Providing Feedback for Surgical Training and Assessment using Artificial Neural Networks

Students: Karah Martin and Annabelle Shen / **Research Mentors:** Erim Yanik, Ph.D. and Suvranu De, Sc.D.
College of Medicine, College of Arts and Sciences, FAMU-FSU College of Engineering

Abstract

Problem Statement: Up to 100,000 deaths annually from preventable surgical errors due to subjective, time-consuming training and assessments.

Proposed Solution: Development of objective, time-efficient, automated **Artificial Intelligence (AI)** techniques.

Methodology:

- ✓ Developed a surgical rubric for consistent, structured feedback from experts (surgeons) with both categorical and open-ended responses.
- ✓ Collected high-quality laparoscopic suturing videos under the Fundamentals of Laparoscopic Surgery (FLS) program.
- ✓ Extracted tool motion sequences from videos to predict surgeons' multiple-choice feedback using an AI model.
- ✓ Demonstrated a Proof of Concept (PoC) validating the methodology for predicting multiple-choice feedback.

Future plans: Aim to create an open-source Large Learning Model (LLM) that uses open-ended feedback for personalized feedback per trial without a predefined categories.

Introduction

Importance of Automation: Effective surgical skill evaluation is critical for success, yet current mentor-apprenticeship methods are subjective and inefficient, leading to poor reliability[3,4]. **AI** models offer a solution to enhance consistency and efficiency[4].

Challenges for AI Implementation: For AI to yield reliable outcomes, it requires large-scale, high-quality input data and consistent labels (surgical feedback). However, the scarcity of public data and the lack of standardized collection methods for surgical feedback pose significant challenges. While sensor-based kinematic data is commonly used, it's cumbersome and may interfere with surgeries. Alternatively, surgical videos provide a scalable and less intrusive data source, with tool motion tracking proving to correlate well with surgical skill.

Study Approach: This study focused on laparoscopic suturing, collecting videos to extract tool motion sequences and developing a rubric for streamlined, high-quality feedback as labels. Collaborating with surgeons from the College of Medicine and Tallahassee Memorial Hospital, feedback collection is ongoing due to the job's sensitive nature. The methodology's validity was tested on the JIGSAWS dataset, with results detailed in subsequent sections.

Dataset Characteristics

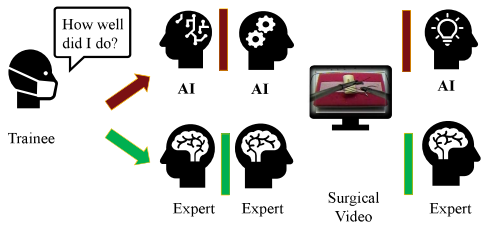
The proposed dataset for our study is Laparoscopic Suturing within the context of FLS program. The task involved suturing a Penrose drain as provided in **Figure 1a.** Tasks were performed by a group of 18 medical professionals, comprising 8 residents (5 males and 3 females) and 10 surgeons (5 males and 5 females), with an average age of 31 years and a standard deviation of 7.9, participated in a total of 63 trials.

Proof of Concept

Methods

Given the high-stakes work environment, we could not collect the feedback from surgeons on time. To provide a Proof of Concept (PoC), we decided to use the prominent publicly available dataset – JIGSAWS. JIGSAWS contains categorical feedback in 5 categories based on the Likert Scale for 6 questions. For details, please check the Methods section. JIGSAWS, however, does not contain open-ended feedback. Hence, we provided our PoC for the categorical responses only. We used the same methodology as detailed under the Methods section, i.e., extract tool locations --> establish tool motion sequences --> use an **AI** model to predict categorical scores via these sequence. We used the current state-of-the-art surgical skill assessment model – **the VBA-Net -** as our **AI** model.

Results



1. Unnecessary force on tissue. **2.** Excellent suture control. **3.** Efficient time/motion. **4.** Obviously planned course of operation with efficient transition between moves. **5-6.** Very Competent performance and final product quality.

1. Careful tissue handling. **2.** Majority of the knots placed correctly. **3.** Efficient time/motion. **4.** Demonstrated some forward planning. **5-6.** Competent performance and final product quality.

Table 1. Predictive performance achieved for each category. The category numbers follow the same order as the categories detailed in Methods section.

Category	1	2	3	4	5	6
Predictive Performance	0.28	0.46	0.61	0.48	0.40	0.43

Discussion

- ✓ The results show that our preliminary analysis the provide a PoC yields promising results on the publicly available JIGSAWS dataset.
- ✓ Our AI achieved the best predictive performance in generating feedback on the time and motion of the operation (0.61) while generated promising results for predicting the right feedback category for tool usage (Category 2), flow of operation (Category 4), and the overall assessments of the operation and the product quality (Categories 5-6). The AI, however, achieved a poor predictive performance when it comes to tissue handling. This can be attributed to the fact that the dataset uses a synthetic tissue and not a real one.
- ✓ Overall, while there is a significant work ahead to optimize our AI, the results are promising enough to be considered as PoC.

Conclusion

- ✓ This study proposes a framework for advancing the evaluation of technical proficiency in surgical procedures and fostering a more efficient patient care experience. Specifically, we developed a rubric that is tailored towards both categorical and open-ended feedback.
- ✓ We showed via our PoC that our pipeline has the potential to provide categorical feedback to the user with a promising accuracy.
- ✓ Our efforts are ongoing towards data collection, and we will continue to contribute to this project after the official deadline set by the program.
- ✓ The research laboratory that we are a part of is currently developing the necessary **LLM** technologies and our rubric will provide the valuable data to achieve this technology to provide open-ended personalized feedback.
- ✓ **For references and supplementary materials please scan the QR code (Top right corner).**

Methods

Rubric Development

To seek systematic and structured feedback from expert surgeons on the suturing dataset, we have developed a rubric based on the modified Objective Structured Assessment of Technical Skills (OSATS) rubric. We also developed demographic collection and consent documentation adhering to the ethics and safety regulations. Aiming to develop **AIs** that can provide both categorical and open-ended feedback, our rubric queried both multiple-choice with Likert-Scale responses and open-ended questions in the following 6 categories:

1. "Explain how the tissue was respected during the presentation"
2. "In terms of application, how well were the suture, needle, and grasper utilized?"
3. "What are your opinions on time and motion of the operation?"
4. "In what ways did the flow of operation compare to a standard procedure"
5. "Please describe the quality of the final product observed"
6. "Overall, how well was the suturing demonstration performed".

Artificial Intelligence Modeling

In this study, we propose to develop an **AI pipeline** that can process the surgical suturing videos and generate feedback based on the rubric we use to collect informative data from the surgeons. Notably, we propose two distinct **AI pipelines.** First is the traditional approach where an **instance segmentation network** is used to extract tool locations and another **AI** uses this information to generate the categorical feedback as seen in **Figure 1a**. The second is to use **Large-Language Models (LLMs)** such as **Chat-GPT** or it's open-source variants, e.g., **GPT-3, Mistral, LLAMA**, to directly provide open-ended and personalized feedback as illustrated in **Figure 1b**.

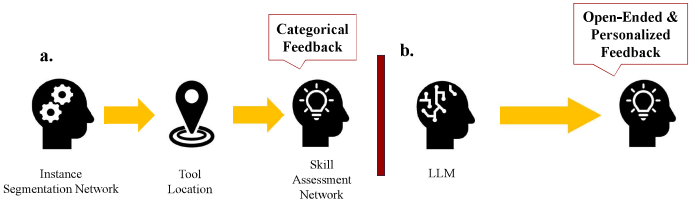


Figure 1 | a. The traditional approach. **b.** The LLM approach.

In the traditional approach, we first developed an instance segmentation model, namely Mask Region-Based Convolutional Neural Network[7] (Mask R-CNN) to extract tool locations for each video to establish tool motion sequences. For this we annotated each tool in several hundred randomly selected frames individually as depicted in **Figure 2b**. Once the annotations were completed Mask R-CNN was trained on these frames to learn to differentiate the tool locations from its background **(Figure 2c)**.

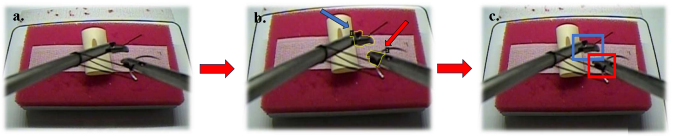


Figure 2 | a. The suturing task. **b.** The suturing frames are being annotated by using an Image Annotator. **c.** The bounding boxes generated by the Mask R-CNN.

When it comes to LLM training however, it is not feasible for small organization and/or research laboratories to muster enough computational power to effectively train an LLM. For instance, with one common GPU, the training from scratch would take more than 30 years. Thus, we are currently investigating the open-source LLMs and their benchmark performance with the collaboration of other researchers in our research laboratory.