Surgical Endoscopy
https://doi.org/10.1007/s00464-022-09509-y

ORIGINAL ARTICLE

Developing artificial intelligence models for medical student suturing and knot-tying video-based assessment and coaching

Madhuri B. Nagaraj[1,2] · **Babak Namazi**[1] · **Ganesh Sankaranarayanan**[1] · **Daniel J. Scott**[1,2]

Received: 19 April 2022 / Accepted: 23 July 2022

Abstract

Background Early introduction and distributed learning have been shown to improve student comfort with basic requisite suturing skills. The need for more frequent and directed feedback, however, remains an enduring concern for both remote and in-person training. A previous in-person curriculum for our second-year medical students transitioning to clerkships was adapted to an at-home video-based assessment model due to the social distancing implications of COVID-19. We aimed to develop an Artificial Intelligence (AI) model to perform video-based assessment.

Methods Second-year medical students were asked to submit a video of a simple interrupted knot on a penrose drain with instrument tying technique after self-training to proficiency. Proficiency was defined as performing the task under two minutes with no critical errors. All the videos were first manually rated with a pass-fail rating and then subsequently underwent task segmentation. We developed and trained two AI models based on convolutional neural networks to identify errors (instrument holding and knot-tying) and provide automated ratings.

Results A total of 229 medical student videos were reviewed (150 pass, 79 fail). Of those who failed, the critical error distribution was 15 knot-tying, 47 instrument-holding, and 17 multiple. A total of 216 videos were used to train the models after excluding the low-quality videos. A k-fold cross-validation ($k = 10$) was used. The accuracy of the instrument holding model was 89% with an F-1 score of 74%. For the knot-tying model, the accuracy was 91% with an F-1 score of 54%.

Conclusions Medical students require assessment and directed feedback to better acquire surgical skill, but this is often time-consuming and inadequately done. AI techniques can instead be employed to perform automated surgical video analysis. Future work will optimize the current model to identify discrete errors in order to supplement video-based rating with specific feedback.

Keywords Artificial intelligence · Video-based review · Convolutional neural network · Skills assessment · Suturing and knot-tying simulation

The introduction of surgical skills to pre-clinical medical students has long been supported for a variety of reasons including early exposure, early skill acquisition in a simulation environment, and evidence of the resulting positive attitude towards surgery [1–3]. More specifically, a large proportion of students are given the opportunity to suture while on clerkship and engagement in operations has been significantly associated with positive student perceptions on their clerkship [4].

Feedback and distributed practice have both been identified as critical needs for deliberate practice and skill acquisition [5]. Previous medical student suturing curricula, even those using video-based assessment, have utilized human resources for assessment and feedback [6, 7]. However, the large time investments required for surgical faculty to gain or teach skills with directed feedback have been a longstanding concern [7, 8]. A key component of deliberate practice is however "detailed immediate feedback" [5]. Video-based training has demonstrated many benefits, including relieving

Presented at Society of American Gastrointestinal and Endoscopic Surgeons Next Big Thing Poster Presentation. August 31-September 3, 2021.

✉ Madhuri B. Nagaraj
Madhuri.nagaraj@gmail.com

[1] Department of Surgery, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9159, USA

[2] University of Texas Southwestern Simulation Center, 2001 Inwood Road, Dallas, TX 75390-9092, USA

Published online: 18 August 2022

some of the personnel burden and scheduling issues, and allowing for retrospective review while maintaining equivalent learning outcomes [9–11]. There remains, however, limitations in the ability to provide directed, immediate feedback [10]. There has been some work with variable results demonstrating the accuracy in learner self-assessment of skill acquisition [12, 13]. Furthermore, there has been limited previous work examining the use of artificial intelligence in assessing and providing feedback for medical student skills training [14].

Compared with manual video-based assessment, machine learning and AI-based methods are faster, require less instructor-hours, more reproducible, and potentially less subjective. The flexibility and accessibility of these methods have thus resulted in numerous successful applications in surgical education [15–18]. Nevertheless, the majority of existing work has focused on minimally invasive techniques; within this area, AI work has been developed for instrument tracking, work-flow analysis, the identification of critical anatomy, and basic surgical skills such as robotic suturing and knot-tying [19–23].

We previous reported positive learner outcomes associated with an open suturing and knot-tying curriculum for our second-year medical students which we adapted to an at-home video-based assessment format due to the social distancing requirements of COVID-19 [11]. Students recognized the at-home training environment as a low-stress and convenient practice environment. Additionally, examination of that curriculum demonstrated reduced personnel needs associated with the virtual platform [11]. However, the need for more frequent and directed feedback was also reported. Based on these data, we aimed to use artificial intelligence to develop a Deep Learning (DL) model that would (1) distinguish pass-fail performances to develop an automatic assessment tool, as well as (2) identify the reason for task failure.

Materials and methods

Task details and materials

The task and assessment criteria for this curriculum had been predetermined from previous curriculum content [11]. The medical students were to perform a simple-interrupted knot on a penrose drain model using an instrument tying technique (Fig. 1). All students were provided with the same take-home suturing kits that contained instruments (surgical needle driver, forceps, and scissors), suture material (2–0 silk suture, SH needle), penrose drain models with premarked targets and velcro for model fixation. An instructional 25-min video was provided that explained the task with demonstrations of passing performance, errors, and pitfalls. Students were instructed to self-train to proficiency

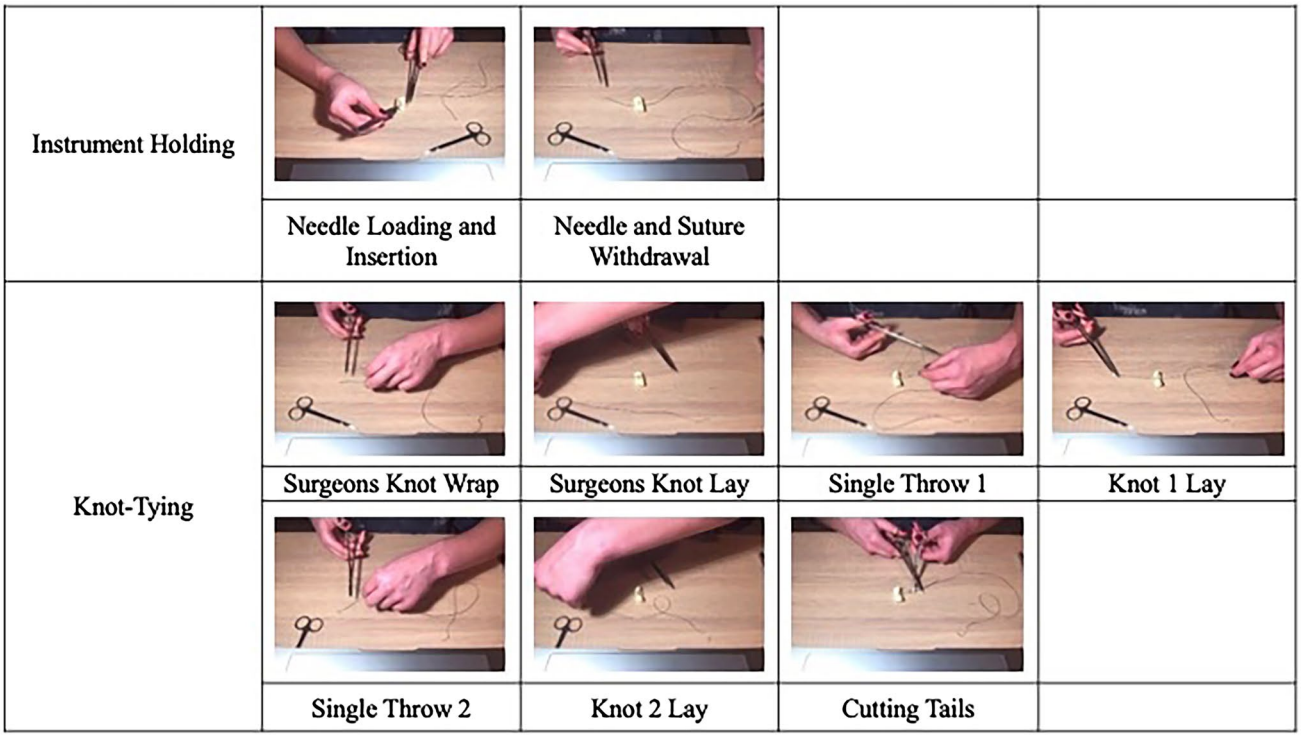

Fig. 1 Snapshots of the task segmentation

and then provide a video of their performance for assessment. The instructional video also included demonstrations of the desired video capture using a laptop webcam or cellphone camera on a tripod with the intended opposite facing 45-degree downward angle view.

Metrics and errors

Students were assigned a task rating using a pass/fail scale that was developed in the previous curriculum [11]. Critical errors, which consisted of a safety or technical error, resulted in a rating of "fail." Safety errors included touching the needle with one's hands. Technical errors were classified as knot-tying errors, instrument holding errors, or combined errors. To develop a more robust error detection for the AI model, the errors were classified in a more granular fashion based on step-wise delineation of the task (Table 1). This study was reviewed and deemed exempt by the University of Texas Southwestern Institutional Review Board.

Video collection

At the completion of their self-training period, students submitted their video for evaluation to an online learning platform D2L (desire to learn). These videos were then downloaded and deidentified. Subsequently, all videos were manually annotated to mark the nine steps of open suturing task (Table 1, Fig. 1).

There existed large variation in the quality of videos with regards to file size, aspect ratio, camera angle, camera motion, field of view, lighting, and the background color. Though a large variation in input data can improve an AI model's robustness, it introduces additional challenges for designing a computer vision-based algorithm such as low lighting, poor contrast between suture and background, and other distractions (clutter/moving objects).

Deep-learning models

Advanced deep-learning techniques and in particular Convolutional Neural Networks (CNNs) have proven effective in capturing high-level representations in images and videos and have been successfully used in many computer vision tasks such as image and video classification and activity recognition [24, 25]. Inspired by the human visual cortex and neural system, a deep CNN consists of several trainable convolutional layers, which sequentially update the input representation through extracting important discriminating patterns. The weights of all the layers are iteratively adjusted by minimizing an error function (also called loss function) during training in an end-to-end fashion using the so-called back-propagation algorithm. Once trained with a large set of image/video instances, the model can perform the desired input–output mapping (such as classification) in new and unseen data points.

To identify an error using deep learning, a training dataset is needed that has a large number of instances of the specific error. Since there are only a few videos for some of the errors (Table 3), we combined multiple error types and categorized the errors as either instrument holding or knot-tying errors. Two deep learning models were trained to detect these errors in the corresponding steps. The rationale for developing separate models was that each error required distinct automated detection methods; while still images could be sufficient for identifying instrument holding errors, detecting the knot-tying errors required tracking and capturing hand movements to assess performance. The description of the models are as follows:

(1) Instrument Holding Error Detection: In the instrument holding error detection model, three types of errors were annotated: incorrect forcep hold, incorrect needle driver hold, and incorrect needle load. Images of correct and incorrect forcep and needle driver holding are shown in Fig. 2. The input frames for this model

Table 1 Error classification system

Curriculum error	AI model error
Knot-tying error	Surgeon's knot: incorrect direction of wrap
	Surgeon's knot: incorrect number of wraps
	Surgeon's knot: incomplete crossing of hands when laying the knot
	Square Knot #1: incorrect direction of wrap
	Square Knot #1: incomplete crossing of hands when laying the knot
	Square Knot #2: incorrect direction of wrap
	Square Knot #2: incomplete crossing of hands when laying the knot
Instrument holding error	Forceps: incorrect hold
	Needle Driver: incorrect hold
	Needle: incorrect load
Combination error	Combination error

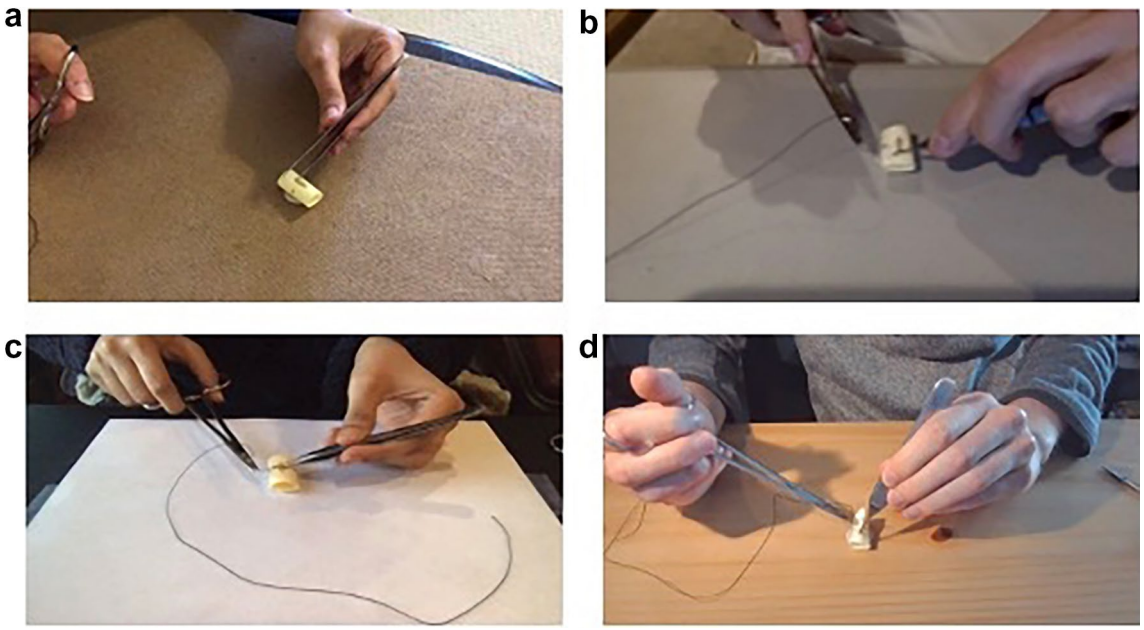

Fig. 2 Examples of instrument holding errors. **a** Correct forceps hold, **b** incorrect forceps hold, **c** correct needle driver hold, and **d** incorrect needle driver hold

were randomly taken from the first two steps of the task (needle loading and insertion through needle and suture withdrawal).

The instrument holding model consisted of a CNN to extract spatial features for each input frame, an attention block to aggregate the extracted features, and a classifier layer to output the probability of pass/fail. The block diagram of the model is shown in Fig. 3. The role of the attention block was to calculate the importance of the input frames in making the pass/fail probability estimation. Using the attention mechanism (Luong et al. 2015), the output vectors of the CNN that are taken from a sequence of frames are given importance weights and the weighted sum of these extracted visual features is used to classify the videos based on the presence of an error. The classifier layer is a fully connected networks that generates the probability of pass or fail. The length of the input sequence was 16 frames. The architecture of the CNN was based on the Efficient-net model that was pre-trained on a large dataset of real-world images named "Imagenet" [26, 27].

(2) Knot-Tying Error Detection: In the knot-tying error detection model, seven types of errors were recognized and annotated (Table 1). As tracking the hand movement is essential in identifying incorrect knot-tying, we developed a model that considers temporal and motion-based features. Therefore, we computed optical flow for all the frames of the videos and used them as the input to the deep learning model. Optical flow methods approximate the motion for each individual pixel in consecutive images and can be used as a measure for objects' velocity in a 2D direction [28, 29]. The advantage of using optical flow input as opposed to the original colored images is that the movements of the objects are considered directly and irrespective of the color, texture or pose of the hand or background. A sequence of 64 optical flow images from consecutive frames (approximately 1 frame per second) were chosen from the next seven knot-tying steps (surgeon's knot wrap to second knot lay). For shorter videos, the sequence was padded with blank (white) images.

The block diagram of the knot-tying error detection model is shown in Fig. 4. As Three-Dimensional (3D) CNNs are powerful models for learning representation from 3D volumetric data and sequence of 2D images such as in videos, a 3D CNN was adopted to capture the temporal features of the sequential input. The architecture of the 3D CNN was based on the X3D model that was pre-trained on a large video dataset [30]. The loss function used for training both models was sigmoid cross-entropy, which is suitable for binary classification (0 for pass, 1 for fail).

Data augmentation

Data augmentation is a commonly utilized and referenced technique for improving the performance and generalizability of deep learning models [31]. In our study, the original dataset used was imbalanced as it contained far more instances of "pass" than "fail" videos. This was especially

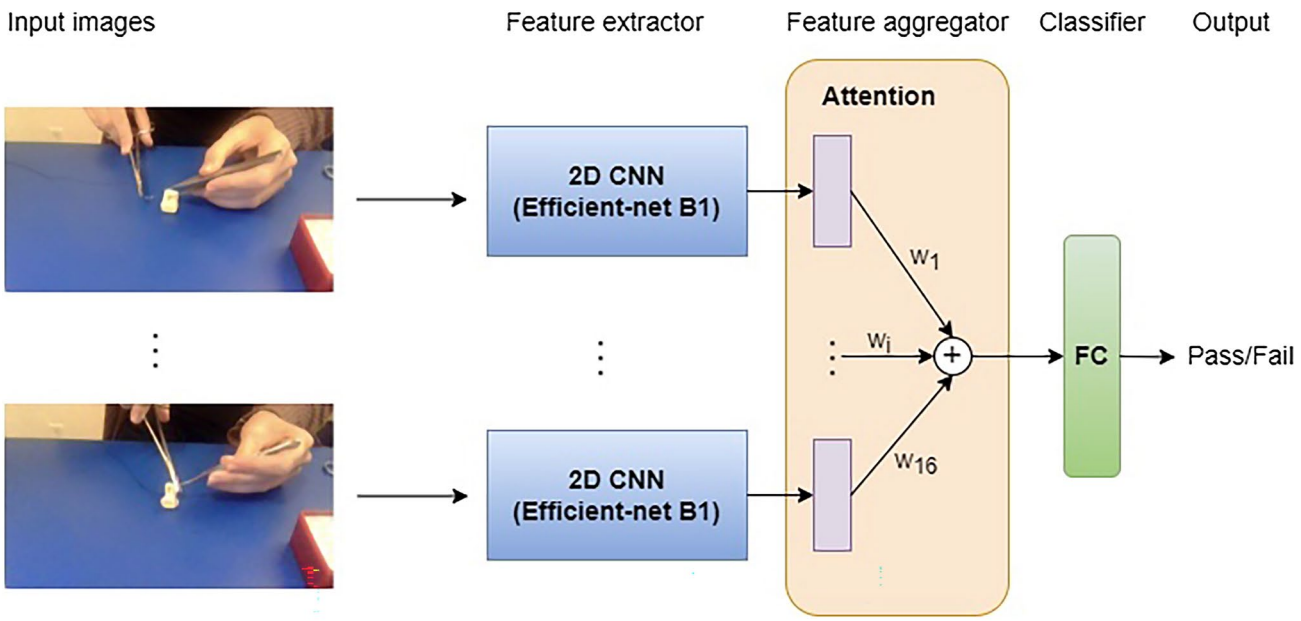

Fig. 3 The block diagram of our instrument holding error detection model. The model's inputs were several randomly chosen images from steps 1–2. The visual features of all the input frames were extracted using a CNN and aggregated in the attention block. Using the attention mechanism, the features were weighted by the trainable parameter w and summed before the fully connected (FC) classfication layer. The output as the probability of the presence/absence of an error (pass/fail)

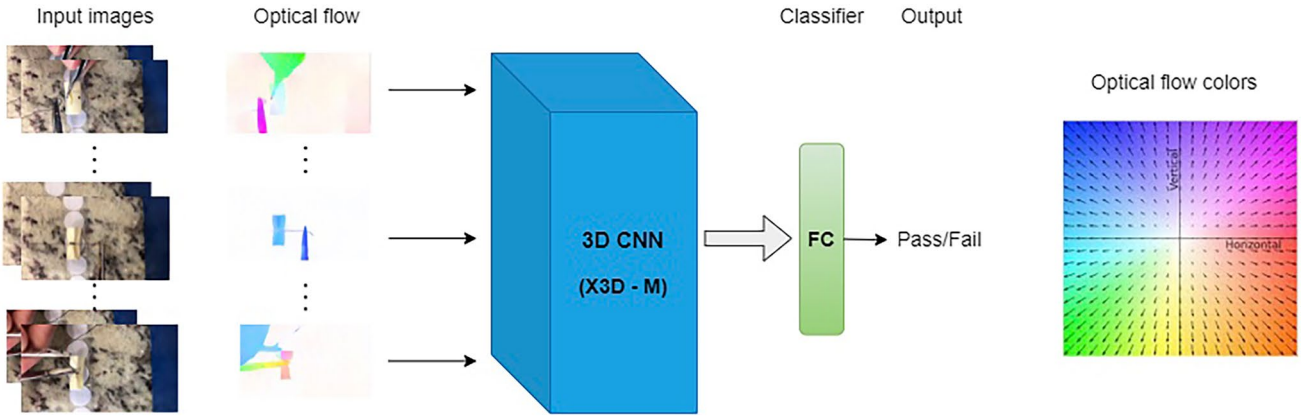

Fig. 4 The block diagram of our knot-tying error detection model. The model's inputs were optical flow images, which represented the movement of the objects/hand (the colored images are shown for visualization only). A 3D CNN extracted the spatial and temporal pattern and generated the probability of presence/absence of an error (pass/fail)

true for the knot-tying error detection model. This imbalance can affect the model performance negatively; to mitigate this issue, one investigator (MBN) deliberately committed a variety of errors while performing the task. These 45 "fail" videos were used to supplement the dataset and reduce the imbalance ratio during model training. We utilized other well-known data augmentation methods in the development of these models; these included random image crop, translation, rotation, temporal speed change, and color and brightness adjustment.

Deep-learning training and evaluation

The models were implemented using the Keras library, a popular deep learning library [32]. The batch size for training the instrument holding error detection model and the knot-tying error detection model were 16 and 8 respectively.

Both of the models were trained using the A100 Graphic Processing Unit (GPU) with 40 GB of memory. An SGD optimizer was used for both models with the learning rate of 0.001 and decay rate of 0.7 after 10 epochs (https://portal.biohpc.swmed.edu/content/).

To evaluate the performance of the trained models, we used a tenfold cross-validation method. The extra videos of fail instances were used in all the 10 folds training sets. For overall results, a video was labeled as fail if at least one type of error occurred. We defined True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) as correctly predicted as fail, correctly predicted as pass, pass incorrectly predicted as fail, and fail incorrectly predicted as pass, respectively. Based on these definitions, the metrics used for validation (accuracy, precision, sensitivity and F1-score) were calculated for both models as well as the overall pass/fail grade:

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

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

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

To further analyze and optimize the prediction probabilities, Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were plotted. The ROC curve summarizes the trade-off between the true positive rate (precision) and false-positive rate ($\frac{FP}{FP+TN}$), whereas the PR curve summarizes the trade-off between precision and sensitivity. Both plots are used for analyzing the probability threshold for classification models.

Results

Video-based ratings and characteristics

As previously reported, a total of 229 medical students participated in the curriculum and provided video performances for manual assessment [11]. Of these, 65.5% (150/229) resulted in a passing rating. A rating of fail was given to 34.5% (79/229) of which the most common, 59.5% (47/79), were instrument holding errors (Tables 2 and 3). The majority of trainees performed the task with their right-hand, 97.4% (223/229), as compared to the left-hand 2.6% (6/229). It is unclear how many are left-hand dominant and performed the task with right non-dominant hand. Videos were shot opposite to the learner for a front angle in 87.3% (200/229) of the videos, while over-the-shoulder angles made up the remaining 12.7% (29/229).

Table 2 Student performance ratings by pass/fail and error type

Student performance rating	Error classification	Number of videos (*n*)
Pass grade		150
Fail grade		79
	Knot-tying errors	15
	Instrument-holding error	47
	Combined error	17
		Total = 229

Table 3 Total number of videos for each error type

Error	Number of videos (*n*)
Incorrect forcep hold	53
Incorrect needle driver hold	1
Incorrect needle load	2
Surgeons knot direction	5
Surgeons knot wrap #	2
Surgeons knot lay	7
Knot #1 wrap	16
Knot #1 lay	6
Knot #2 wrap	4
Knot #2 lay	12

Certain characteristics made for poor video quality. These included a moving/non-stationary camera (usually held manually by another individual), low lighting, poor contrast with a dark or patterned background, and distractions (clutter or other moving objects). After deliberation, 7.0% (16/229) of the videos were excluded from use for model training due to poor quality. The remaining videos were cropped and resized to 480*480 to better focus on the moving objects while maintaining the aspect ratio. All videos were saved at 30 frames per second.

Task performance time in median [interquartile range] was significantly shorter for passing group at 53 s [43.25–64.75] as compared to the failing group at 62 s [51–76.5] (p-value < 0.01). The descriptive statistics of the performance times of each steps is reported with a mean time for instrument holding at 24.13 + 7.52 s and knot-tying at 35.66 + 6.44 seconds (Table 4).

Table 4 Performance times (seconds) for each step

Task segmentation	Mean	STD	Min	Max
1. Needle loading and insertion	15.02	6.4	5	37
2. Needle and suture withdrawal	9.11	3.95	3	28
3. Surgeons knot wrap	7.95	3.77	2	26
4. Surgeons knot lay	5.81	2.45	2	16
5. Single throw 1	4.85	2.85	1	22
6. Knot 1 lay	4.33	2.08	1	17
7. Single throw 2	4.02	2.08	2	14
8. Knot 2 lay	4.37	2.13	1	14
9. Cutting tails	8.93	3.49	1	30

Deep learning model results

The cross-validation results demonstrate 0.83 overall pass/fail accuracy with an F1-score of 0.69 (Table 5). The accuracy of the instrument holding error model was 0.89 and the F1-score 0.74. The accuracy of the knot-tying error detection model was 0.91 and the F1-score 0.54. The area under the ROC curve for the instrument holding error detector and knot-tying error detector was 0.84 (95% CI 0.76–0.91) and 0.67 (95% CI 0.55–0.79) respectively; additionally, the average precision was 0.66 (95% CI 0.54–0.82) and 0.49 (95% CI 0.30–0.67) respectively (Fig. 5). Confusion matrices for the two models are shown in Fig. 6.

Discussion

Optimizing learner assessment and feedback remains a large focus of ongoing medical education given the proven benefits despite high person-hour requirements. This was an innovative study which aimed to develop a DL model for automatic rating development (assessment) and error detection (feedback) of a medical student knot-tying task. Our work demonstrated some interesting findings including the high accuracy and low sensitivity of our preliminary model, the impact of imbalanced video distributions in CNN training, and the need for video standardization. Because of the limited number of videos for some error types, training and evaluating DL models independently for all types of errors was not possible. Therefore, we grouped the errors into two categories of instrument holding and knot-tying errors and separate models were developed for each of these categories.

For a DL model to be useful in assessment and feedback, it requires high accuracy and precision. Fortunately, our models had high accuracy scores (89% and 91%), which demonstrate the potential to both (1) identify the presence/absence of the two types of errors and (2) to provide the pass/fail score. The relatively lower F1-score and

Table 5 Results of the deep learning models*

	Accuracy	Precision	Sensitivity	F1-score
Instrument holding error detection	0.89 ± 0.05	0.85 ± 0.28	0.65 ± 0.29	0.74 ± 0.27
Knot-tying error detection	0.91 ± 0.03	0.92 ± 0.44	0.38 ± 0.31	0.54 ± 0.32
Overall pass/fail	0.83	0.86	0.58	0.69

*Data reported as mean and standard deviations where appropriate

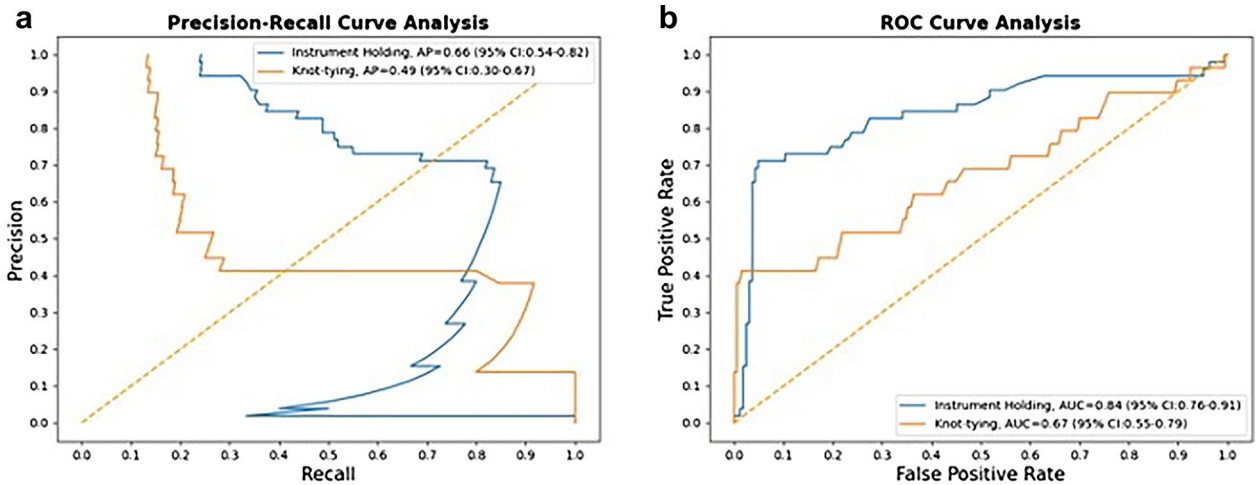

Fig. 5 Precision-Recall (**a**) and Received Operating Characteristic (**b**) curves for the instrument holding and knot-tying error detection models

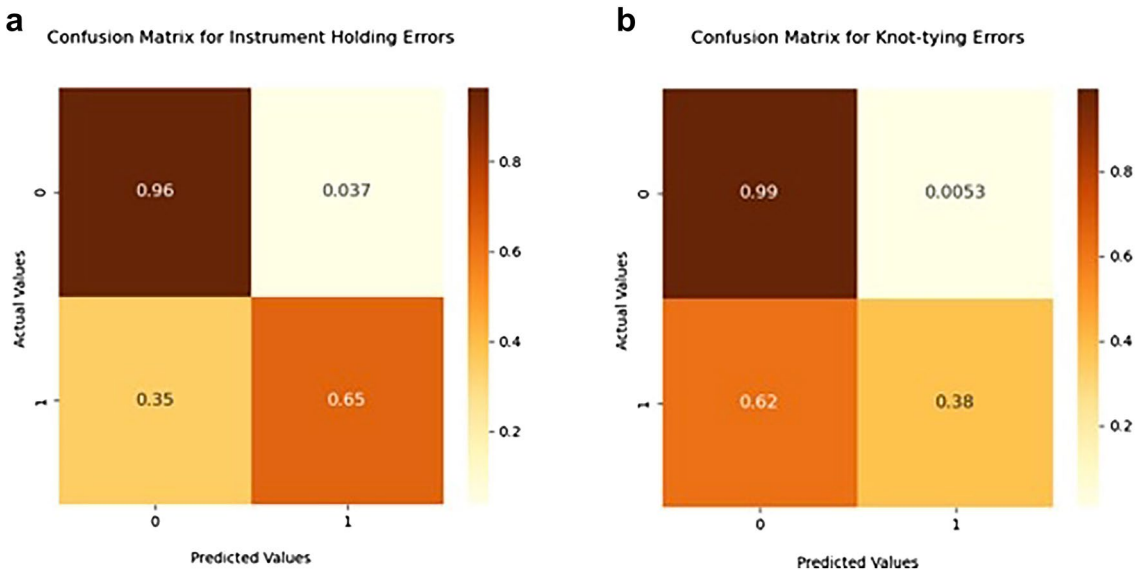

Fig. 6 Confusion matrices for the (**a**) instrument holding and (**b**) knot-tying error detection models

the large gap between accuracy and F1, however, indicated worse performance on fail videos. This was likely due to a high imbalance of pass versus fail videos in our dataset for both models.

As compared to the instrument holding error detection model (three errors), the knot-tying error detection model (seven errors) performed worse overall. This was further evident when observing the difference between precision and sensitivity. The knot-tying model had a higher precision (0.92), while having a lower sensitivity. We suspect that this was likely due to the exaggeration of the imbalance problem again across multiple different error types, which was more severe in knot-tying errors (Table 5).

The high precision of the two models indicated a low FP rate; however, the low sensitivity indicated a higher FN rate (Fig. 6). In practice, this is an important limitation. While FPs would likely get a human review or secondary evaluation, FNs may allow false trainee certification without further assessment. Incorrectly predicting a fail as a pass rating could undermine a robust proficiency-based curriculum.

The person-hours required for someone to evaluate and provide feedback for all TP and FP videos would likely be dramatically reduced from having to manually grade the entire cohort of student videos. Indeed, in previous published data regarding the same curriculum, it was estimated that 10 person-hours were required to manually grade all videos and a subsequent 24 person-hours to perform remediation for groups ranging 2–9 students [11]. However, identifying and subsequently minimizing FNs would be difficult. It is thus essential to analyze the trade-off between precision and sensitivity to minimize the FNs. Indeed, the ROC and PR curves are powerful tools in identifying the right threshold in our binary classifiers' prediction (pass/fail).

Another finding was the lack of standardization in the videos that impacted the quality and quantity of model training. Indeed, 7.0% (16/229) of the originally submitted videos were excluded from model training for various reasons. This was despite discrete instructions being given to learners regarding how to arrange their workspace and record their performance. Problems included moving/non-stationary cameras (often held by another individual), low lighting, minimal contrast of the materials from the background with the use of dark or marble patterned counter spaces, and distractions (other moving objects or clutter in the background). While the video quality of these submissions may have been sufficient for human assessment, these conditions provided inadequate quality for AI assessment models. Thus, these findings demonstrated the importance of standardization in video-based and AI assessment.

The strengths of this study were primarily in the seminal use of and the potential for AI-based assessment and feedback in simulation-based education. Previous studies largely focused on using AI based models to evaluate expert versus novice performance. However, our work provided a novel use of AI models in developing trainee (medical student) evaluation and feedback tools to augment the video-based education learner's experience for open skills. Constructive student feedback on the original curriculum raised a concern in learning a task incorrectly and then subsequently cementing incorrect technique into habits with repetitive practice without feedback or until final assessment [11]. The benefits of AI models included their speed and automaticity, which may allow for not only summative feedback but

also the potential for formative feedback in the setting of limited instructor person-hours. Future efforts will include developing task deconstruction and error analysis to not only provide "pass"/"fail" metrics, but rather more discrete error detail such as "incorrect forceps hold (step 1)" or "incorrect direction of surgeon's knot wrap (step 3)." This automated form of formative assessment could allow for the early recognition and correction of incorrect technique while also optimizing the use of directed feedback for all learners. Additional work will include performing data augmentation for balanced distribution of "fails" using methods such as obtaining pre-training performance videos, recruiting more novice learners, and standardizing video capture (tripod, white paper background, measured setup distances, camera quality) to limit the number of videos that are removed from analysis.

The authors recognize several limitations. First, the poor video quality and the removal of 16 videos from analysis could have skewed the data, but in which manner remains unclear. Next the imbalance of videos across both pass/fail and early/late practice performance leads to underdeveloped model training. Future efforts to mitigate such imbalance issues include training the model with more "fail" videos as a method of intentional data augmentation beyond what was attempted to further reduce the accuracy-F1 gap and improve the overall performance. Finally, our analysis generated an overall accuracy of 83%, which fell short of the generally accepted threshold of 90% or greater.

Conclusions

We aimed to create an AI model that would provide an automatic assessment tool for our previously reported suturing and knot-tying curriculum. Our preliminary data showed positive progress in accuracy of error detection by performance improvement methods including designing a multi-stream architecture and training it in end-to-end fashion. Identified needs include having additional videos with more errors for training, standardization of video capture, and balanced rating distribution for improved model training. We hope that this work will guide further efforts in automated assessment for summative as well as eventually formative training feedback. We anticipate that such innovations will help reduce the burden of relying on instructor support, while improving efforts in simulation-based education.

Acknowledgements The authors gratefully acknowledge support provided by the UT Southwestern Simulation Center.

Author contributions MN: data collection, data annotation, data analysis and manuscript drafting and finalization. BN: data annotation, data analysis and manuscript drafting and finalization. GS: data annotation, data analysis and manuscript drafting and finalization. DS: data collection, data analysis and manuscript drafting and finalization.

Declarations

Disclosure Drs. Madhuri B. Nagaraj, Babak Namazi, Ganesh Sankaranarayanan, and Daniel J. Scott have no conflicts of interest or financial ties to disclose.

References

1. McAnena PF, O'Halloran N, Moloney BM, Courtney D, Waldron RM, Flaherty G, Kerin MJ (2018) Undergraduate basic surgical skills education: impact on attitudes to a career in surgery and surgical skills acquisition. Ir J Med Sci 187(2):479–484. https://doi.org/10.1007/s11845-017-1696-7 (**Epub 2017 Oct 17**)
2. Manning EP, Mishall PL, Weidmann MD, Flax H, Lan S, Erlich M, Burton WB, Olson TR, Downie SA (2018) Early and prolonged opportunities to practice suturing increases medical student comfort with suturing during clerkships: Suturing during cadaver dissection. Anat Sci Educ 11(6):605–612. https://doi.org/10.1002/ase.1785 (**Epub 2018 Mar 30**)
3. Luhoway JA, Ryan JF, Istl AC, Davidson J, Christakis N, Bütter A, Mele T (2019) Perceived barriers to the development of technical skill proficiency in surgical clerkship. J Surg Educ 76(5):1267–1277. https://doi.org/10.1016/j.jsurg.2019.03.020 (**Epub 2019 Apr 16**)
4. McKinley SK, Cassidy DJ, Mansur A, Saillant N, Ghosh A, Evenson A, Askari R, Haynes A, Cho N, James BC, Olasky J, Rangel E, Petrusa E, Phitayakorn R (2020) Identification of specific educational targets to improve the student surgical clerkship experience. J Surg Res 254:49–57. https://doi.org/10.1016/j.jss.2020.03.066 (**Epub 2020 May 11**)
5. Ericsson KA (2004) Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 79(10 Suppl):S70-81. https://doi.org/10.1097/00001888-200410001-00022
6. Olasky J, Kim M, Muratore S, Zhang E, Fitzgibbons SC, Campbell A, Acton R, ACS/ASE Medical Student Simulation-based Research Collaborative Group (2019) ACS/ASE medical student simulation-based skills curriculum study: implementation phase. J Surg Educ 76(4):962–969. https://doi.org/10.1016/j.jsurg.2019.01.014 (**Epub 2019 Feb 21**)
7. Jowett N, LeBlanc V, Xeroulis G, MacRae H, Dubrowski A (2007) Surgical skill acquisition with self-directed practice using computer-based video training. Am J Surg 193(2):237–242. https://doi.org/10.1016/j.amjsurg.2006.11.003
8. Jaffe TA, Hasday SJ, Knol M, Pradarelli J, Pavuluri Quamme SR, Greenberg CC, Dimick JB (2018) Strategies for new skill acquisition by practicing surgeons. J Surg Educ 75(4):928–934. https://doi.org/10.1016/j.jsurg.2017.09.016 (**Epub 2017 Sep 30**)
9. Summers AN, Rinehart GC, Simpson D, Redlich PN (1999) Acquisition of surgical skills: a randomized trial of didactic, videotape, and computer-based training. Surgery 126(2):330–336
10. Rogers DA, Regehr G, Yeh KA, Howdieshell TR (1998) Computer-assisted learning versus a lecture and feedback seminar for teaching a basic surgical technical skill. Am J Surg 175(6):508–510. https://doi.org/10.1016/s0002-9610(98)00087-7
11. Nagaraj MB, Campbell KK, Rege RV, Mihalic A, Scott DJ (2022) At-home medical student simulation: achieving knot-tying

proficiency using video-based assessment. Global Surg Educ. https://doi.org/10.1007/s44186-022-00007-2 (**Epub 1 Jan 2022**)
12. MacDonald J, Williams RG, Rogers DA (2003) Self-assessment in simulation-based surgical skills training. Am J Surg 185(4):319–322. https://doi.org/10.1016/s0002-9610(02)01420-4
13. Evans AW, Leeson RM, Newton-John TR (2002) The influence of self-deception and impression management on surgeons' self-assessment scores. Med Educ 36(11):1095. https://doi.org/10.1046/j.1365-2923.2002.134612.x
14. Emmanuel T, Nicolaides M, Theodoulou I, Yoong W, Lymperopoulos N, Sideris M (2021) Suturing skills for medical students: a systematic review. In Vivo 35(1):1–12. https://doi.org/10.21873/invivo.12226
15. Maier-Hein L, Eisenmann M, Sarikaya D, März K, Collins T, Malpani A, Fallert J, Feussner H, Giannarou S, Mascagni P, Nakawala H, Park A, Pugh C, Stoyanov D, Vedula SS, Cleary K, Fichtinger G, Forestier G, Gibaud B, Grantcharov T, Hashizume M, Hackmann-Nötzel D, Kenngott HG, Kikinis R, Mündermann L, Navab N, Onogur S, Roβ T, Sznitman R, Taylor RH, Tizabi MD, Wagner M, Hager GD, Neumuth T, Padoy N, Collins J, Gockel I, Goedeke J, Hashimoto DA, Joyeux L, Lam K, Leff DR, Madani A, Marcus HJ, Meireles O, Seitel A, Teber D, Ückert F, Müller-Stich BP, Jannin P, Speidel S (2022) Surgical data science-from concepts toward clinical translation. Med Image Anal 76:102306. https://doi.org/10.1016/j.media.2021.102306 (**Epub 2021 Nov 18**)
16. Yanik E, Intes X, Kruger U, Yan P, Miller D, Van Voorst B, Makled B, Norfleet J, De S (2021) Deep neural networks for the assessment of surgical skills: a systematic review. J Def Model Simul. https://doi.org/10.1177/15485129211034586
17. Ward TM, Mascagni P, Madani A, Padoy N, Perretta S, Hashimoto DA (2021) Surgical data science and artificial intelligence for surgical education. J Surg Oncol 124(2):221–230
18. Rogers MP, DeSantis AJ, Janjua H, Barry TM, Kuo PC (2021) The future surgical training paradigm: virtual reality and machine learning in surgical education. Surgery 169(5):1250–1252
19. Namazi B, Sankaranarayanan G, Devarajan V (2022) A contextual detector of surgical tools in laparoscopic videos using deep learning. Surg Endosc 36(1):679–688. https://doi.org/10.1007/s00464-021-08336-x (**Epub 2021 Feb 8**)
20. Garrow CR, Kowalewski KF, Li L, Wagner M, Schmidt MW, Engelhardt S, Hashimoto DA, Kennegott HG, Bodenstedt S, Speidel S, Müller-Stich BP, Nickel F (2021) Machine learning for surgical phase recognition: a systematic review. Ann Surg 273(4):684–693
21. Madani A, Namazi B, Altieri MS, Hashimoto DA, Rivera AM, Pucher PH, Navarete-Welton A, Sankaranarayanan G, Brunt LM, Okrainec A, Alseidi A (2021) Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Ann Surg. https://doi.org/10.1097/SLA.0000000000004594 (**Epub ahead of print**)
22. Mascagni P, Vardazaryan A, Alapatt D, Urade T, Emre T, Fiorillo C, Pessaux P, Mutter D, Marescaux J, Costamagna G, Dallemagne B, Padoy N (2021) Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Ann Surg. https://doi.org/10.1097/SLA.0000000000004351 (**Epub ahead of print**)
23. Ahmidi N, Tao L, Sefati S, Gao Y, Lea C, Haro BB, Zappella L, Khudanpur S, Vidal R, Hager GD (2017) A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Trans Biomed Eng 64(9):2025–2041
24. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
25. Luong T, Pham H, Manning CD (2015). Effective approaches to attention-based neural machine translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421. https://doi.org/10.18653/v1/D15-1166
26. Tan M, Le Q (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, PMLR 97:6105–6114, 2019.
27. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. IEEE Conf Comput Vis Pattern Recognit. https://doi.org/10.1109/cvprw.2009.5206848
28. Horn BK, Schunck BG (1981) Determining optical flow. Artif Intell 17(1–3):185–203. https://doi.org/10.1016/0004-3702(81)90024-2
29. Pérez JS, Meinhardt-Llopis E, Facciolo G (2013) TV-L1 optical flow estimation. Image Process Line 2013:137–150. https://doi.org/10.5201/ipol.2013.26
30. Feichtenhofer C (2020). X3d: Expanding architectures for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 203-213. Doi: DOI:https://doi.org/10.1109/cvpr42600.2020.00028
31. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6:60. https://doi.org/10.1186/s40537-019-0197-0
32. Chollet F (2015). Keras. GitHub. https://github.com/fchollet/keras. Accessed 4 Apr 2022