music, and exploring creative projects that bring together his technical and artistic interests. He is excited to continue learning and working on projects that use AI to solve meaningful problems.

Finn can be contacted at flk16@rutgers.edu.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.