International Journal of Computer Assisted Radiology and Surgery (2020) 15:2079–2088
https://doi.org/10.1007/s11548-020-02269-x

ORIGINAL ARTICLE

Bidirectional long short-term memory for surgical skill classification of temporally segmented tasks

Jason D. Kelly[1] · Ashley Petersen[2] · Thomas S. Lendvay[3] · Timothy M. Kowalewski[1]

Received: 13 March 2020 / Accepted: 23 September 2020 / Published online: 30 September 2020

Abstract

Purpose The majority of historical surgical skill research typically analyzes holistic summary task-level metrics to create a skill classification for a performance. Recent advances in machine learning allow time series classification at the sub-task level, allowing predictions on segments of tasks, which could improve task-level technical skill assessment.

Methods A bidirectional long short-term memory (LSTM) network was used with 8-s windows of multidimensional time-series data from the Basic Laparoscopic Urologic Skills dataset. The network was trained on experts and novices from four common surgical tasks. Stratified cross-validation with regularization was used to avoid overfitting. The misclassified cases were re-submitted for surgical technical skill assessment to crowds using Amazon Mechanical Turk to re-evaluate and to analyze the level of agreement with previous scores.

Results Performance was best for the suturing task, with 96.88% accuracy at predicting whether a performance was an expert or novice, with 1 misclassification, when compared to previously obtained crowd evaluations. When compared with expert surgeon ratings, the LSTM predictions resulted in a Spearman coefficient of 0.89 for suturing tasks. When crowds re-evaluated misclassified performances, it was found that for all 5 misclassified cases from peg transfer and suturing tasks, the crowds agreed more with our LSTM model than with the previously obtained crowd scores.

Conclusion The technique presented shows results not incomparable with labels which would be obtained from crowd-sourced labels of surgical tasks. However, these results bring about questions of the reliability of crowd sourced labels in videos of surgical tasks. We, as a research community, should take a closer look at crowd labeling with higher scrutiny, systematically look at biases, and quantify label noise.

Keywords Surgical skill · Crowd sourcing · Bidirectional LSTM · Surgical technical skill · Machine learning

Introduction

Computationally assessing the skill of a surgeon in an objective manner using tool motion has proven a complex problem with many challenges. Previous research has relied mostly on summary performance metrics from kinematic data [1–3]. Unfortunately, these metrics typically failed to completely discriminate novices from experts, that is to never misclassify "obvious" novices vs. "obvious" experts—the so-called minimally acceptable classifier (MAC) criterion [4]. Recent advances in machine learning techniques have expanded the possibility of using time series datasets to evaluate surgeries and surgical tasks [5]. These techniques may aid in successfully classifying obvious novices and experts, as this is the next incremental step to take in correctly discriminating skill groups.

The de facto gold standard for determining the level of skill in a surgical performance is video-based evaluation by an expert surgeon [6]. Skills assessment is normally evaluated with the use of one of several possible established assessment schemes that utilize Likert-scale scoring of sub-domains relevant to skill, such as bimanual dexterity, or tissue handling. An example of one of these assessment methods is the Global Operative Assessment of Laparoscopic Skills (GOALS) evaluation scheme, which evaluates surgeons on a

✉ Jason D. Kelly
kell1917@umn.edu

1 Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, USA

2 Division of Biostatistics, University of Minnesota, Minneapolis, MN, USA

3 Department of Urology, Seattle Children's Hospital, Seattle, WA, USA

scale of 1–5 in four categories: depth perception, bimanual dexterity, efficiency, and tissue handling [7]. This is both a time-consuming and laborious process for experienced surgeons whose valuable time could rather be spent performing life-saving surgeries. In recent years, it was found that crowd-based evaluation of surgical performance is able to achieve almost the same standard of accuracy in predicting technical skill in surgery [8]. This may still be a subjective measure of technical skill and prone to bias [9]. Crowd-sourced labels of surgical task videos are comparable, but not equivalent to a ground truth classification of surgical expertise. An objective technique to computationally evaluate the technical skill of a surgeon would be beneficial.

There have been multiple approaches to using kinematic data in the past for surgical skill research. One such method uses surgical procedures to both learn a vocabulary of common surgical activities and frequent patterns which were used for hierarchical clustering based off the procedural meaning to classify expertise [10]. Similar studies attempted to extract and recognize surgical gestures or phases using a variety of methods [11,12]. More machine learning oriented approaches have also been used, by using deep convolution neural networks as well as video-based deep ranking methods [13–15]. Other methods have used the frames of video data to obtain temporal information, through the use of machine learning methods by performing clustering calculations on frames as time progresses [16]. No previous work to the authors knowledge has analyzed the temporal information given from kinematic data alone, as opposed to using video or using summary performance metrics.

Long short-term memory networks (LSTMs) are an adaptation of recurrent neural networks which are capable of analyzing past events in a time series to learn how they might affect a present time index [17]. This is possible through a series of gates in the architecture of the network which hypothesize, learn, and forget predictions by deducing what information to ignore and which to emphasize in the training process. These networks have been further improved through the implementation of bidirectional networks [18]. These behave by training a network in the forward direction, while also training a network in the reverse direction, thereby connecting two hidden layers in different directions to create one output from twice as much information, and allowing the LSTM to use this increase in information to achieve better results.

The objective of this investigation was to evaluate the feasibility of temporal segmentation of surgical tasks for quantifying the skill of a surgeon, and whether crowd-sourced labels may be used to accurately train such a technique. The hypothesis of this work is that experts do not behave in an expert-like manner throughout the entirety of a task, and likewise for novices, but instead that the main factor in deciding upon a surgeon's overall technical skill is the number of expert-like to novice-like segments in footage of a surgical task, and that bidirectional LSTMs have the ability to learn this information from kinematic tool motion data.

Methods

Dataset

This study used the Basic Laparoscopic Urologic Study (BLUS) dataset, described in detail in [19], with a summary re-iterated here for convenience. This dataset arose from a gap in the field, in which no educational surgical certification process existed for urologic surgery, as opposed to how the Fundamentals of Laparoscopic Surgery (FLS) exists for general surgical procedures [20–22]. The BLUS training curriculum aimed to address urology appropriate skills improvement by recording video performances in an initial validation project of over 450 videos [23].

This dataset contains 454 videos of surgical performances consisting of four surgical tasks (110 peg transfer, 110 pattern cutting, 115 suturing, 119 clipping), which are performed by medical students, urology residents, fellows, and faculty surgeons from eight academic urology training centers in the USA [24]. Each trial of a surgeon performing one of the four tasks was recorded at 30 fps with a fixed camera-position of the laparoscopic tools interacting with the training field. Each trial additionally has kinematic data, sampled at 30 Hz, logging the tooltip positions, grasping force, and the jaw angles during the performance, as well as demographic information for each performer being obtained. A GOALS score was obtained for each video via crowd evaluation, and an expert evaluation was obtained for a subset of videos which were randomly selected.

Previous research regarded suturing the most clinically relevant of the four tasks, as these performances require the mastering of needle and suture handling which are more similar to what is encompassed in real surgery vs. transferring synthetic blocks or gauze cutting. However, all four tasks were used in this study in an attempt to provide a classification scheme which can successfully separate experts from novices. Here, a novice was defined solely as an "obvious novice", or someone who should never be allowed to operate and experts solely as "obvious experts", or surgeons who should never be disqualified from operating [4]. An obvious expert was chosen such that the performer was in the top 15% of previously obtained GOALS scores for that particular BLUS task. The obvious novices were chosen in the same fashion such that they were in the bottom 15% of these domains. These sets of skill levels were chosen for model training due to the large differences in skill, allowing a machine learning model to learn characteristics which most

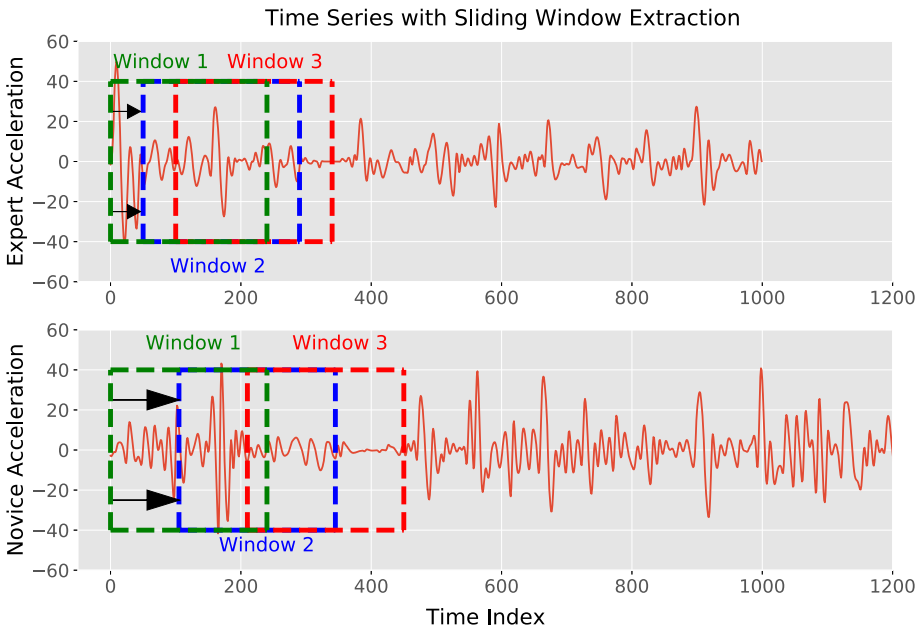


Fig. 1 Example left-hand tool acceleration time series with the sliding time window extraction method applied to both a novice series and an expert series. A novice performance is indexed with a window overlap of 105 time indices, and an expert performance is indexed with a window overlap of 50 time indices

differentiate the two labels. This method aimed to provide two well-discriminated clusters of skill levels that should demand no misclassifications.

Skill classification from temporal segmentation

Data partitioning

In the preliminary analysis, a sliding window data partitioning technique was used in which windows of varying sizes with varying overlapping lengths were tested to find the optimal parameters. The network was trained by using the novices and experts from the top and bottom 15% in each of the four tasks such that the classifier could more easily find separating features to define each class. Using intermediate performers (performers who are neither obvious experts or obvious novices) would not allow the model to be able to discriminate performances as easily, by not being able to focus model parameters on learning characteristics of novices and experts. These values are shown in Table 2, where the minimum possible GOALS score is 4, and the maximum is 20. These labels are then converted into binary variables based on whether they are considered an expert or novice. This results in 16 expert videos and 16 novices videos for peg transfer, suturing, and cutting tasks, and 16 expert videos and 17 novice videos for the clipping task.

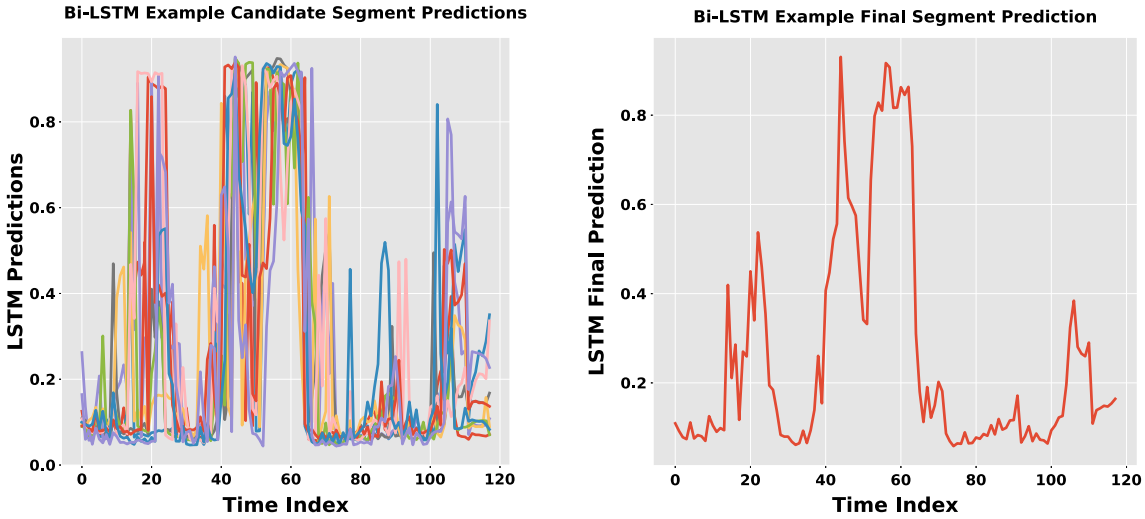
Window parameter selection

Due to the relatively small dataset resulting from subselecting only obvious novices vs. obvious experts, a stratified cross-validation scheme was computed on the experts and novices, in which two performances were left out of training. Each of these groups of two performances consisted of an expert and a novice, in an effort to keep the training response values at equal proportions and prevent overfitting of the dataset. In addition, to avoid oversampling of novice performances, as novices consisted of about 70% of the total data, being that novices usually take longer to perform, the expert-labeled performances were sampled with a step size half as large as the novice step size, to simulate more training data. This sliding window and the specified sampling control technique are illustrated in Fig. 1.

LSTM parameters and architecture

The bidirectional LSTM network consisted of a one-layer 32-unit bidirectional cell, followed by a 50% dropout layer, and a sigmoid activation function. L2 regularization was also used to further prevent overfitting. Each network was trained to 100 epochs using a binary cross-entropy loss function, and the Adam optimizer. Each segment of a test performance was fed to the network and evaluated. After all segments were evaluated, the LSTM outputs a probability of how likely a segment is to be expert-like versus novice-like. The task as a whole was considered an expert-level performance if the mean of the predictions resulted in a prediction of greater than 0.50, and vice versa.

After the initial results were obtained from using experts and novices as training and test data only, additional testing was done using intermediates as test data on these models. During this trial, the same training technique was used as discussed previously, this time alternatively using intermediate-level performers as the test data. This training



(a) 10 Test Predictions for a suturing performance.

(b) Final averaged prediction with outliers removed.

Fig. 2 Process of computing final suturing performance prediction. All ten candidate predictions of a performance are averaged at each index to compute the final prediction

was repeated ten times per validation set, with different random number seeds being assigned at the initialization of each trial. For each intermediate performance prediction, the ten different predictions were averaged for each segment of the performance. An example of this is shown in Fig. 2.

Using these intermediate performance prediction values, it was now possible to get a rough approximation of the correlation between the previously obtained GOALS scores and the average of the performance's LSTM predictions. These values were obtained for the experts and novices, as well as the intermediate performers.

Crowd reassessment

In order to delve further into the reliability of crowd-sourced labels and test whether scores would be reliable, the authors chose to reassess the surgical videos which were misclassified by the LSTM, to learn whether there was a disagreement between new and previous crowd scores, i.e., do crowds agree more with the current LSTM ranking, or with previous crowd scores. As suturing and peg transfer tasks are agreed to be the two most separable and clinically relevant tasks from the BLUS dataset, these two were chosen to have their misclassified obvious expert and novice videos reassessed by crowd workers. In addition, 5 obvious experts and 5 obvious novices from each task were randomly selected for reassessment, to validate whether any possible discrepancy was only occurring for misclassified videos, or for all videos.

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid an average of $0.50 to watch and evaluate the short video of the surgical task. The goal was to compensate the evaluators at a rate of approximately $10/h. A user interface was created which asked crowds to rate the

Table 1 Main hyperparameters tested during model evaluation in the "Skill classification from temporal segmentation" section, with most optimal results in bold

Dropout	Batch size	L2 Regularization
0.1	120	0.1
0.2	**240**	**0.2**
0.4	360	0.3
0.5	720	0.4

Table 2 Accuracy results computed by the bidirectional LSTM for each BLUS task. Additionally, the threshold scores for obvious experts and novices in each of the tasks are displayed

Task	Accuracy	Expert threshold	Novice threshold
(a) Accuracy for each BLUS task's experts and novices, from the "Skill classification from temporal segmentation" section			
Suturing	**96.88%**	15.40+	9.47−
Peg Transfer	87.50%	15.89+	10.69−
Cutting	87.50%	16.14+	10.03−
Clipping	73.33%	16.86+	12.28−

Task	Novice-specific acc.	Expert-specific acc.
(b) Novice- and expert-specific accuracy for each BLUS task, also known as sensitivity and specificity		
Suturing	**100%**	**93.75%**
Peg Transfer	**93.75%**	81.25%
Cutting	87.50%	87.50%
Clipping	68.75%	87.50%

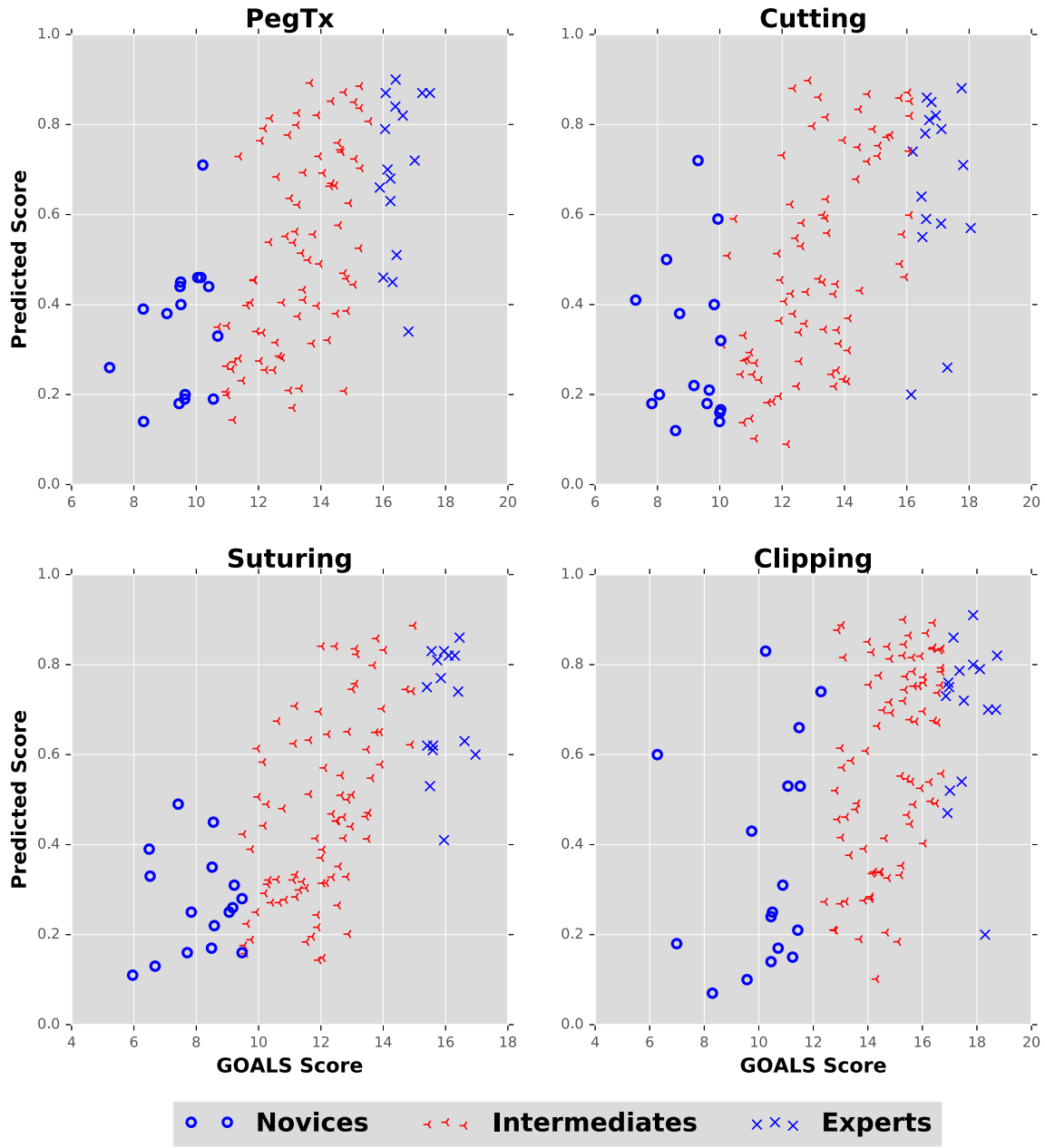


Fig. 3 The LSTM's prediction for all performances from each of the tasks in the BLUS dataset. The LSTM was only trained on experts and novices (summary score labels with cross-validation), from the "Skill classification from temporal segmentation" section

skill level of the performance using the GOALS assessment method, as obtained previously. These videos were given to crowds one at a time, using 40 crowd workers per video. The mean of the ratings for each video was then taken and compared to the previously obtained ratings for the misclassified videos, to find the level of agreement.

Results

Skill classification from temporal segmentation

After performing a grid search with different hyperparameter values, the network was found to perform optimally with a

Table 3 Correlation coefficients for the relationship between GOALS scores and predicted scores from the LSTM for every performance from each of the four BLUS tasks, including intermediate-level performers, from the "Skill classification from temporal segmentation" section

Task	Spearman	Pearson
Suturing	0.76	**0.86**
Peg Transfer	0.72	0.76
Cutting	0.61	0.69
Clipping	0.60	0.63

window sample size of 240 time indices, approximately equal to 8 s, with a 3.5 s overlap at each window for novices and approximately a 1-second overlap for expert performances, which results in the number of samples from experts and novices to be roughly equal. Most of the values tested are shown in Table 1.

Table 2 illustrates the accuracy of these networks in classifying overall skill in surgical task settings, achieving over 96.88% accuracy for the suturing task, which is usually seen as the most clinically relevant task. Table 2 also reports the expert- and novice-specific accuracies, showing that all of the novices for suturing were correctly classified. Of the 32 suturing videos labeled as experts or novices, only one was mislabeled. By getting the average of all predicted segments in a specific performance, the algorithm results in a number between 0 and 1, codifying the LSTM's prediction of skill for the performance. These predictions can then be used to arrive at a correlation coefficient signifying the degree of correctness in the neural network at evaluating intermediate performers in addition to experts and novices, when compared to crowds. The expert and novice classifications are combined with intermediate performer classifications in Fig. 3 for brevity. Correlation values for each task are in Table 3. Suturing had the highest Pearson correlation coefficient with a value of 0.86.

Given the high accuracy of classification using this method, it appears the hypothesis that performance metrics may not be consistent throughout entire procedures may be true. Specifically, Fig. 2b illustrates an example time series prediction which appears to exhibit periods of both novice-like and expert-like performance. This information would not be available from summary performance metrics.

The suturing task was the most accurate of the four tasks in the BLUS dataset. These were twelve randomly selected suturing performances which were rated by faculty surgeons. As faculty surgeon review is the gold standard for the field, comparing the LSTM's predictions to the gold standard could provide more information for the accuracy of the method. Figure 4 shows the faculty scores plotted against the skill predictions obtained from the bidirectional LSTM. As can be seen, there is a strong positive correlation, which has a Spearman coefficient of 0.89 and 91.67% accuracy.

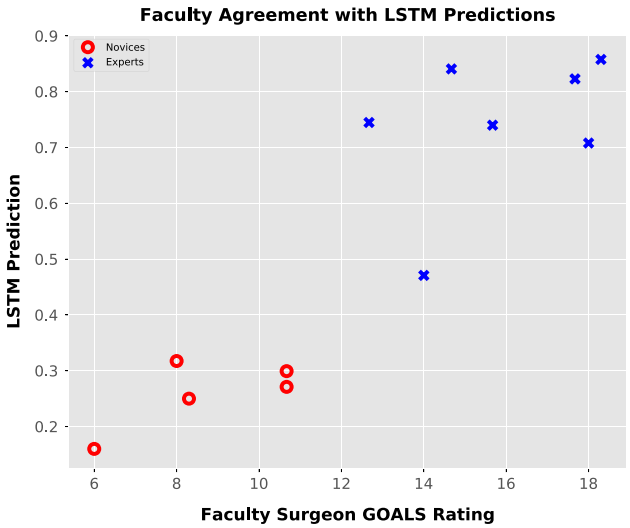


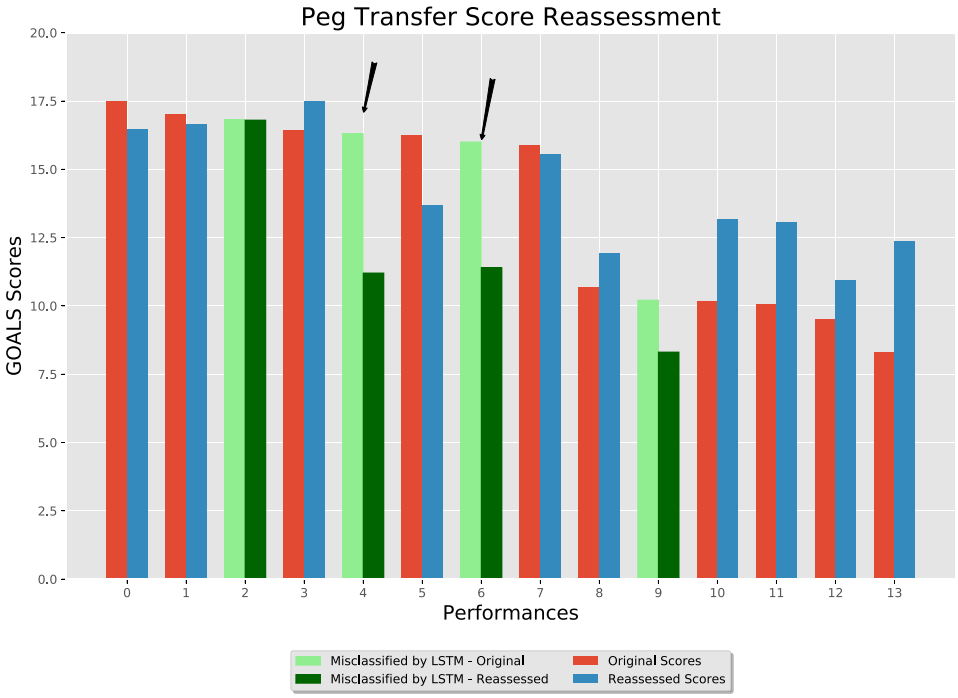
Fig. 4 The LSTM predictions trained on crowd scores, for performances of suturing tasks as rated by faculty surgeons, which has a Spearman correlation of 0.89, from the "Skill classification from temporal segmentation" section

Crowd reassessment

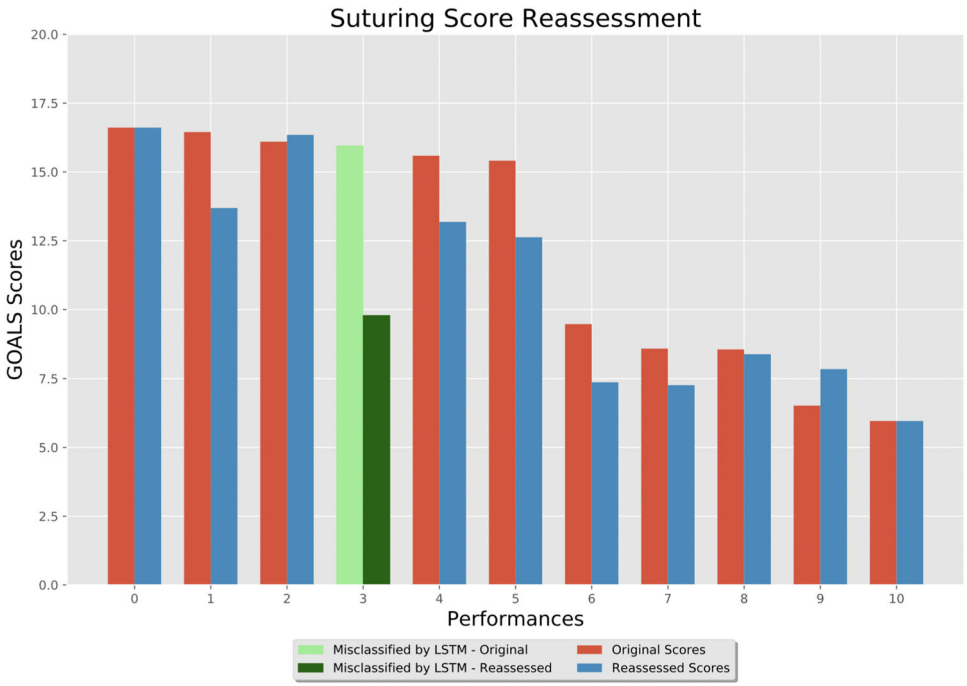
The suturing and peg transfer tasks were the two tasks with both high levels of classification accuracy and having a stronger correlation between scores and prediction levels, with a total of five misclassifications among the two tasks. The performances which were misclassified by the algorithm, which was trained on previously obtained crowd scores, were re-assessed by crowds. In addition to those performances, 5 randomly selected obvious novices and 5 randomly selected obvious experts were additionally chosen for reassessment.

Surprisingly, the range in the reassessed scores was quite lower than in the original assessments. This could have been caused from a variety of reasons, such as the newer user interface used as well as having had each video individually evaluated separately compared to the previous method in which videos were evaluated in batches. However, the reassessed ratings do still have a general agreement in ranking of performances as the original scores. Interestingly, if the new evaluations are normalized to be in the range of the old scores (5.96–16.61 for suturing and 8.31–17.5 for peg transfer, compared to the new scores having 12.07–16.09 for suturing and 13.99–16.48 for peg transfer), 3 of the 5 performances which were misclassified by the LSTM were reassessed to no longer be classified as having the skill level initially given to that performance, as shown in Figs. 5 and 6. Figure 6 illustrates the propensity of the reassessed crowd scores to agree with previously obtained scores or

Fig. 5 Reassessment of misclassified suturing and peg transfer performances suggest crowds agree more with LSTM than previous crowd ratings, from the "Crowd reassessment" section



(a) GOALS scores of misclassified peg transfer tasks before and after reassessment, as well as 5 obvious experts and 5 obvious novices, chosen randomly. Two of the 4 misclassified performances were reassessed to be more in line with what the LSTM predicted, indicated by the arrows.



(b) GOALS scores of misclassified suturing tasks before and after reassessment, as well as 5 obvious experts and 5 obvious novices, chosen randomly. The only misclassified performance was reassessed to be more in line with what the LSTM predicted.

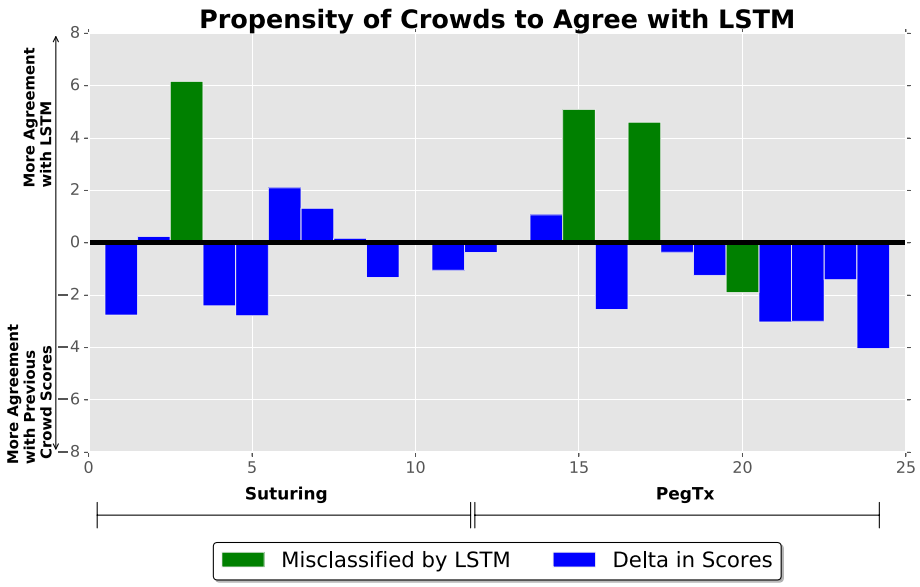


Fig. 6 A bar plot of the magnitude of attenuation of the crowds to agree more with the LSTM on score reassessment, than with the original scores obtained

to agree with the LSTM, based on the skill level predicted by the algorithm, and whether there was a misclassification. These figures suggest that the LSTM's classifications for those 3 performances could have been more accurate than the original ratings given to those performers. From this, it appears that current crowds agree more with the LSTM score than their own previous score. However, because there is no definitive ground truth, it is unclear which approach is more accurate. Regardless, this result does raise questions about the reliability of crowd-sourced labels as giving a reasonable quantitative measure of a surgeon's expertise.

Conclusion

The proposed method of evaluating technical skill using a bidirectional LSTM displayed an ability to correctly classify expert and novice surgical performances in a discriminative manner. This shows that possible future iterations of machine learning methods which track and predict temporal kinematic data may be able to improve and inform surgeons to their skill at different time points of a surgical task.

The results from the bidirectional LSTM ("Skill classification from temporal segmentation" section) provide evidence in support of our hypothesis that novice and expert surgeons do not exhibit expert performance metrics continuously throughout a task (and similarly for obvious novices), as evidenced in Fig. 2b. This suggests that temporal segmentation of kinematic tool motion analysis could provide more informative feedback of the skill of a surgeon, as compared to static summary performance metrics. The results, which also show good correlations between predicted score and actual score for several intermediate level performers, provide some additional evidence against overfitting, since these tasks were held out during the training phase. It is likely that the lower accuracy for cutting and clipping tasks (and to a lesser degree, peg transfer) was due to those movements being less repetitive and therefore harder for a model to generalize. However, these are also the two tasks which are considered less clinically relevant of the four BLUS tasks.

Other popular methods of evaluating surgical skill such as computer vision algorithms which train on the frames of video, or popular automated performance metrics [2] could possibly be combined with this technique to create an even better classification model. This would further alleviate concerns about tool motion alone lacking important context that is still present in the video. Future iterations could enable giving correct predicted ratings of a performance, even during the surgical task performance, leading to near real-time skill feedback, by accurately assessing small sub-task-level segments of operations.

This study included some limitations. First and foremost, these tasks are simulated procedures, and the proposed algorithms and techniques may perform differently on real surgeries. We intend to test these hypotheses in the future on robot-assisted surgical data obtained from practicing surgeons. The proposed method only analyzes tool motion data, which may not contain sufficient data required for complete skill classification [25]. The scores of the model testing on intermediate performances were included to illustrate the need for future iterations to improve upon the method, so that intermediate performers may be successfully classified. The need for the reassessed scores to be normalized in order to have a comparison of crowd scores could be due to the techniques used in obtaining crowd evaluations which were different than the original evaluation methods. The authors

acknowledge the results from the crowd reassessment could be in part due to differences in the user interface of the evaluation web pages used for the two different occurrences of testing, although this serves as a further argument for not equating crowd-sourced labels as a label-noise-free objective ground truth for surgeon skill. Future work should take a closer look at crowd labeling with higher scrutiny and systematically look at biases and quantifying label noise.

Future work includes A/B testing and statistical analysis on the different user interfaces used during the reliability of crowd ratings due to different user interfaces, and the original crowd ratings received. Future work will further investigate what may bias crowds to rate surgical skill more highly in certain layouts, combinations of assessed videos, or evaluation tools.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

Ethical standard All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Funding This work was supported, in part, by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-15-2-0030, the National Science Foundation M3X CAREER grant under Award No. 1847610, as well as the National Institutes of Health's National Center for Advancing Translational Sciences, Grant UL1TR002494. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense, the National Science Foundation, or the National Institutes of Health's National Center for Advancing Translational Sciences.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Balasubramanian S, Melendez-Calderon A, Burdett E (2012) A robust and sensitive metric for quantifying movement smoothness. IEEE Trans Biomed Eng 59(8):2126–2136
2. Hung A, Chen J, Che Z, Nilanon T, Jarc A, Titus M, Oh PJ, Gill IS, Liu Y (2018) Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. J Endourol 32(5):438–444
3. Kowalewski TM, White LW, Lendvay TS, Jiang IS, Sweet RS, Wright A, Hannaford B, Sinanan MN (2014) Beyond task time: automated measurements augments fundamentals of laparoscopic skills methodology. J Surg Res 192(2):329–338
4. Dockter R, Lendvay TS, Sweet RM, Kowalewski TM (2017) The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. Int J Comput Assist Radiol Surg 12:1151–1159
5. Lin HC, Shafran I, Murphy TE, Okamura AM, Yuh DD, Hager GD (2005) Automatic detection and segmentation of robot-assisted surgical motions. In: Duncan JS, Gerig G (eds) Medical image computing and computer-assisted intervention: MICCAI 2005. Lecture notes in computer science, vol 3749. Springer, Berlin
6. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442
7. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondre K, Stanbridge D, Fried GM (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190(1):107–113
8. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, Kuksenok K, Aragon C, Holst D, Lendvay T (2013) Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. J Surg Res 187(1):65–71
9. Kelly JD, Peterson A, Lendvay TS, Kowalewski TM (2020) The effect of video playback speed on surgeon technical skill perception. In: International proceedings of computer-assisted interventions—IPCAI 2020. Munich, Germany.
10. Huaulme A, Voros S, Riffaud L, Forestier G, Moreau-Gaudry A, Jannin P (2017) Distinguishing surgical behavior by sequential pattern discovery. J Biomed Inform 67:34–41
11. Forestier G, Petitjean F, Senin P, Despinoy F, Huaulme A, Fawaz HI, Weber J, Idoumghar L, Muller PA, Jannin P (2018) Surgical motion analysis using discriminative interpretable patterns. Artif Intell Med 91:3–11
12. Malpani A, Lea C, Chen CCG, Hager GD (2016) System events: readily accessible features for surgical phase detection. Int J Comput Assist Radiol Surg 11(6):1201–1209
13. Lea C, Reiter A, Vidal R, Hager GD (2016) Segmental spatiotemporal cnns for fine-grained action segmentation and classification. arXiv:1602.02995
14. Wang Z, Fey AM (2018) Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. Int J Comput Assist Radiol Surg 13:1959–1970
15. Doughty H, Damen D, Mayol-Cuevas WM (2017) Who's better, who's best: skill determination in video using deep ranking. arXiv:1703.09913
16. Zia A, Zhang C, Xiong X, Jarc A (2017) Temporal clustering of surgical activities in robot-assisted surgery. Int J Comput Assist Radiol Surg 12:1171–1178
17. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
18. Schuster M, Paliwal KP (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(5):2673–2681
19. Kowalewski T, Comstock B, Sweet R, Schaffhausen C, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, McDougall E, Lendvay TS (2015) Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills (BLUS) tasks. J Urol 195(6):1859–1865
20. Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL (1998) Development of a model for training and evaluation of laparoscopic skills. Am J Surg 175:482
21. Fried GM (2008) FLS assessment of competency using simulated laparoscopic tasks. J Gastroenterol Surg 12:210
22. Peters JH, Fried GM, Swanstrom LL, Soper NJ, Silin LF, Schirmer B, Hoffman K (2004) Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery 135:21
23. Seete RM, Beach R, Sainfort F, Gupta P, Reihsen T, Poniatowski LH, McDougall EM (2012) Introduction and validation of the American urological association basic laparoscopic urology surgery skills curriculum. J Endourol 26:190

24. Kowalewski TM, Seet R, Lendvay TS, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, McDougall E (2016) Validation of the AUA BLUS tasks. J Urol 195:998
25. French A, Seidel K, Lendvay TS, Kowalewski TM (2018) Role of contextual information in skill evaluation of minimally invasive surgical training procedures. In: Hamlyn symposium on medical robotics, London, United Kingdom