


Original Article

Open Access



Automatic assessment of robotic suturing utilizing computer vision in a dry-lab simulation

Sarah Choksi^{1,2} , Sanjeev Narasimhan³, Mattia Ballo², Mehmet Turkcan⁴, Yiran Hu⁴, Chengbo Zang⁴, Alex Farrell², Brianna King², Jeffrey Nussbaum², Adin Reisner², Zoran Kostic⁴, Giovanni Taffurelli⁵, Filippo Filicori^{2,6}

¹Department of General Surgery, Albany Medical Center, Albany, NY 12208, USA.

²Department of Surgery, Northwell Health, New Hyde Park, NY 11042, USA.

³Department of Computer Science, Columbia University, New York, NY 10027, USA.

⁴Department of Electrical Engineering, Columbia University, New York, NY 10027, USA.

⁵Department of Surgery, Ospedale S. Maria delle Croci, AUSL Romagna, Ravenna 48121, Italy.

⁶Zucker School of Medicine at Hofstra/Northwell Health, Hempstead, NY 11549, USA.

Correspondence to: Dr. Sarah Choksi, Department of General Surgery, Albany Medical Center, 43 New Scotland, Albany, NY 12208, USA. E-mail: Sarah.choksi@outlook.com

How to cite this article: Choksi S, Narasimhan S, Ballo M, Turkcan M, Hu Y, Zang C, Farrell A, King B, Nussbaum J, Reisner A, Kostic Z, Taffurelli G, Filicori F. Automatic assessment of robotic suturing utilizing computer vision in a dry-lab simulation. *Art Int Surg*. 2025;5:160-9. <https://dx.doi.org/10.20517/ais.2024.84>

Received: 29 Sep 2024 **First Decision:** 9 Jan 2025 **Revised:** 19 Feb 2025 **Accepted:** 3 Mar 2025 **Published:** 1 Apr 2025

Academic Editors: Eyad Elyan, Thomas Schnellendorfer **Copy Editor:** Ting-Ting Hu **Production Editor:** Ting-Ting Hu

Abstract

Aim: Automated surgical skill assessment is poised to become an invaluable asset in surgical residency training. In our study, we aimed to create deep learning (DL) computer vision artificial intelligence (AI) models capable of automatically assessing trainee performance and determining proficiency on robotic suturing tasks.

Methods: Participants performed two robotic suturing tasks on a bench-top model created by our lab. Videos were recorded of each surgeon performing a backhand suturing task and a railroad suturing task at 30 frames per second (FPS) and downsampled to 15 FPS for the study. Each video was segmented into four sub-stitch phases: needle positioning, targeting, driving, and withdrawal. Each sub-stitch was annotated with a binary technical score (ideal or non-ideal), reflecting the operator's skill while performing the suturing action. For DL analysis, 16-frame overlapping clips were sampled from the videos with a stride of 1. To extract the features useful for classification, two pretrained Video Swin Transformer models were fine-tuned using these clips: one to classify the sub-stitch phase and another to predict the technical score. The model outputs were then combined and used to train a Random Forest Classifier to predict the surgeon's proficiency level.



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



Results: A total of 102 videos from 27 surgeons were evaluated using 3-fold cross-validation, 51 videos for the backhand suturing task and 51 videos for the railroad suturing task. Performance was assessed on sub-stitch classification accuracy, technical score accuracy, and surgeon proficiency prediction. The clip-based Video Swin Transformer models achieved an average classification accuracy of 70.23% for sub-stitch classification and 68.4% for technical score prediction on the test folds. Combining the model outputs, the Random Forest Classifier achieved an average accuracy of 66.7% in predicting surgeon proficiency.

Conclusion: This study shows the feasibility of creating a DL-based automatic assessment tool for robotic-assisted surgery. Using machine learning models, we predicted the proficiency level of a surgeon with 66.7% accuracy. Our dry lab model proposes a standardized training and assessment tool for suturing tasks using computer vision.

Keywords: Automatic surgical skill assessment, computer vision, surgical education, simulation

INTRODUCTION

Deep learning (DL) models, particularly those in Computer vision (CV), have rapidly advanced over the last five years. CV, a form of artificial intelligence (AI), enables machines to recognize and interpret images using DL algorithms. In recent years, CV models have been increasingly applied in the healthcare field. Given the large amount of video-based data, minimally invasive surgery remains an apt field for applying these CV models. In the last few years, CV in minimally invasive surgery has already been able to achieve task segmentation, object detection, and gesture recognition^[1-4].

Technical skill assessment remains integral to surgical training. Surgical performance in the operating room is key to good patient outcomes^[5]. Current technical skill evaluation such as video-based assessment remains time-consuming and provides subjective, sometimes unactionable feedback. Due to this, constructing a framework for automated skills assessment is of the utmost importance. In open surgery, other methods, such as tracking hand movement, have also been utilized for automatic assessment. Grewal *et al.* utilized an inertial measurement unit to collect data on hand movements to automatically assess surgical skills, while Azari *et al.* assessed surgical skills using CV on hand movements^[6,7].

Robot-assisted surgery is becoming widespread. However, currently, there is no validated, universal robot training curriculum or assessment for surgical trainees. Fundamentals of laparoscopic surgery (FLS) has been created for the training and assessment of surgical residents in laparoscopy. Residents are required to pass the FLS assessment before being board-eligible^[8]. Accreditation for robotic surgery, however, lacks standardization and is based largely on case experience, leading to a large amount of variability in robotic training. Multiple international studies have made attempts to advance robotic curriculums; however, there is still a need for standardization, especially with rapidly developing technologies^[9-12]. Past studies have worked on creating automatic assessment algorithms using CV for FLS. Lazar *et al.* utilized CV to differentiate between experts and novices performing the Peg Transfer task on the FLS trainer^[13], while Islam *et al.* designed a video-based system capable of providing task-specific feedback during the FLS task^[14].

However, for robotic-assisted surgery, limited automatic assessment tools exist. Ma *et al.* have been able to create an algorithm to automatically provide feedback for robotic suturing^[15,16]. These studies are the first to provide automatic assessment and feedback of robotic suturing in simulation and dry lab models for vesicourethral anastomosis.

In our study, we also aim to design a dry lab model for basic robotic suturing skills and create DL CV models capable of automatically assessing the performance of a trainee on suturing tasks. We further aim to determine if the participants are proficient at robotic suturing or need more practice.

METHODS

Study design and participants

Twenty-seven surgeons were recorded while completing two repetitions of two robotic tasks, backhand suturing and railroad suturing, on a bench-top model created by our lab. The bench top model consisted of artificial skin padded by packaging foam taped to a wooden block, as shown in [Figure 1](#). This was placed inside a robotic simulation abdominal cavity.

The railroad suturing task consisted of performing a running stitch by driving the needle from one side to the opposite of the wound, and then re-entering next to where it exited. The backhand suturing exercise involves performing a continuous stitch by guiding the needle from the side closest to the operator to the opposite side.

Videos were recorded at 30 frames per second (FPS) and downsampled to 15 FPS for the study. This prospective cohort study was approved by the Institutional Review Board of Northwell Health (IRB 23-069).

Video segmentation and data labeling

Video of each suturing task was broken down into a sequence of four sub-stitch phases: needle positioning, needle targeting, needle driving, and needle withdrawal [[Figure 1](#)]. Using Encord (Cord Technologies Limited, London, UK), every video was first temporally labeled into the four sub-stitch phases and then each sub-stitch was annotated with a binary technical score (ideal or non-ideal) based on a previously validated model^[15,17]. The ideal/non-ideal classification reflects the operator's skill while performing the suturing action. The annotators were all surgical residents and trained by a senior surgical resident. The annotators and engineers were blinded to the participants and experience levels when performing the annotations and building the model. Sub-stitch annotations (labels) were mapped into frame annotations.

CV model

Our proposed system leverages a multi-model approach to surgical skill assessment. We employ two distinct video DL models that are trained on overlapping 16-frame clips with a stride of 1 extracted from the videos: the first model classifies the clip into one of the four sub-stitch phases or a background (no action) class, while the second model classifies the clip into a binary technical score. These models use the Video Swin Transformer architecture to capture spatiotemporal features within the surgical workflow. To generate a comprehensive skill assessment, the individual model predictions are aggregated and fed into a Random Forest Classifier, which produces a final classification of trainee skill level, categorized as either "Proficient" or "Trainee". Proficient in suturing was based on a case experience of 50 robotic cases. This was then confirmed by a video review by two minimally invasive fellowship-trained attendings. This multi-faceted approach aims to provide a robust and objective evaluation of surgical competency in robotic surgery training [[Figure 2](#)].

At 15 fps, 16 frames correspond to 1.067 s of video time, which constitutes a very short period for the models to identify what suturing sub-stitch is taking place, and if any mistakes are made during the sub-stitch. Rather than naively increasing the clip size, which results in a larger memory footprint, we introduced a dilation (frame skip) of 15 frames between consecutive frames of the clip, resulting in an effective clip length of ~17 s. We employ dilation in both train and test phases, which improved sub-stitch

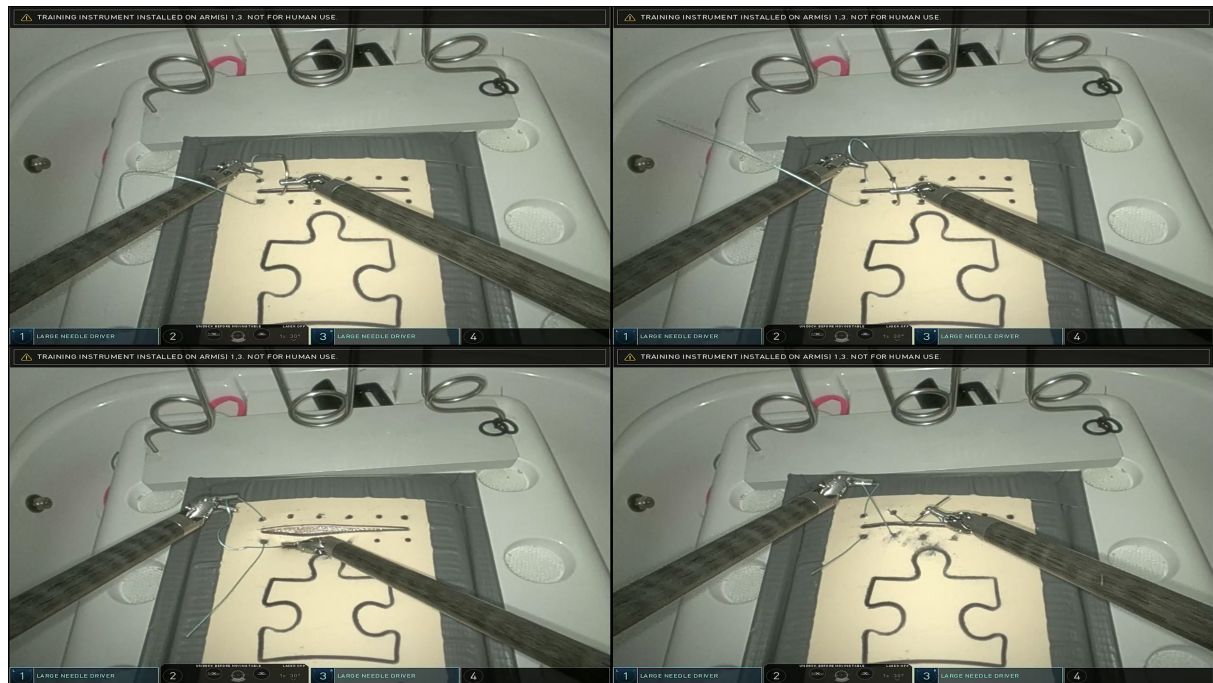


Figure 1. (Top-left to Bottom-right) Example frames for Needle Positioning, Needle Targeting, Needle Driving, and Needle Withdrawal sub-stitch actions from the backhand suturing task.

classification accuracy by ~11% and technical score prediction by ~5%.

RESULTS

Twenty-seven surgeons participated in the suturing tasks [Table 1]. The surgeons ranged from post graduate year (PGY) 1 to attendings with greater than 10 years' experience. The average robotic case experience was 50 cases +/- 156 cases. A total of 102 videos, consisting of 51 videos for the backhand suturing task and 51 videos for the railroad suturing task, spanning 891,384 frames, were evaluated. A total of 862,038 frames were annotated. We employed 3-fold cross-validation across the surgeons and averaged the results from the held-out surgeons across the folds. Performance was assessed on sub-stitch classification accuracy, technical score accuracy, and surgeon proficiency prediction. The clip-based Video Swin Transformer models achieved an average accuracy of 70.23% for sub-stitch classification and 68.4% for technical score prediction on the test folds [Table 2]. The confusion matrix for sub-stitch classification and technical score prediction across all videos is presented in Figure 3, with the darker cells representing more samples, labeled by their proportion within the dataset. Combining the model outputs, the Random Forest Classifier achieved an average accuracy of 66.7% in predicting surgeon proficiency [Figure 4]. The importance of input features to the Random Forest Classifier was analyzed using mean decrease in impurity (MDI), revealing that the Needle Driving feature with a technical score of "Ideal" is the most significant in determining surgeon proficiency [Figure 5].

DISCUSSION

This study shows the feasibility of creating a dry lab model and DL-based automatic assessment tool for robotic-assisted surgery. Our results show that our CV algorithm is capable of assessing the proficiency level of a surgical trainee with an accuracy of 66.7% utilizing surgical videos.

Table 1. Participant demographics

	All (n = 27)	Trainees (n = 20)	Attendings (n = 7)
Female n (%)	10 (37%)	9 (45%)	1 (14%)
Right dominant hand n (%)	24 (88.9%)	18 (90%)	6 (86%)
Robotic case experience (median +/- SD)	50 +/- 156	45 +/- 32	186 +/- 231
PGY level (for residents) (median +/- SD)	N/A	3 +/- 1.27	N/A
Years experience after training (attendings) (median +/- SD)	N/A	N/A	8.7 +/- 6.01

PGY: Post graduate year; N/A: not applicable.

Table 2. Performance of each model in the system using 3-fold cross-validation across held-out surgeons. For technical score prediction models, we assume the type of suturing exercise (backhand, railroad) is known prior and apply the corresponding model

Technique	Model	Average weighted F-1 score	Average macro F-1 score	Average accuracy
Sub-stitch classification	Video swin transformer	0.6452	0.6400	0.7023
Technical score prediction - backhand	Video swin transformer	0.7185	0.7155	0.7259
Technical score prediction - railroad	Video swin transformer	0.6430	0.6364	0.6411
Surgeon proficiency prediction	Random forest classifier	0.6266	0.5805	0.6665

This study represents one of the first to utilize AI to automatically assess surgical trainees on a specific robotic surgery task. This study underlines the ability of AI-assisted assessment tools as an effective educational tool for surgical trainees in identifying their proficiency and potentially providing feedback. While Ma *et al.* developed the first AI-based video feedback tool for robotic suturing, their study participants had no robotic surgical experience and, therefore, focused on improving tasks rather than determining proficiency^[16]. Our model is able to provide feedback while also determining the skill level of the trainee.

Suturing is a fundamental surgical skill, and proficiency in this skill implies mastery of many technicalities, such as needle angulation, insertion point, depth, and tissue manipulation. By breaking down the suturing tasks into four sub-stitches: needle positioning, needle targeting, needle driving, and needle withdrawal, trainees can understand what specific needle movements they need to practice while maintaining a standardized taxonomy. This specific suturing taxonomy, based on prior research, allows us and future researchers to have a reproducible methodology around automatic supervised learning suturing assessment^[15,17].

These surgical techniques are vital for surgical trainees to practice more specific movements in a controlled setting before they perform surgery on a patient. It also allows surgical attendings to gain trust in their trainees prior to operating in a real clinical setting based on their robotic proficiency score.

This preliminary study shows only the feasibility of creating this type of model to assess the skill level of a trainee, with an accuracy of 66.7%. While the model does need to be improved, its current accuracy allows for identifying which residents need extra practice. Only six of the participants were false negatives in our study [Figure 3]. Although the accuracy rate is only 66.7%, having residents practice more and then attempt the dry lab again will only improve their skills and therefore the model can be utilized to pick out those trainees who need more practice. However, we do aim to improve our accuracy in the future with a larger

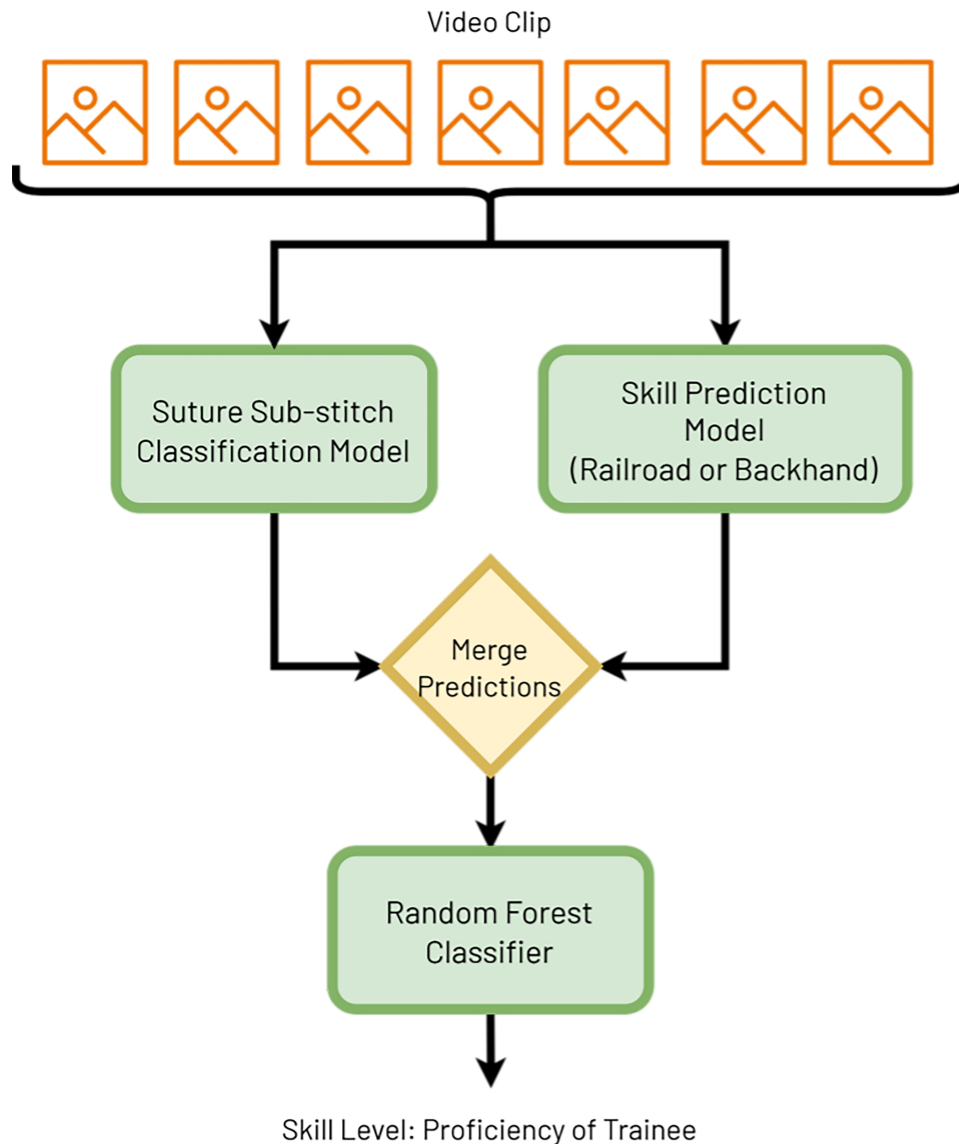


Figure 2. Illustration of the end-to-end system to predict trainee proficiency level.

sample size and more annotations to train our model.

This study was limited by the number of participants. With a larger video training set, we aim to be able to improve our model to enhance accuracy. This dry lab model was also a preliminary model. The model can be improved for better object and task recognition by CV. For example, utilizing a larger needle or larger suture size may allow more differentiation between the background and the needle for the CV algorithm to better pick up the needle movements. In the future, this could be assessed by a separate pre-assessment model to determine how well CV is doing in object identification. With the rapid improvement in CV models, this model could even be further optimized with the use of not only object identification but also gesture recognition. Otiato *et al.* showed that surgical gestures during dissection can vary based on experience and correlate with patient outcomes^[18]. Using a multi-modal model on a dataset with an

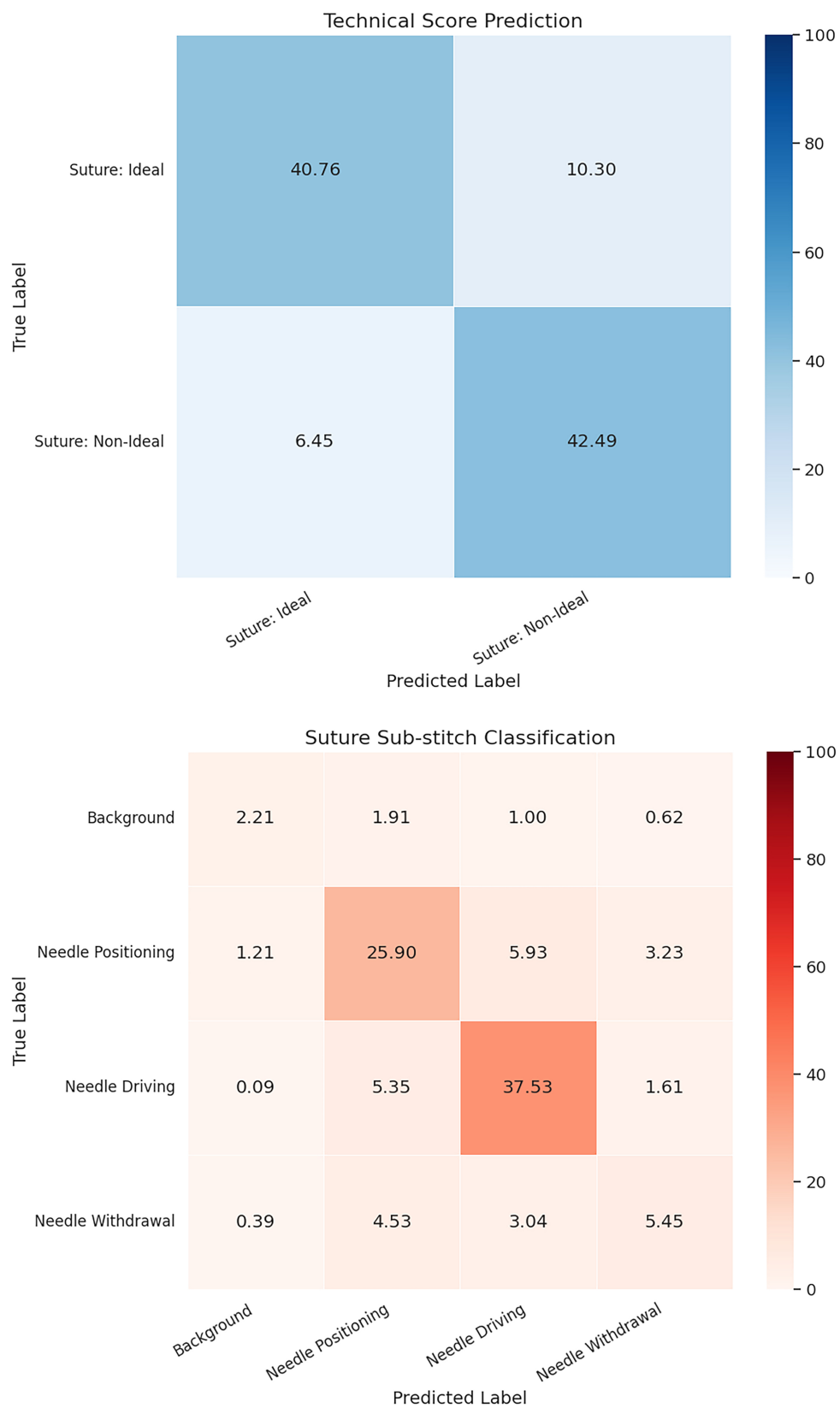


Figure 3. Confusion matrix for: (Top) Technical Score Prediction across both Backhand and Railroad Tasks; (Bottom) Suturing Sub-stitch Classification. Values have been normalized to represent the percentage of total data. The X-axis represents the prediction while the Y-axis represents the actual sub-stitch.

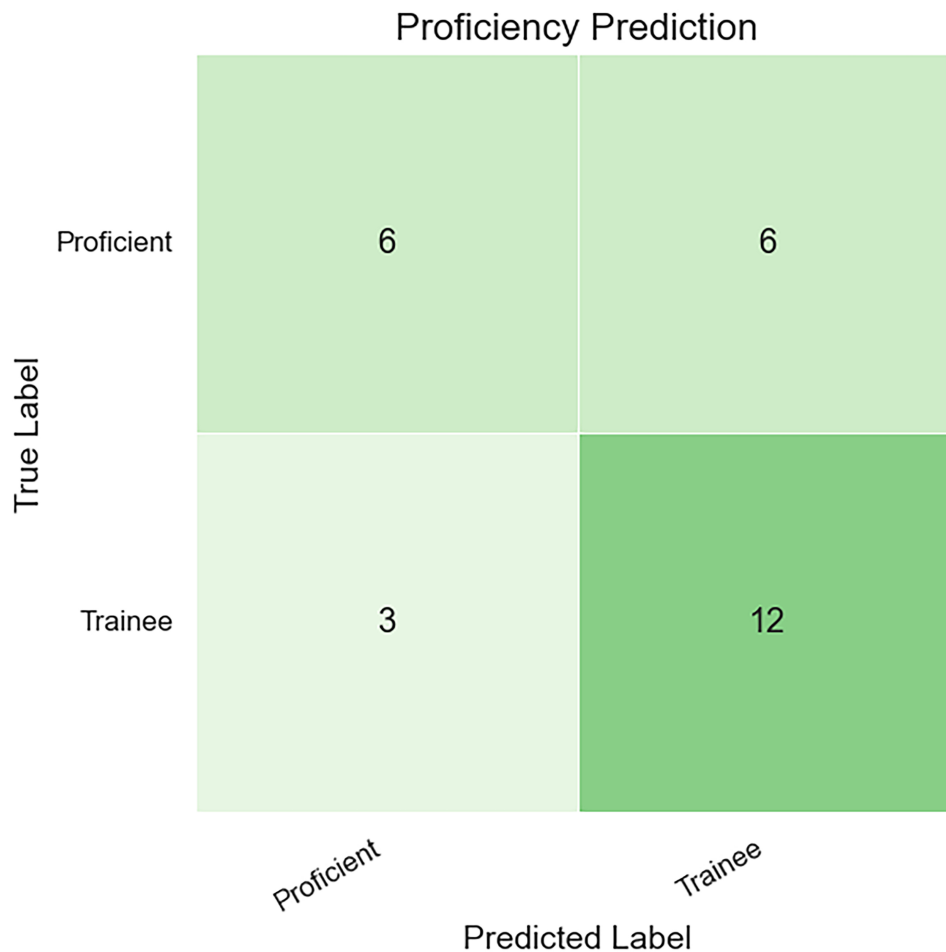


Figure 4. Confusion matrix for surgeon proficiency prediction based on 27 participants. The X-axis shows predicted labels and the Y-axis shows true labels.

increased sample size in the future will help us increase the accuracy of our model. However, this study is a first step to creating an automatic assessment and training tool for surgical trainees that does not require a large time burden for expert surgeons.

With its continued prevalence and popularity, robotic surgery needs a standardized curriculum and assessment for training. Specifically, an objective evaluation method with a limited time burden on expert surgeons is a crucial need in robotic training. This study provides a clinically relevant proposal to improve robotic surgical education. This dry lab model proposes a standardized training tool for suturing tasks utilizing CV for automatic assessment. This relieves the time burden on surgical experts of video-based assessment and removes the subjectivity of an individual expert. It also allows for standardized feedback based on needle movements. This model can be utilized in the future to ensure trainees are at an adequate level before operating on patients to improve surgical safety.

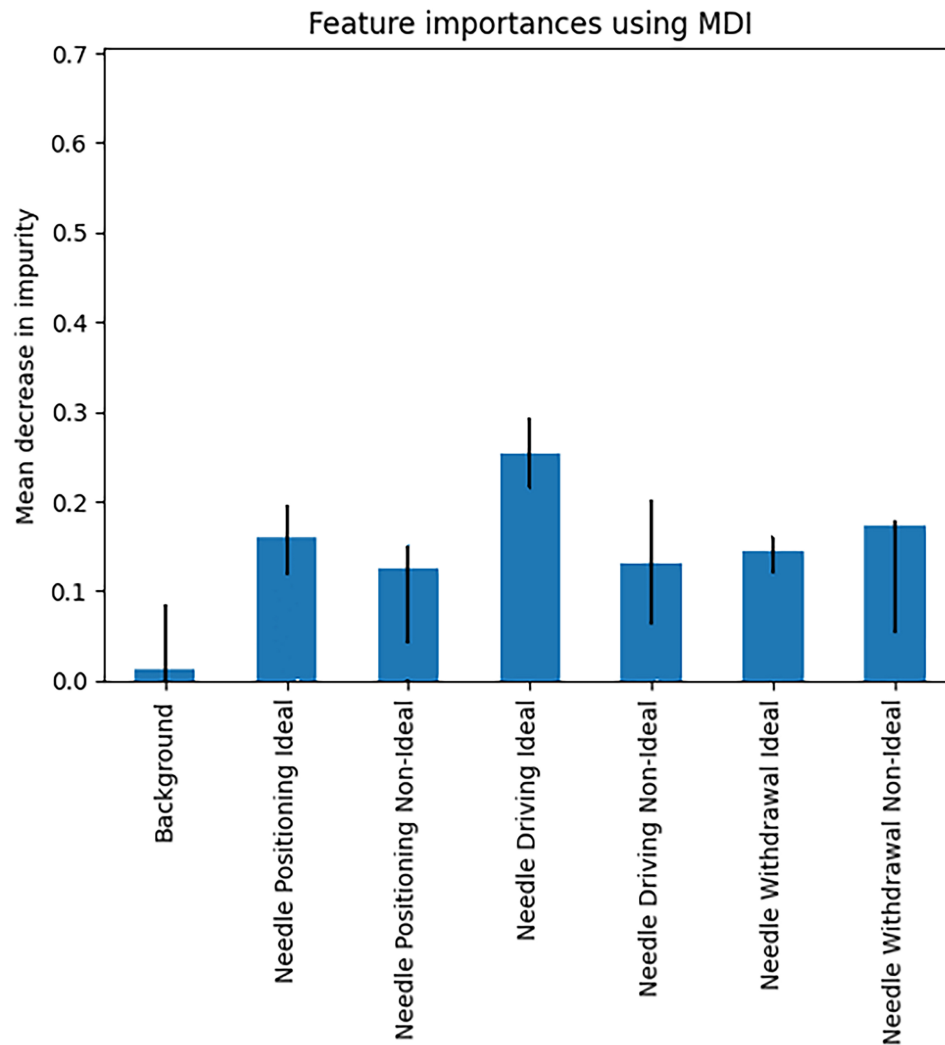


Figure 5. Feature importance analysis from the Decision Tree Classifier for surgeon proficiency prediction using MDI. MDI: Mean decrease in impurity.

DECLARATIONS

Authors' contributions

Made substantial contributions to the conception and design of the study: Choksi S, Filicori F

Performed data acquisition and assisted with data analysis: Ballo M, Farrell A, King B, Nussbaum J, Reisner A, Taffurelli G

Performed data analysis and interpretation: Choksi S, Turkcan M, Narasimhan S, Hu Y, Zang C, Kostic Z

Drafted the manuscript and contributed to significant revisions: Choksi S, Ballo M, Turkcan M, Narasimhan S, Filicori F, Kostic Z

Availability of data and materials

The data are available from the corresponding author upon reasonable request.

Financial support and sponsorship

None.

Conflicts of interest

Filicori F is a paid consultant for Boston Scientific. The other authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

This study was approved by the Institutional Review Board of Northwell Health (IRB 23-069). All participants signed a consent form to participate in the study.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2025.

REFERENCES

1. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surg Endosc*. 2020;34:4924-31. DOI PubMed
2. Cai T, Zhao Z. Convolutional neural network-based surgical instrument detection. *Technol Health Care*. 2020;28:81-8. DOI PubMed PMC
3. Luongo F, Hakim R, Nguyen JH, Anandkumar A, Hung AJ. Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery. *Surgery*. 2021;169:1240-4. DOI PubMed PMC
4. Choksi S, Szot S, Zang C, et al. Bringing artificial intelligence to the operating room: edge computing for real-time surgical phase recognition. *Surg Endosc*. 2023;37:8778-84. DOI PubMed
5. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369:1434-42. DOI
6. Grewal B, Kianercy A, Gerrah R. Characterization of surgical movements as a training tool for improving efficiency. *J Surg Res*. 2024;296:411-7. DOI PubMed
7. Azari DP, Frasier LL, Quamme SRP, et al. Modeling surgical technical skill using expert assessment for automated computer rating. *Ann Surg*. 2019;269:574-81. DOI PubMed PMC
8. Welcome to FLS. Fundamentals of laparoscopic surgery. Available from: <https://www.flsprogram.org/about-fls/>. [Last accessed on 27 Mar 2025].
9. Fuchs HF, Collins JW, Babic B, et al. Robotic-assisted minimally invasive esophagectomy (RAMIE) for esophageal cancer training curriculum-a worldwide Delphi consensus study. *Dis Esophagus*. 2022;35:doab055. DOI PubMed
10. Stegemann AP, Ahmed K, Syed JR, et al. Fundamental skills of robotic surgery: a multi-institutional randomized controlled trial for validation of a simulation-based curriculum. *Urology*. 2013;81:767-74. DOI PubMed
11. Satava RM, Stefanidis D, Levy JS, et al. Proving the effectiveness of the fundamentals of robotic surgery (FRS) skills curriculum: a single-blinded, multispecialty, multi-institutional randomized control trial. *Ann Surg*. 2020;272:384-92. DOI PubMed
12. Ayoub-Charette S, McGlynn ND, Lee D, et al. Rationale, design and participants baseline characteristics of a crossover randomized controlled trial of the effect of replacing SSBs with NSBs versus water on glucose tolerance, gut microbiome and cardiometabolic risk in overweight or obese adult SSB consumer: strategies to oppose SUGARS with non-nutritive sweeteners or water (STOP sugars NOW) trial and ectopic fat sub-study. *Nutrients*. 2023;15:1238. DOI PubMed PMC
13. Lazar A, Sroka G, Laufer S. Automatic assessment of performance in the FLS trainer using computer vision. *Surg Endosc*. 2023;37:6476-82. DOI PubMed
14. Islam G, Kahol K, Li B, Smith M, Patel VL. Affordable, web-based surgical skill training and evaluation tool. *J Biomed Inform*. 2016;59:102-14. DOI PubMed
15. Hung AJ, Bao R, Sunmola IO, Huang DA, Nguyen JH, Anandkumar A. Capturing fine-grained details for video-based automation of suturing skills assessment. *Int J Comput Assist Radiol Surg*. 2023;18:545-52. DOI PubMed PMC
16. Ma R, Kiyasseh D, Laca JA, et al. Artificial intelligence-based video feedback to improve novice performance on robotic suturing skills: a pilot study. *J Endourol*. 2024;38:884-91. DOI PubMed
17. Raza SJ, Field E, Jay C, et al. Surgical competency for urethrovesical anastomosis during robot-assisted radical prostatectomy: development and validation of the robotic anastomosis competency evaluation. *Urology*. 2015;85:27-32. DOI PubMed
18. Otiato MX, Ma R, Chu TN, Wong EY, Wagner C, Hung AJ. Surgical gestures to evaluate apical dissection of robot-assisted radical prostatectomy. *J Robot Surg*. 2024;18:245. DOI PubMed PMC