International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-022-02683-3

ORIGINAL ARTICLE

Scoring metrics for assessing skills in arthroscopic rotator cuff repair: performance comparison study of novice and expert surgeons

Doga Demirel[1] · Bryce Palmer[1] · Gunnar Sundberg[1] · Bayazit Karaman[1] · Tansel Halic[2] · Sinan Kockara[2] · Nizamettin Kockara[3] · Mark Edward Rogers[4] · Shahryar Ahmadi[5]

Received: 18 December 2021 / Accepted: 17 May 2022

Abstract

Purpose We aim to develop quantitative performance metrics and a deep learning model to objectively assess surgery skills between the novice and the expert surgeons for arthroscopic rotator cuff surgery. These proposed metrics can be used to give the surgeon an objective and a quantitative self-assessment platform.

Methods Ten shoulder arthroscopic rotator cuff surgeries were performed by two novices, and fourteen were performed by two expert surgeons. These surgeries were statistically analyzed. Two existing evaluation systems: Basic Arthroscopic Knee Skill Scoring System (BAKSSS) and the Arthroscopic Surgical Skill Evaluation Tool (ASSET), were used to validate our proposed metrics. In addition, a deep learning-based model called Automated Arthroscopic Video Evaluation Tool (AAVET) was developed toward automating quantitative assessments.

Results The results revealed that novice surgeons used surgical tools approximately 10% less effectively and identified and stopped bleeding less swiftly. Our results showed a notable difference in the performance score between the experts and novices, and our metrics successfully identified these at the task level. Moreover, the F1-scores of each class are found as 78%, 87%, and 77% for classifying cases with no-tool, electrocautery, and shaver tool, respectively.

Conclusion We have constructed quantitative metrics that identified differences in the performances of expert and novice surgeons. Our ultimate goal is to validate metrics further and incorporate these into our virtual rotator cuff surgery simulator (ViRCAST), which has been under development. The initial results from AAVET show that the capability of the toolbox can be extended to create a fully automated performance evaluation platform.

Keywords Arthroscopic surgery performance metrics · Deep learning · Surgeon skills · Rotator cuff · Arthroscopy

Introduction

Arthroscopy is a minimally invasive surgical procedure performed via small incisions in the patient's skin to examine, diagnose, and repair the injuries inside a joint. Arthroscopic Rotator Cuff (ARC) is a surgical treatment for muscles and tendons that connect the upper arm to the shoulder blade. Surgeons insert pencil-sized instruments with a small lens and lighting into the joint and see the anatomy on a 2D monitor screen streamed from an arthroscope, a small rigid fiber optic camera with a light source. The goal of the anchor placement is to increase strength for tensile stress. The surgeon uses multiple anchors evenly distributed over the humeral head area to divide the load equally and then sutures the torn rotator cuff with threads attached to the anchor. One of the fundamental skills that a surgeon needs to master for the arthroscopic rotator cuff treatment includes arthroscopic navigation and anatomical landmark detection, bursectomy, bone drilling, anchor placement, suture-to-bone fixation, suture-to-tendon fixation, abrasion resistance of suture, suture strength, knot tying, and knot security tasks

✉ Doga Demirel
ddemirel@floridapoly.edu

1 Computer Science Department, Florida Polytechnic University, 4700 Research Way, Lakeland, FL 33805, USA

2 Computer Science Department, University of Central Arkansas, Conway, AR, USA

3 Department of Orthopedics and Traumatology, Erzincan University Medical School, Erzincan, Turkey

4 Alabama Ortho Spine and Sports in Birmingham, Birmingham, AL, USA

5 Memorial Orthopaedic Surgical Group, Long Beach, CA, USA

Published online: 07 June 2022

to name a few. Based on the tear and its location above-mentioned tasks can be performed multiple times, such as suturing happening three times for a triple-loaded suture anchor. If there need to be multiple anchor placements, these tasks may be repeated multiple times.

Arthroscopic rotator cuff repair has seen a notable increase in the frequency of procedures performed per year. From 2007 to 2015, rotator cuff repair surgeries have increased by 188% [1]. While there has been a drastic increase in the volume of surgeries, surgery training remains lacking and is based on traditional methods. Traditional training methods for residents and surgeons include cadavers, mannequins, and the apprenticeship model, which are limited in their cost, realism, and associated with high-risk factors (e.g., practicing on actual patients) [2–5]. The problem becomes more acute considering working hour restrictions which lead to the acquisition of proficient skills in shorter than necessary time [6]. We envision that virtual reality (VR) based simulation training can provide a valuable aid to traditional methods in training arthroscopic surgeries.

We have been developing a VR simulator to diagnose and repair rotator cuff tears called Virtual Rotator Cuff Arthroscopic Skill Trainer (ViRCAST). Our long-term aim is to provide a high fidelity, low-cost arthroscopic training platform to enhance surgery training and assessment with authentic performance measurements. We envision that surgeons' training and skill level at any level of experience can be quantified. However, current assessments of surgeons in arthroscopy training are very subjective and primarily based on the opinion of the supervising surgeon. The performance feedback might not stem from evaluating essential procedural cognitive or psychomotor skills.

On the other hand, VR simulators can provide unbiased and detailed procedural feedback and categorize the surgeon's skill level. However, this objective measurement requires the development and validation of metrics that directly map to the surgeon's performance in the operating room [7–10]. The metrics need to be well defined and objective to capture details of all the tasks (e.g., including discretionary or cognitive tasks) specific to the procedure. The existing metrics in the literature are either general and very subjective [11, 12] or not validated with actual operating room performance [13].

The processes or tasks by which experts outperform novices have been studied in various contexts such as chess matches, solving physics problems, or nursing [14]. Experts tend to make cognitive decisions and perform tasks rapidly without thinking as much before performing an action. This decision process in novices, due to a lack of experience and repetition, can result in slow cognition and misjudgment [14].

In our prior study [15], the performance of expert surgeons in a variety of complex cases of arthroscopic rotator cuff surgery is measured. This study aims to validate our proposed arthroscopy metrics [15] that aim to distinguish surgery performances based on quantitative measurements. Thus, we hypothesize that the proposed metrics could objectively assess the surgery skills between the novice and the expert surgeons. In this study, we have compared the performance of a group of novice surgeons with expert surgeons to validate the hypothesis further. Moreover, we created a preliminary deep learning-based model to automate the quantification and assessment of the surgeons based on their surgery performances.

Methods

We have analyzed surgeons' performances from arthroscopic view recordings of rotator cuff surgeries to develop and validate our metrics. We first mapped segments of the videos to the tasks performed and measured the total time spent on each of these tasks. For instance, knot tying was mapped to the average of each knot tie time and knot cut time, while bursectomy was mapped to the pre-clean time. In some tasks, the completion time of a task can indicate the surgeon's experience level. The ideal outcome of a task (e.g., without any errors) can be measured with checklist items associated with numerical scores. The final performance score is computed as the average score out of the list of all these measures, which indicates the skill levels (e.g., skills in arthroscope manipulation) and quality of the overall procedure.

The metrics were primarily developed for arthroscopic rotator cuff repair surgery tasks. To validate our metrics, we merged the previously proposed metrics in the literature, Arthroscopic Surgical Skill Evaluation Tool (ASSET) [11] and Basic Arthroscopic Knee Skill Scoring System (BAKSSS) [12], mapped our objective metrics to the merged metrics and compared the summary of skill outcomes.

Subjects

We analyzed 24 arthroscopy rotator cuff videos from two novices and two experts. The videos were recorded from an arthroscopic surgery view. Ten of these videos were performed by the novices, each novice surgeon performing five surgeries. Novice surgeons are defined as surgeons with extensive residency training, while expert surgeons are those who have undergone fellowship programs for rotator cuff repair procedures. We conducted questionnaires about the surgeons to identify and quantify their skill levels. The number of surgeries they had completed, the number of rotator cuff surgeries they had seen in the last six months (e.g., mostly applicable for novice surgeons), and the frequency (e.g., number of surgeries per month) they have performed, etc. We also inquired about surgeons' training in the questionnaire. The novice subjects in our study performed the

surgery approximately fifty times, had seen the surgery performed more than fifty-five times in the last six months, and over a hundred times overall. Expert surgeons performed the surgery more than two hundred times and observed the surgery performed more than two hundred times in the last six months.

Surgery questionnaire

We require surgeons to provide detailed information regarding the surgeries used in our analysis. The questionnaire's content includes the type, location, and size of the rotator cuff tear, the difficulties of the procedure, scar tissue and cleaning amount, and suture passing and tying. All video-recorded surgeries were performed in the 'beach chair' position [16]. Recorded shoulder arthroscopy surgeries were common procedures with no unexpected complications such as a change in blood pressure during surgery, infection, or dislocation. All ten novice videos were crescent-shaped tears, while thirteen of the expert videos were crescent-shaped, and one was an L-shaped tear. The tear sizes ranged up to 3 cm. All tears were located on supraspinatus, subscapularis, and/or infraspinatus. Even though the tear's location, size, and shape play a role in the difficulty of the surgery, the procedures were randomized and not predetermined according to a surgeon's level of expertise.

Video timing analysis

Three raters performed surgery video analysis, and each of them was blinded to the scores of the other raters. All raters were given specific guidelines and instructions about each performance metric to minimize the inter-rater inconsistencies and ambiguities due to misunderstanding. Upon completing the rating, start and end timings, and scores of each task for each video from the raters were collected for further statistical analysis. An inter-rater reliability test was performed to see the degree of agreement between the raters.

Metrics

In rotator cuff repair, the major phases of surgery are diagnostic, pre-cleaning (preparation of the joint and space, debridement), anchoring (anchor placement), and suturing (passage of the suture and knot tying). A complete task tree and derivation of the phases can be seen in the hierarchical task analysis work given in [15]. The metrics for a task also include the exact definitions of every salient action's start and end times. It is essential to understand any actions performed in the tasks in order and score them. For instance, the suturing task starts when an additional portal is opened, and the needle/tool is first seen in the surgery video.

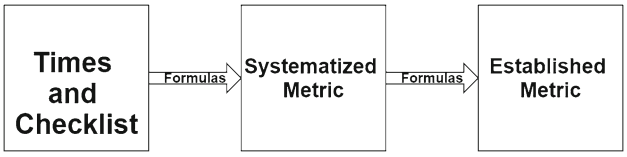

Fig. 1 Workflow illustrating the mapping of systematized metric to established metric

The rater timings are used to compute indirect measures (such as knot tying, anchoring, etc.) using our systematized metrics (see Table 1). To validate the consistency of our systematized metric results, we merged BAKSSS [12] and ASSET [11] evaluation metrics and mapped our systematized metrics to these merged/established metrics, as seen in Fig. 1. Results from each category of the established metric are compared using our systematized metrics and an inspection checklist (see Table 2). In Tables 1 and 2, the maximum score for each category is five points.

Systematized metric

We based systematized metrics on phase/task times, checklist (e.g., a task performed or not), and our metrics [15]. The systematized metrics are derived to capture the overall quality of the task, as seen in Table 1. The metrics are also useful to capture the skill level. For instance, for the proper position of the anchor, the angle (e.g., 45 degrees) and location of the anchors (e.g., uniform distribution of the anchors) are critical for achieving the best score. In the placement of the anchor task, we also considered the time spent per anchor. We hypothesized that the efficiency of the placement of the suture anchor task also reflects the skill level for efficient suture passing since sutures are fixed at the top of the anchor. Therefore, these tasks are dependent on each other and always conducted in the same order. Moreover, recall that the same arthroscopic task may be repeated multiple times, such as in the case of multiple anchor placements.

In Table 1, the time and efficiency splits were determined by analyzing each expert surgeon's videos. The expert surgeon's time for placing the anchor, knot tying, and knot safety (shorter tasks) was 30 s, while bursectomy (a more extended task) was 60 s.

Mapping to the existing metrics

Each inspection checklist item (see the first row in Table 2) is given one point. Due to having six inspection checklist items, the average is computed and normalized to the maximum score of five, depending on the number of measures for each task. *Safety* criteria correlates to a checklist of actions that surgeons need to complete successfully. These are: stopping or controlling bleeding within one minute, identifying

Table 1 Formulas for each task score in our systematized metrics. Notice that time is in seconds and threshold values for timings are mapped to Likert scale from 1 to 5 where 5 corresponds to expert performance while 1 corresponds to novice performance. These threshold values are empirically determined by an expert arthroscopy surgeon

Identification of Suture Locations	SUM (Bleeding not Stopped in 1 min, Identify Anchor Location with Needle, Ensure Anchor Drill in Right Position, Clear View of Task Items, Checking Location of Sutures, Final Check of All Sutures) where all items are 0 or 1
Bursectomy (Preclean)	$\begin{cases} 5, & preclean\,time < 540 \\ 4, & 540 \leq preclean\,time < 600 \\ 3, & 600 \leq preclean\,time < 660 \\ 2, & 660 \leq preclean\,time < 720 \\ 1, & 720 \leq preclean\,time < 780 \\ 0, & otherwise \end{cases}$
Placing of the Anchor	$\begin{cases} 5, & anchor\,time < 120 \\ 4, & 120 \leq anchor\,time < 150 \\ 3, & 150 \leq anchor\,time < 180 \\ 2, & 180 \leq anchor\,time < 210 \\ 1, & 210 \leq anchor\,time < 240 \\ 0, & otherwise \end{cases}$
Efficient Suture Passing	$\begin{cases} 5, & total\,task\,efficiency > 0.9 \\ 4, & 0.9 \geq total\,task\,efficiency > 0.81 \\ 3, & 0.81 \geq total\,task\,efficiency > 0.72 \\ 2, & 0.72 \geq total\,task\,efficiency > 0.63 \\ 1, & 0.63 \geq total\,task\,efficiency > 0.54 \\ 0, & otherwise \end{cases}$ where *task efficiency* $= \frac{Task\ time\ each\ iteration}{Task\ time\ for\ all\ iterations}$, the splits were chosen according to expert surgeon videos
Knot Tying	$\begin{cases} 5, & AVERAGE(total\,knot\,tie\,time) + (total\,knot\,cut\,time) < 150 \\ 4, & 150 \leq AVERAGE(total\,knot\,tie\,time) + (total\,knot\,cut\,time) < 180 \\ 3, & 180 \leq AVERAGE(total\,knot\,tie\,time) + (total\,knot\,cut\,time) < 210 \\ 2, & 210 \leq AVERAGE(total\,knot\,tie\,time) + (total\,knot\,cut\,time) < 240 \\ 1, & 240 \leq AVERAGE(total\,knot\,tie\,time) + (total\,knot\,cut\,time) < 270 \\ 0, & otherwise \end{cases}$
Preparation of the Footprint	$\begin{cases} 5, & postclean\,time < PFBT \\ 4, & PFBT \leq postclean\,time < PFBT + PFBT \\ 3, & PFBT + PFBI \leq postclean\,time < PFBT + PFBI * 2 \\ 2, & PFBT + PFBI * 2 \leq postclean\,time < PFBT + PFBI * 3 \\ 1, & PFBT + PFBI * 3 \leq postclean\,time < PFBT + PFBI * 4 \\ 0, & otherwise \end{cases}$ where *PFBT = Preparation of the Footprint Base Time,* and *PFBI = Preparation of the Footprint Base Interval*
Transition Speed	$\begin{cases} 5, & 0.15 > TBTSS \\ 4, & 0.2 > TBTSS \\ 3, & 0.25 > TBTSS \\ 2, & 0.3 > TBTSS \\ 1, & otherwise \end{cases}$ where *TBTSS = % Time Between Tasks in the Suturing Stage*. *Transition speed* indicates the delay between suturing tasks, the splits were chosen according to expert surgeon videos

Table 1 (continued)

Safety	$\begin{cases} 5, & 30\,seconds > TBNS \\ 4, & 60\,seconds > TBNS \\ 3, & 90\,seconds > TBNS \\ 2, & 120\,seconds > TBNS \\ 1, & 150\,seconds > TBNS \\ 0. & otherwise \end{cases}$ where *TBNS* = *Time of Bleeding Not Stopped. Safety measure*, indicates the aggregated total time of the surgeon when no action is taken during bleeding

Table 2 Formulas for established metric. These categorizations in each row and their evaluation formulas are designed and carefully evaluated based on the guidance received from expert surgeons in arthroscopic procedures

Inspection Checklist	Inspect Superior Labrum and Biceps: ISLB
	Inspect Anterior Labrum and Capsule: IALC
	Inspect Rotator Cuff Muscles: IRCM
	Inspect Glenohumeral Ligament: IGL
	Inspect Rotator Interval: IRI
	Inspect Glenoid and humeral head: IG
Knowledge of Instruments	AVERAGE (Bursectomy + Preparation of the Footprint + Efficient Suture Passing)
Field of View	(Identification of Suture Location + AVERAGE (ISBL + IALC + IRCM + IGL + IRI + IG) *5)/2
Camera Dexterity	(Number of Portals + Identifying Portals + AVERAGE (ISLB + IALC + IRCM + IGL + IRI + IG) *5 + Identification of Suture Location)/4
Instrument Dexterity	AVERAGE (Bursectomy + Preparation of the Footprint + Position of Anchor + Knot Tying)
Bi-Manual Dexterity	AVERAGE (Knot Tying + Efficient Suture Passing + Identification of Suture Location)
Efficiency	AVERAGE (Bursectomy + Preparation of the Footprint + Efficient Suture Passing + Knot Tying)
Flow of Procedure	AVERAGE (Transition Speed + Knot Tying + Efficient Suture Passing)

anchor location with a needle, ensuring the anchor drill is in the proper position, establishing a clear view with the arthroscope, checking the location of sutures (equidistance), and verifying all sutures at the end. *Field of view* correlates to the identification of suture locations and inspection checklist, while *camera dexterity* associates with the number of portals, inspection checklist, identifying portals, and suture locations. *Knowledge of instruments*, *instrument dexterity*, *bi-manual dexterity*, *and efficiency* correlate to tool efficiency. Therefore, the tool efficiency is calculated by dividing the tool's active time (a tool may be idle sometimes) inside the shoulder by the total time the tool appears in the scene. Another factor affecting *bi-manual dexterity* is the knot tying time and knot quality/safety. *The flow of procedure* correlates to the transition speed, knot tying, and efficient suture passing. *Quality of procedure* is the optimal final product with no flaws, and *autonomy* is the successful completion of the procedure without any assistance.

Automation of surgery quality metrics: a case study

Along with the proposed Systematized metric for evaluating the quality of surgery, this study also explores the possibility of using deep learning to create an automated arthroscopic video evaluation tool (AAVET). Manual assessment using the proposed metrics is a labor-intensive and time-consuming process. Therefore, automating the quantitative assessments is crucial for scaling the process of surgical evaluation. Several of the metrics within the proposed system are related to the amount of time required to complete a surgical task, with each task being associated with a set of tools. Proper autonomous evaluation of surgery necessitates some form of tool classification (as seen in Fig. 2) and scene detection from surgery videos to determine what task is being performed. Figure 2 shows the presence of the shaver tool and the electrocautery tool in two different scenes.

This study builds the foundations of an automated assessment framework by performing automatic tool classification using a subset of the surgical tools used in arthroscopic surgery. The widely used electrocautery and shaver tools were selected for initial classification studies, with more to be added in future. Deep learning has been shown to excel at

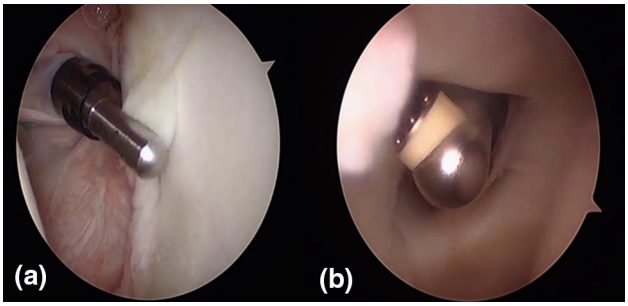

Fig. 2 **a** Shaver tool and **b** Electrocautery tool from the surgery videos to automate tool detection. AAVET automatically detects and identifies surgical instruments present in the scene. AAVET also keeps track of entrance and exit times of the instrument

complex object detection and classification tasks, and it has been used successfully in the classification of electrocautery and shaver tools [17], making it a natural fit for this study.

In our previous study [17], deep learning was used to identify the electrocautery and shaver tools in the context of individual surgical images. However, this study introduces a new deep learning model to automatically classify and subsequently determine the amount of time that the electrocautery and shaver tools appeared in the scene throughout the surgery with the goal of automating the assessment process in future. The automation of the assessment process will allow for objective and quantitative self-assessment. Training and self-assessment are expected to decrease the learning [18].

Deep learning data acquisition

Deep learning data was acquired by editing the videos used for surgery video analysis into one-second clips. First, the videos were processed to create a dataset of video clips of the shaver and electrocautery tools in the scene and clips in which no tools are present. The training dataset contained 3,086 one-second clips, with 1,936 clips of the electrocautery tool and 815 clips of the shaver tool, and 335 with no tools from 16 videos. A small validation dataset was created containing 36 video clips; 15 electrocautery tool, 11 shaver tool, and ten no tools. This set was used to check for the deep learning model's overfitting (high bias) or underfitting (high error rate). A separate test dataset of 7 videos ranging from 30 s to a few minutes was used to test the model's accuracy in a real-world scenario.

Deep learning model

For this project, we used a 3D Convolutional Neural Network (3DCNN) [19] to classify one-second clips as either containing no tools, the electrocautery tool, or the shaver tool. The model consists of three convolutional blocks followed by fully connected layers. 3DCNN architecture used for the model is illustrated in Fig. 3. For the activation function, Rectified Linear Unit (ReLU) [20] is used in hidden layers except for the output layer. The output layer uses Softmax [21] for binary classification with two output nodes where one node detects the existence/nonexistence of electrocautery and the other node detects the existence/nonexistence of the shaver tool. In this study, we used Adam [22] as the optimizer with sparse categorical cross-entropy [23] as the error/loss function to optimize the deep learning model's training process. Adam was used due to its performance over other optimizers [22]. Dropout was applied after each convolutional block and between the second and last fully connected layers to ensure that the neural network learns a more robust set of features that perform equally well with random subsets of the nodes selected. L2 regularization was applied to every convolutional and fully connected layer to combat overfitting by penalizing the weights that become too large for some set of features (see Fig. 3 for a diagram of the model).

Deep learning training

The 3DCNN was trained on a virtual machine at Google cloud. The virtual machine used an n1-standard-16 machine type, which includes 16 virtual CPUs and 60 GB of RAM. The virtual machine was also configured to use an NVIDIA TESLA P100 GPU and a 150 GB SSD. The operating system used by the virtual machine was Ubuntu 18.04 LTS. As mentioned in "Deep learning data acquisition" section, 3086 one-second clips were used as the training dataset. The model discussed in this study is a preliminary model trained for 150 epochs, with a single batch with the size of the full training dataset. All images were resized to 128 × 128x3. During the training process, a learning rate of 0.001 was used.

Results

Systematized metrics results

In every systematized metric category, expert surgeons scored better than novice surgeons, except for the position of the suture anchor, where both groups performed the same score. Table 3 shows the average systematized metric results for both expert and novice surgeons for the suturing task. Figures 4 and 5 show the box plots of each suturing task in our systematized metric. These graphs demonstrate the variation among novice surgeons.

Identification of suture locations

An essential checklist item in our metrics is identifying the anchor location with a needle before deploying the anchor.

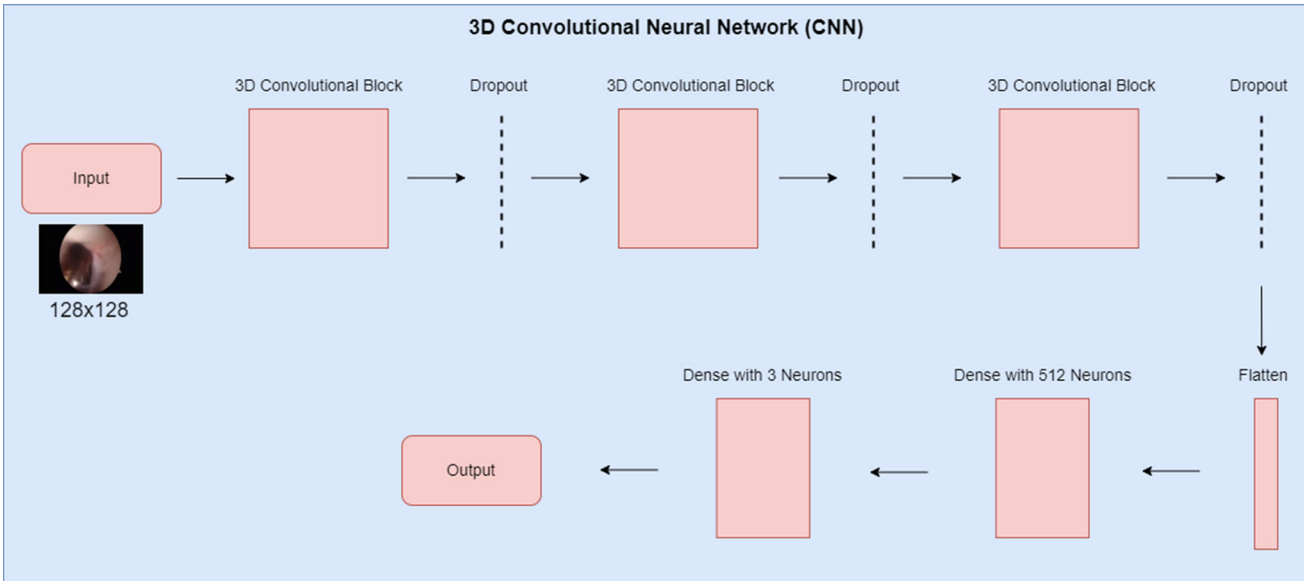

Fig. 3 Diagram of the 3D Convolution Neural Network architecture used in AAVET. The model consists of three convolutional blocks followed by fully connected layers

Table 3 Systematized metric results for each suturing task. For each task given in the first column of the table average expert and novice scores are presented. Scores are calculated using our systematized metric. Different columns indicate percentile difference between the average scores of experts against novices. For instance, for there is no difference observed between experts and novices for positioning anchors after suture anchors' locations are marked with a needle

Suturing task	Expert average (points)	Novice average (points)	Difference (%)
Identification of Suture Anchor Locations	4.64	3.5	22.8
Position of Anchor	5	5	0
Placing of the Anchor	4.57	3.1	29.4
Efficient Suture Passing	4.36	2.3	41.6
Knot Tying	4.29	3	25.8
Preparation of the Footprint	5	4.9	2
Transition Speed	3.72	2.9	36.4
Safety	4.5	2.6	38.0

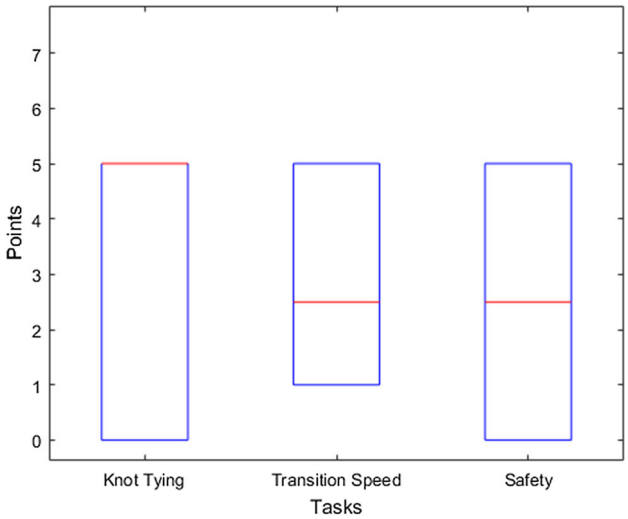

Fig. 4 Distribution of Novice performance results for *Knot Tying, Transition Speed* and *Safety* metrics. The highest possible score is 5

The expert surgeons were found to always complete this task; however, novice surgeons only completed the task in 40% of the novice videos. Also, we noticed that the expert surgeons clearly inspected all sutures before placing the next one. In two novice videos, the final check of all sutures and clear view of task items were not completed, which caused a 22.8% lower score. In all expert videos, bleeding was spotted and stopped within 60 s, which is the case only in 40% of the novice videos. The ability to establish a clear view of the tasks in the procedure was also determined in the checklist. In expert cases, the view was clearly visible during the entire surgery in each video, while the clear visibility was decreased to only 3 cases out of 10 novice videos. During the bleeding, novice surgeons proceeded at a slower pace to stop the bleeding, and in three of the videos, the novices scored zero points due to overlong bleeding.

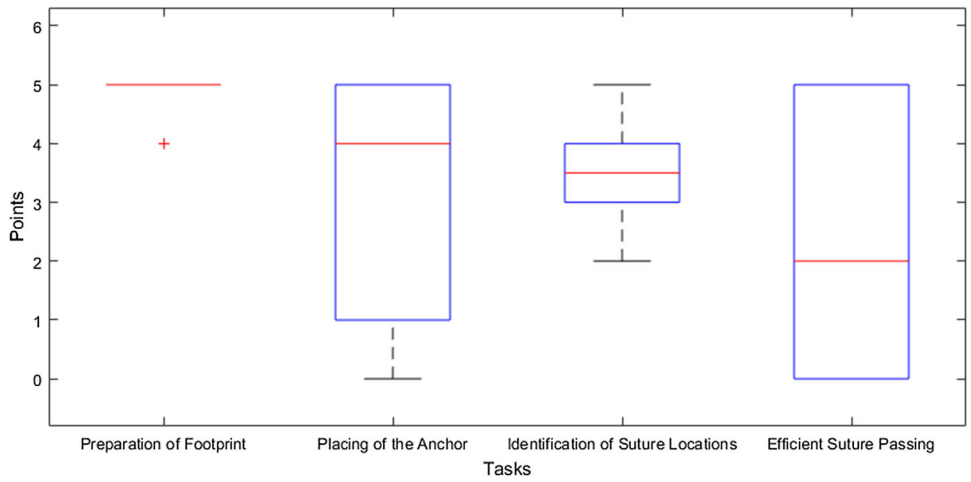

Fig. 5 Distribution of Novice performance results for *Preparation of the Footprint, Placing of the Anchor, Identification of Suture Locations* and *Efficient Suture Passing metrics*. The highest possible score is 5

Placing the anchor and knot-tying

In two of the novice videos, the anchoring task took more than four minutes due to the excessive time spent during the identification of a location and inserting the anchor in the humerus, which, as a result, caused a 29.4% lower score for novices than experts.

In the knot-tying task, we determined the type of the knot (e.g., square knot) and the knot tying time. All knots by novice surgeons were square knots, but in four novice videos, knot tying times were longer due to the knots being tied out of order.

Transition speed

Five of the novice videos received a score of one point in transition speed criteria, which suggests that additional suturing training might be necessary for novice surgeons. The effect of this training can be noticeable and reduce the transition time between sutures during the suturing task.

Established metrics results

In every established metric category, expert surgeons scored better than or the same as novice surgeons. Autonomy is the only category in which both groups performed the same score. Our results demonstrated that all the indirectly computed measures (systematized metrics) for novice and expert surgeons have notable differences. Table 4 shows the average established metric results both for expert and novice surgeons. The variation in the metric category (as seen in Figs. 6 and 7) was notably more significant for novice surgeons (min 2.93, max: 5) than comparing to expert surgeons (min 4.27, max: 5).

Table 4 Established metric result comparison for expert and novice surgeons. The first column lists the skills to be assessed based on the established metric

Established metric category	Expert average (points)	Novice average (points)	Difference (%)
Safety	5	3.7	26.0
Knowledge of instruments	5	4.2	16
Field of View	5	4.25	15
Camera Dexterity	5	4.625	7.50
Instrument Dexterity	4.62	4.5	2.4
Bi-Manual Dexterity	4.5	2.93	31.4
Efficiency	4.44	3.96	9.6
Flow of Procedure	4.27	2.73	30.8
Knowledge of Specific Procedure	5	4	20
Quality of Procedure	5	4.2	16
Autonomy	5	5	0

Safety

In the safety analysis for established metrics, two videos from novice surgeons received a score of two. In all these cases, the final verification of all sutures, clear view during the several tasks such as suturing and anchoring, and identification of anchor location with a needle were not performed, which caused a 26% lower score. On one occasion, in one of the lower-rated *safety* criteria videos, during the post-cleaning phase, the sutures were accidentally burned with the electrocautery tool.

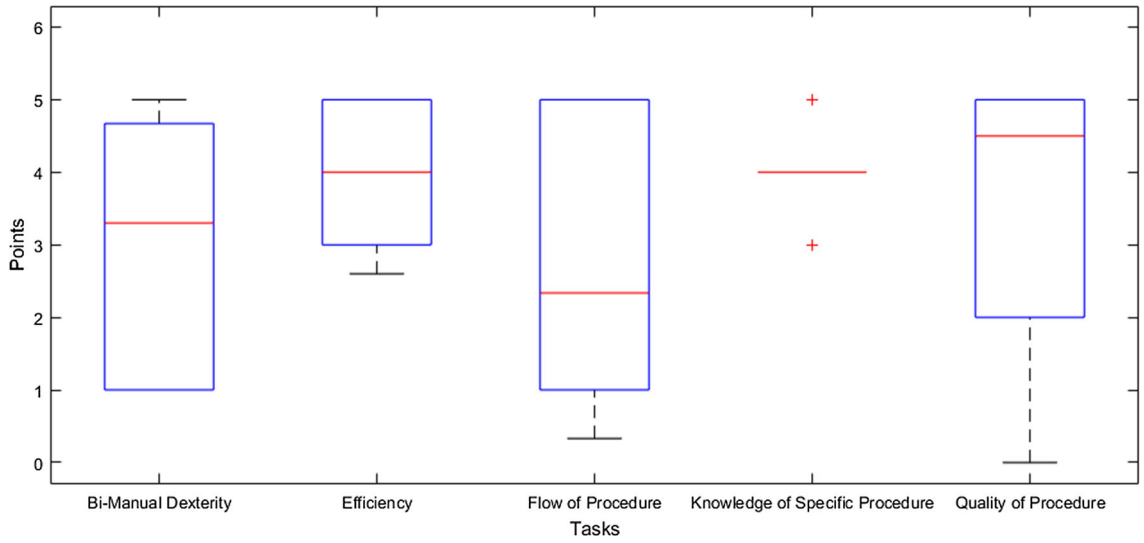

Fig. 6 Distribution of novice and expert performance results for established metric category: *Bi-Manual Dexterity, Efficiency, Flow of Procedure, Knowledge of Specific Procedure,* and *Quality of Procedure*. The highest possible score is 5

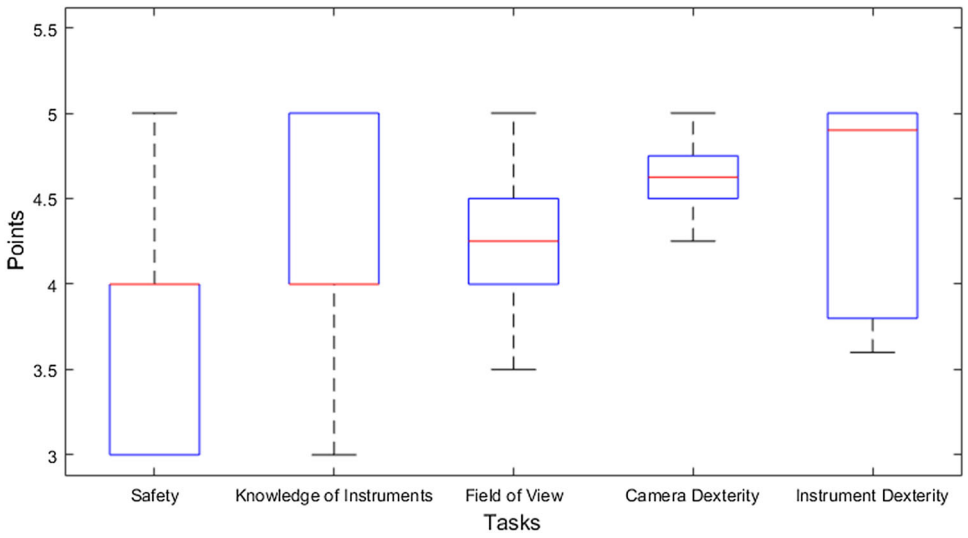

Fig. 7 Distribution of novice and expert performance results for established metric category: *Safety, Knowledge of Instruments, Field of View, Camera Dexterity,* and *Instrument Dexterity*. The highest possible score is 5

Efficiency

Efficiency had a 9.6% average score difference between the expert and novice scores. While all expert efficiency grades were consistent, out of ten novice videos, three had a task efficiency of less than 54%, which was the main contributing factor for overall lower novice scores.

We also calculated the tool efficiency of shaver and electrocautery tools for expert and novice surgeons in the cleaning phases, as shown in Fig. 8. The results showed that the average expert tool efficiency for shaver and electrocautery was 86.25% and 87.95%, respectively, while the average novice tool efficiency for shaver and electrocautery was 77.47% and 78.86%, respectively.

Bi-manual dexterity

Bi-manual dexterity had the most significant differentiation in novice and expert scores with 31.4%. This was caused by the substantial difference between experts and novices in knot tying (difference: 25.8%), efficient suture passing (difference: 41.6%), and identification of suture location (difference: 22.8%).

AAVET results

The confusion matrix of AAVET is presented in Fig. 9. This figure shows that most of the no-tool test videos were predicted correctly (precision = 0.91%). Conversely, 108

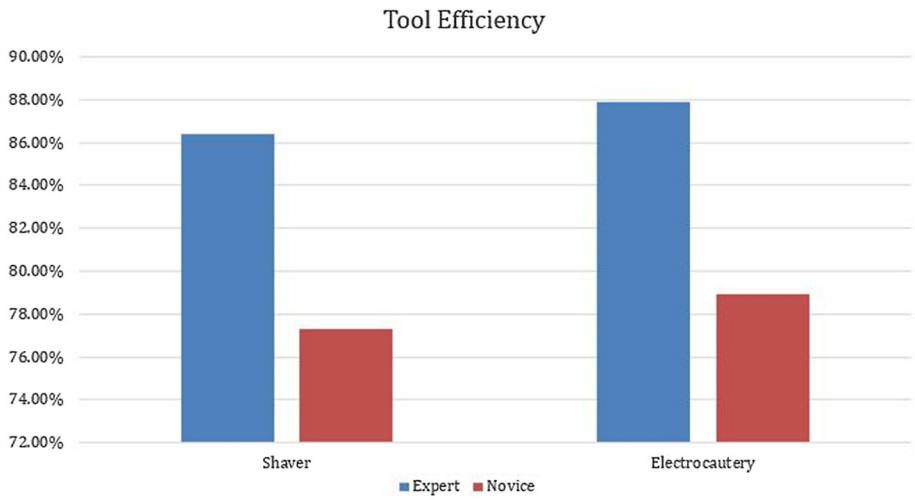

Fig. 8 Tool Efficiency comparison for expert and novice surgeons in pre- and post-cleaning phases

Predicted / Actual	No Tools	Shaver Tool	Electrocautery Tool
No Tools	243	6	18
Shaver Tool	28	543	104
Electrocautery Tool	80	172	1232

Fig. 9 Confusion Matrix of the AAVET showing the predicted and actual values of no tools, shaver tool, and electrocautery

electrocautery or shaver tools were predicted as no tool (sensitivity = 0.69%). The highest specificity and accuracy are found to be for the no-tool classification, which was 99% and 95%, respectively. Moreover, the precision of predicting the shaver tool was 80%. The rest, 20%, was incorrectly predicted as the shaver tool (sensitivity = 75%). Similar to no tool classification, the shaver tool had high specificity (92%) and accuracy (87%) results. The highest sensitivity (0.91%) is found for the electrocautery tool. It shows that 91% of electrocautery tool predictions were classified correctly over all samples that are predicted as electrocautery tool. The electrocautery tool had the lowest specificity (76%) and accuracy (85%) results, while the precision of the electrocautery tool was 83%. Accuracy, precision, F1-score, sensitivity, and specificity for each class are given in Table 5.

Discussion

Our hypothesis was that we could objectively assess arthroscopic surgery skills between the novice and the expert surgeons with minimal human input. To validate our hypothesis, we merged ASSET and BAKSS evaluation metrics and mapped our objective metrics to the merged metric. We further validated our metrics by comparing the performance of a group of novice surgeons against expert surgeons. Since our end goal is to automate the quantification and assessment of the surgeons based on their surgery performances, we also attempted to create a preliminary deep learning-based model for detecting the presence or absence of the surgical instruments along with the accurate classification of the tool.

Systematized metrics

To the best of our knowledge, there aren't any performance metrics for arthroscopic surgeries that can objectively assess surgeon performance with minimal human input. As mentioned in [24], only using checklists doesn't improve assessment validity over global rating scales such as BAKSS, but they obtain other information. Thus, we utilized checklist items and time data to derive our own systematized metrics. Our systematized metric was able to differentiate between expert and novice performances. The metrics showed that novice surgeons have the most challenging time with the

Table 5 Performance evaluation of the automated arthroscopic video evaluation tool (AAVET)

	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
No Tools	0.95	0.91	0.69	0.99	0.78
Shaver Tool	0.87	0.80	0.75	0.92	0.77
Electrocautery Tool	0.85	0.83	0.91	0.76	0.87

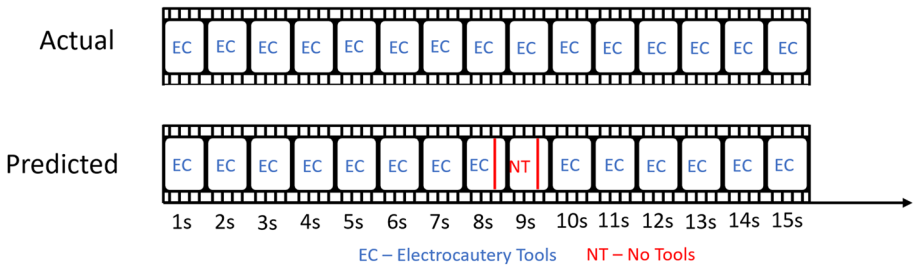

Fig. 10 Illustration of the electrocautery tool being predicted as no tool due to the electrocautery tool not being fully present in the scene

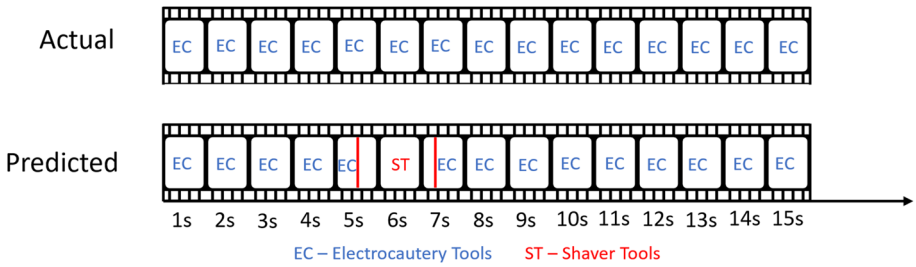

Fig. 11 Illustration of the electrocautery tool being predicted as the shaver tool due to only the backside of the electrocautery tool being visible to the camera

suture passing task, as the point difference between expert and novice average scores was 41.6%. The other two tasks where novice surgeons had the largest performance percentage difference were safety (38%) and transition speed (36.4%). Novices performed the same as experts in the positioning of the anchor task. Also, novices performed almost as well as experts in the preparation of the footprint task with only a difference of 2%. These results show that novices had the correct theoretical knowledge (anchor position and footprint preparation) while missing practical knowledge (suture passing, transition speed, and safety).

Established metrics

In the established metrics, the average expert performance for all metric categories was over 4.27 points. While for novices, the average performance score for the metric categories varied from 5 to 2.73. For seven metric categories, average expert performance was five, for novices, this number was only one which was autonomy. The two most significant differences between expert and novice performance were bi-manual dexterity (difference of 31.4%) and flow of the procedure (difference of 30.8%). These results were expected as bi-manual dexterity has been shown to be higher in experts than novices [25, 26]. Autonomy was given as 5 to everyone as all procedures were completed.

AAVET

Minimally invasive surgeries such as laparoscopy and arthroscopy help patients to recover faster and reduce blood loss. However, the lack of direct visual contact and limited real-time feedback are disadvantages of the procedure for the surgeons [27]. Computer vision methods have been used to overcome these limitations, but most studies have been for laparoscopy instead of arthroscopy [28–30]. This is due to the small field of view in the joint space and the debris obstructing the field of view [27]. To the best of our knowledge, this is the first study that predicts the tools used in arthroscopy surgery in real-time.

The end goal of automating the objective assessment metrics proposed in this study was limited by several factors while pointing to promising future advancements. When evaluating the predictive deep learning model, it became apparent that specific sets of circumstances within a video clip led to lower model performance. For instance, when a tool in the frame is heavily occluded or mostly out-of-frame, the model's accuracy decreases because the defining features of the tools are not visible. As seen in Fig. 10, the electrocautery tool can be predicted as no tool due to the tool not being fully present in the scene. In Fig. 11, only the backside of the electrocautery was present in the scene, which caused AAVET to predict it as the shaver tool. This was due to the similar nature of the backside of both tools. A possible future solution to this problem is developing an algorithm that analyzes the confidence level of the model's prediction to determine if the context of surrounding time intervals in the video should be used to make a decision (rather than a decision based solely on the current time interval). For example, if only a portion of a tool is in view within a given interval (leading to a low confidence value), the algorithm would adjust past predictions if, in a subsequent interval, a defining feature becomes visible (leading to a high confidence value). This algorithm will also need to determine if the tool has left the frame entirely (this is possible using the model's ability to predict that no tool is present) and for how long to ensure that the tool in use has not changed. Another limiting

factor in the automation case study was training data. Manually defining the ground truth for the amount of time that a tool is present in a video is laborious and limits the size of the dataset. Increasing the dataset size could lead to better results.

Additionally, we plan to expand the set of tools in future studies. Expanding the set of tools would automatically determine which part of the surgery is being performed. Combining the ability to identify more tools and an algorithm for determining what portion of the surgery is being performed would allow for initial studies on complete automation of evaluation using a subset of the proposed surgical quality metrics.

Video analysis

The limitation of the study is that we analyzed a total of 24 videos from two novices and two experts. The level of expertise varies among novice and expert surgeons. For this reason, we still need to validate the scoring metrics further and tune the automated tool detection with a more extensive study that involves more surgeons and surgeries.

Conclusion

In conclusion, we are currently developing a virtual simulator for arthroscopic rotator cuff procedures. The virtual simulator can train surgeons to perform arthroscopy and improve their surgical skills without any significant risk to the patient. The training module requires a scoring metric to give constant feedback to the operator. In this study, we performed a preliminary construct validation study for the proposed scoring metrics for shoulder arthroscopy and arthroscopic rotator cuff repair surgery. This metric can be used to give quantitative feedback to trainees. Due to our metrics being specific to arthroscopic rotator cuff repair surgery, it can also be incorporated as a performance evaluation to any VR-based arthroscopic rotator cuff repair surgery simulators.

Acknowledgements The authors would like to thank Seth Cooper-Baer and Mustafa Tunc for their contributions to this publication. This publication was made possible by the Grant NIH/NIAMS R44AR075481-01. This project was also supported by the Arkansas INBRE program (NIGMS, P20 GM103429), NIH/NCI 5R01CA197491, and NIH/NHLBI NIH/NIBIB 1R01EB025241, R56EB026490.

Funding This project was made possible by the Arkansas INBRE program, supported by a grant from the National Institute of General Medical Sciences (NIGMS), P20 GM103429 from the National Institutes of Health (NIH). This project was also supported by NIH/NIAMS R44AR075481-01, NIH/NCI 5R01CA197491, and NIH/NHLBI NIH/NIBIB 1R01EB025241, R56EB026490.

Declarations

Competing interests The authors declare that they have no competing interests in regard to this study.

References

1. Day MA, Westermann RW, Bedard NA, Glass NA, Wolf BR (2019) Trends associated with open versus arthroscopic rotator cuff repair. HSS J 15:133–136
2. Wang EE, Vozenilek JA, Flaherty J, Kharasch M, Aitchison P, Berg A (2007) An innovative and inexpensive model for teaching cricothyrotomy. Simul Healthc 2:25–29
3. Pettineo CM, Vozenilek JA, Wang E, Flaherty J, Kharasch M, Aitchison P (2009) Simulated emergency department procedures with minimal monetary investment: cricothyrotomy simulator. Simul Healthc 4:60–64
4. Cho J, Kang GH, Kim EC, Oh YM, Choi HJ, Im TH, Yang JH, Cho YS, Chung HS (2008) Comparison of manikin versus porcine models in cricothyrotomy procedure training. Emerg Med J 25:732–734
5. Aggarwal R, Ward J, Balasundaram I, Sains P, Athanasiou T, Darzi A (2007) Proving the effectiveness of virtual reality simulation for training in laparoscopic surgery. Ann Surg 246:771–779. https://doi.org/10.1097/SLA.0b013e3180f61b09
6. Jamal MH, Rousseau MC, Hanna WC, Doi SA, Meterissian S, Snell L (2011) Effect of the ACGME duty hours restrictions on surgical residents and faculty: a systematic review. Acad Med 86:34–42
7. Fried MP, Satava R, Weghorst S, Gallagher A, Sasaki C, Ross D, Sinanan M, Cuellar H, Uribe JI, Zeltsan M, Arora H (2005) The use of surgical simulators to reduce errors. In: Henriksen K, Battles JB, Marks ES, Lewin DI (eds) Advances in patient safety: from research to implementation (volume 4: programs, tools, and products). Agency for Healthcare Research and Quality (US), Rockville
8. Seymour NE, Gallagher AG, Roman SA, O'brien MK, Bansal VK, Andersen DK, Satava RM (2002) Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 236:458–464
9. Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P (2004) Randomized clinical trial of virtual reality simulation for laparoscopic skills training. Br J Surg 91:146–150
10. Rosser JC, Rosser LE (1960) Savalgi RS (1997) Skill acquisition and assessment for laparoscopic surgery. Arch Surg Chic Ill 132:200–204
11. Koehler RJ, Amsdell S, Arendt EA, Bisson LJ, Braman JP, Butler A, Cosgarea AJ, Harner CD, Garrett WE, Olson T, Warme WJ, Nicandri GT (2013) The Arthroscopic Surgical Skill Evaluation Tool (ASSET). Am J Sports Med 41:1229–1237. https://doi.org/10.1177/0363546513483535
12. Olson T, Koehler R, Butler A, Amsdell S, Nicandri G (2013) Is there a valid and reliable assessment of diagnostic knee arthroscopy skill? Clin Orthop 471:1670–1676. https://doi.org/10.1007/s11999-012-2744-2
13. Bayona S, Akhtar K, Gupte C, Emery RJ, Dodds AL, Bello F (2014) Assessing performance in shoulder arthroscopy: the imperial global arthroscopy rating scale (IGARS). J Bone Joint Surg Am 96:e112
14. Benner P (1982) From novice to expert. Am J Nurs 82:402–407
15. Demirel D, Yu A, Cooper-Baer S, Dendukuri A, Halic T, Kockara S, Kockara N, Ahmadi S (2017) A hierarchical task analysis of shoulder arthroscopy for a virtual arthroscopic tear diagnosis and evaluation platform (VATDEP). Int J Med Robot 13:e1799. https://doi.org/10.1002/rcs.1799

16. Peruto CM, Ciccotti MG, Cohen SB (2009) Shoulder arthroscopy positioning: lateral decubitus versus beach chair. Arthrosc J Arthrosc Relat Surg 25:891–896
17. Palmer B, Sundberg G, Dials J, Karaman B, Demirel D, Abid M, Halic T, Ahmadi S (2020) Arthroscopic Tool Classification using Deep Learning. In: Proceedings of the 2020 the 4th International Conference on Information System and Data Mining. pp 96–99
18. Thompson BM, Rogers JC (2008) Exploring the learning curve in medical education: using self-assessment as a measure of learning. Acad Med 83:S86–S88
19. Ji S, Xu W, Yang M, Yu K (2012) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35:221–231
20. Ide H, Kurita T (2017) Improvement of learning for CNN with ReLU activation by sparse regularization. In: 2017 International joint conference on neural networks (IJCNN). IEEE, pp 2684–2691
21. Wang M, Lu S, Zhu D, Lin J, Wang Z (2018) A high-speed and low-complexity architecture for softmax function in deep learning. In: 2018 IEEE Asia Pacific conference on circuits and systems (APCCAS). IEEE, pp 223–226
22. Zhang Z (2018) Improved adam optimizer for deep neural networks. In: 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS). IEEE, pp 1–2
23. Kakarla J, Isunuri BV, Doppalapudi KS, Bylapudi KSR (2021) Three-class classification of brain magnetic resonance images using average-pooling convolutional neural network. Int J Imaging Syst Technol 31:1731–1740. https://doi.org/10.1002/ima.22554
24. Insel A, Carofino B, Leger R, Arciero R, Mazzocca AD (2009) The development of an objective model to assess arthroscopic performance. JBJS 91:2287–2295
25. Koskinen J, Huotarinen A, Elomaa A-P, Zheng B, Bednarik R (2021) Movement-level process modeling of microsurgical bimanual and unimanual tasks. Int J Comput Assist Radiol Surg 2021:1–10
26. Narazaki K, Oleynikov D, Stergiou N (2007) Objective assessment of proficiency with bimanual inanimate tasks in robotic laparoscopy. J Laparoendosc Adv Surg Tech 17:47–52
27. Jonmohamadi Y, Takeda Y, Liu F, Sasazawa F, Maicas G, Crawford R, Roberts J, Pandey AK, Carneiro G (2020) Automatic segmentation of multiple structures in knee arthroscopy using deep learning. IEEE Access 8:51853–51861
28. Alshirbaji TA, Jalal NA, Möller K (2018) Surgical tool classification in laparoscopic videos using convolutional neural network. Curr Dir Biomed Eng 4:407–410
29. Jaafari J, Douzi S, Douzi K, Hssina B (2022) The impact of ensemble learning on surgical tools classification during laparoscopic cholecystectomy. J Big Data 9:1–20
30. Wang S, Raju A, Huang J (2017) Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos. In: 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017). IEEE, pp 620–623