Deep neural networks for the assessment of surgical skills: A systematic review

Erim Yanik, Xavier Intes, Uwe Kruger, Pingkun Yan, David Diller, Brian Van Voorst, Basiel Makled, Jack Norfleet, and Suvranu De

Abstract
Surgical training in medical school residency programs has followed the apprenticeship model. The learning and assessment process is inherently subjective and time-consuming. Thus, there is a need for objective methods to assess surgical skills. Here, we use the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to systematically survey the literature on the use of Deep Neural Networks for automated and objective surgical skill assessment, with a focus on kinematic data as putative markers of surgical competency. There is considerable recent interest in deep neural networks (DNNs) due to the availability of powerful algorithms, multiple datasets, some of which are publicly available, as well as efficient computational hardware to train and host them. We have reviewed 530 papers, of which we selected 25 for this systematic review. Based on this review, we concluded that DNNs are potent tools for automated, objective surgical skill assessment using both kinematic and video data. The field would benefit from large, publicly available, annotated datasets representing the surgical trainee and expert demographics and multimodal data beyond kinematics and videos.

Keywords
Deep learning, deep neural network, artificial intelligence, convolutional neural network, LSTM, GRU, RNN, surgical skill assessment, laparoscopic surgery, robotic surgery, virtual surgical simulators

One sentence summary
This paper provides a comprehensive review of the current state of the literature on automated surgical skill assessment based on deep neural networks (DNNs).

1. Introduction
Surgical procedures require high dexterity, fine motor control, and hand-eye coordination. Surgical training has followed the Halstedian apprenticeship model of learning with an expert in the operating room (OR).1,2 Traditional surgical assessment methods, such as direct observations by an experienced trainer to assess the trainee’s skills, are generally subjective and use global rating scales (GRS), e.g., the Objective Structure Assessment of Technical Skills (OSATS), to score competency.3 These observational methods allow experienced surgeons to use structured checklists for technical criteria and rate the surgical performance of the trainee under direct observation.

1Department of Mechanical, Aerospace, and Nuclear Engineering, Rensselaer Polytechnic Institute, USA
2Department of Biomedical Engineering, Rensselaer Polytechnic Institute, USA
3Raytheon BBN Technologies, USA
4Simulation and Training Technology Center, Army Research Laboratory, USA

Corresponding author:
Erim Yanik, M.S., Department of Mechanical, Aerospace, and Nuclear Engineering, Rensselaer Polytechnic Institute, 110 8th St, Low Center for Industrial Innovation, Troy, NY 12180, USA.
Email: yanike@rpi.edu
Surgical training programs also utilize other checklist-based approaches, e.g., In-Training Evaluation Reports (ITERs), to evaluate trainees. These approaches suffer from the well-known limitations of being subjective with poor inter-rater reliability, distribution errors, recall bias, and halo effects.\textsuperscript{4–6} Increasingly, simulators are being used to provide an efficient training environment outside the OR and certify surgeons, eliminating risks to human patients. Notably, the Fundamentals of Laparoscopic Surgery (FLS) program, developed by a joint committee of the American Gastrointestinal and Endoscopic Surgeons (SAGES) and the American College of Surgeons (ACS), is a prerequisite for board certification in general and ob/GYN surgery.\textsuperscript{7,8} The Fundamentals of Endoscopic Surgery (FES) utilizes a virtual reality-based simulator and is required for general surgery board certification.\textsuperscript{9} Moreover, a more recent curriculum, Fundamentals of Robotic Surgery (FRS), offers standardization to assess surgeons in robotic surgery.\textsuperscript{10} Simulators such as these use metrics based on time and error to provide performance scores to certify surgeons. The definition of error in these simulators is subjective and may not correlate with operative errors. The evaluation of these errors, e.g., calculating the deviation of a cut-trajectory from a pre-marked circle in the pattern cutting task of the FLS, is also manually intensive and subjective. Also, time to completion is heavily weighted, with no evidence that completing a task faster indicates surgical proficiency.

There is currently an intense focus on developing data-driven techniques to overcome the limitations of surgical skill assessment. Two of the major categories of data that are collected for skills assessment include kinematic and video-based data. Kinematic data are collected by placing sensors at tool tips, hands, or other locations of the trainee’s body to collect motion-related data.\textsuperscript{11–15} Data for features such as location, velocity, and rotation of a surgical tool can be collected directly using these sensors. This type of data collection is invasive due to the sensor’s presence interfering with the trainee’s motion. The kinematic data can also be collected virtually using software such as EndoVis for virtual training applications, which does not suffer from the limitations of physical sensors.\textsuperscript{16} Unlike kinematic data, video-based data are gathered non-invasively and are easier to work with. However, video-based data require extensive post-processing to extract meaningful information. Crowd-sourcing has been proposed as a method to leverage the crowd’s wisdom and reduce the need for time-intensive work by medical professionals for scoring surgical videos.\textsuperscript{17}

Automated data processing is a prerequisite to automating skill assessment. Earlier works derive statistical representations from the collected data to represent surgical performance.\textsuperscript{18–20} Anh et al.\textsuperscript{21} used the mean, min-max-standard deviation (SD) of each sequence together with more complex representations, e.g., the average absolute difference, the average root sum of squared level, the average root square difference, and binned distribution to represent the kinematic tool motion data collected for elementary surgical tasks performed using surgical robots.\textsuperscript{12} With the statistical features of choice, they differentiated experts from novices with accuracies between 87.25% and 95.42%. Moreover, Kim et al.\textsuperscript{22} employed crowd-sourced tool tip velocities (TTV) and optical flow fields (OFF) to represent tool motion during Capsulorhexis and differentiated the experts from novices with accuracies of 84.8% and 63.4% for TTV and OFF, respectively.

Although statistical representations prove helpful in specific scenarios, manual extraction of the features needed for domain knowledge and time-effectiveness is a critical limitation. Machine learning (ML) techniques, including hidden Markov models (HMM), support vector machines (SVM), and bag of words (BoW), are data-driven models extensively used for surgical skill assessment.\textsuperscript{24–27} In a hybrid study, Fard et al.\textsuperscript{28} extracted statistical features, including duration, path length, and smoothness, from the kinematic tool motion data provided by the DaVinci surgical system (dVSS). The features were selected to represent the surgeon’s dexterity based on tool motion. Logistic regression (LR) and support vector machine (SVM) were applied for binary classification of eight surgeons performing suturing, and a global rating score was utilized to annotate the surgeon’s performance. As a result, a maximum of 85.7% overall accuracy was reported using LR with leave-one-super-trial-out (LOSO) cross-validation (CV). In another study conducted by Zia et al.,\textsuperscript{29} various ML techniques were compared, including HMM, BoW, Motion Textures, and Discrete Cosine Transform (DCT), which performs well with repetitive tasks. They collected video-based data on suturing and knot-tying and used OSATS to establish ground truth for the performance. As a result, they showed the benefit of using frequency domain-based techniques, i.e., DCT, over repetitive tasks mentioned above.

While ML techniques are promising, they rely on task-specific manual features and require significant effort to define and optimize. Also, neither statistical features nor ML techniques are end-to-end, i.e., they need multiple steps to produce the desired output, e.g., extracting information from the collected data, computing features, and then using the ML model or an equation to assess performance. Deep learning (DL) is rapidly emerging as a method of choice to overcome these limitations of ML approaches.\textsuperscript{30} DL is a powerful supervised learning technique that consists of hidden layers with neurons within. The neurons extract and represent complex features, and the complexity can be enhanced by adding more layers and neurons. The utility of DL is used to identify surgical
tools in every frame of surgical videos, automatically, based on the extracted features that are otherwise non-detectible to the human eye.

The availability of powerful algorithms, large labeled datasets, and inexpensive high-power computing is fueling the growth in DL-based approaches in surgical assessment.\textsuperscript{31,32} In recent years complex recurrent neural networks (RNNs) such as Long Short Term Memory (LSTM), Gated Recurring Unit (GRU), and their combinations with convolutional neural networks (CNNs) have been utilized for surgical skill assessment.\textsuperscript{33,34} CNNs were also used to assess skills and generate meaningful feedback for the trainee.\textsuperscript{35–38} Moreover, CNNs such as Hourglass Network, Faster Region-Based CNN (Faster RCNN), Temporal Segment Networks (TSN), and others were used to extract tool locations from surgical videos.\textsuperscript{39–43}

This paper systematically analyzes automated and objective surgical skill assessment techniques based on deep neural networks (DNNs) using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.\textsuperscript{44} While multiple review papers have been recently published in similar areas, they lack technical analysis on DNNs and do not state the benefits of DNNs over traditional ML techniques for surgical skill assessment, such as visual feedback or end-to-end learning.\textsuperscript{31,32}

The rest of the paper is organized as follows. The review criteria are presented in Section 2. Section 3 discusses the significant findings from the review, including the available datasets, deep learning models utilized, prevalent cross-validation methods, and skill assessment techniques. Sections 4 and 5 present discussions and conclusions, respectively.

2. Review criteria

We used Google Scholar as the database and included all the published works through June 2020 in the analysis. We did not have a designated start date, and the earliest work we could find in this field belonged to November 2013. There were 8 different keywords and 14 different keyword combinations used to identify the 527 potential publications. The keywords related to surgical skills, namely “surgical skill assessment,” “surgical skill classification,” and “surgical skill evaluation,” were used together with deep learning terminology, including “deep learning,” “artificial intelligence,” “deep neural networks,” “LSTM,” and “convolutional neural networks.” Each keyword combination was treated individually. Papers cited in the obtained papers were also included in our list. We excluded 364 papers, duplicates, patents, books, theses, and publications in foreign languages. The screening was done by reading the title and the abstract of each paper. The papers excluded were mainly related to different surgical applications or did not utilize DL techniques. We believe that these papers were included in the repository due to the articles they cited. After the screening process, we identified 46 papers for an extensive review. In this step, we read each paper thoroughly. We removed the papers about tool detection and localization, workflow and action recognition, and video segmentation in surgery using DNNs because even though these steps are required for skill assessment, the authors did not utilize their technology to do so. Furthermore, some studies that were eliminated reported results by using ML techniques and claimed that DL was their next study aim. This process finally led to a total of 25 publications fulfilling all criteria (Figure 1). The temporal distribution of the papers is presented in Figure 2.

3. Review results

We observed that the published papers on DNN-based surgical skill assessment could be broadly categorized based on the dataset used — those that utilized the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) and those that used other datasets.\textsuperscript{12} The papers also differed in the data acquisition, DNN models, cross-validation, and skill assessment techniques, as presented below and summarized in Tables 1 and 2.

3.1. Datasets

The availability of relatively large datasets, including publicly available ones, has spurred significant growth of DNL-based techniques. Both kinematic and video-based data were collected and utilized in these datasets. The kinematic data were gathered using sensors, simulators, or surgical robots, whereas the video-based data were obtained using either virtual simulators or cameras. For instance, JIGSAWS is a publicly available dataset that is used in multiple studies.\textsuperscript{12} It is a robotic surgery dataset containing data from three elementary tasks performed using the dVSS, including suturing, needle passing, and knot-tying. Suturing is done by passing a needle through a synthetic tissue. Needle passing, like suturing, is where a needle is passed through metal hoops from one side to the other. Knot-tying is when the subject ties a single loop knot using both ends of the suture. There are 39 suturing, 36 knot-tying, and 28 needle passing trials conducted by eight subjects, including two experts with 100+ hours of experience, four novices with less than 10 hours of experience, and two intermediates between 10 and 100 hours of experience. Here, the hours of expertise are self-proclaimed, and expert
and intermediate subjects are underrepresented compared with novices. In addition, each trial is annotated using a modified version of OSATS. The dataset includes a total of 76 variables representing kinematic data associated with tooltip positions in Cartesian coordinates, tooltip rotation matrix, tip linear and rotational velocities, and gripper angle velocity for four robotic arms including two master and two patient side manipulators. The kinematic data have been collected by dVSS and video-based data recorded by endoscopic cameras at 30 FPS at a resolution.
of 640 x 480. The dataset has gesture-based verbal annotations yet does not possess spatial annotation for the tools used during the tasks.

Another publicly available dataset is Cholec80, which consists of 80 videos of laparoscopic cholecystectomy performed by 13 experts and captured at 25 FPS. \(^{42}\) The binary tool presence annotation was provided at a rate of 1 FPS, offering a suitable environment to study DNN for tool detection. \(^{60}\) In addition, mc2cai16-tool and mc2ai16-workflow are other datasets constructed based on Cholec80 for the 2016 M2CAI Challenge. \(^{32,61}\) Mc2cai16-tool was later expanded by Jin et al. \(^{40}\) who introduced spatial annotation

### Table 1. Summary of objective and automated surgical skill assessment on JIGSAWS.

| Author [year] | Data acquisition | Data processing | cv | Form | Visual-Feedback |
|---------------|------------------|-----------------|----|------|-----------------|
| Benmansour et al. \(^{45}\) [2018] | Kinematic | DNN | N/A | C | N/A |
| Dougherty et al. \(^{46}\) [2018] | Video-Based | Sesame Model + Two-stream TSN | 4-fold | C | N/A |
| Fawaz et al. \(^{38}\) [2018] | Kinematic | CNN with GAP | LOSO* | C | CAM |
| Wang et al. \(^{35}\) [2018] | Kinematic | CNN | LOSO | C | N/A |
| Wang et al. \(^{47}\) [2018] | Kinematic | CNN-GRU, CNN-LSTM, LSTM, PCA, Autoencoder, Statistical, DFT, DCT | LOSO | C | N/A |
| Anh et al. \(^{31}\) [2019] | Kinematic | CNN-LSTM, LST,PCA, Autoencoder, Statistical, DFT, DCT | LOSO | C | N/A |
| Benmansour et al. \(^{48}\) [2019] | Kinematic | 3 RNNs for KT, NP, and ST | LOSO* | C | N/A |
| Castro et al. \(^{37}\) [2019] | Kinematic | CNN with SELU + GMP + Quaternion Convolutions | LOSO* | R and C | CAM |
| Fawaz et al. \(^{36}\) [2019] | Kinematic | CNN with GAP | LOSO* | C | N/A |
| Funke et al. \(^{41}\) [2019] | Video-Based | 3DCNN + TSN (First end-to-end) and LOUO | LOSO | C | AM |
| Li et al. \(^{49}\) [2019] | Video-Based | RNN-based attention network | 4-fold | C | AM |
| Nguyen et al. \(^{50}\) [2019] | Kinematic | CNN-LSTM | LOSO | C | N/A |
| Zhang et al. \(^{51}\) [2020] | Kinematic | CNN | LOSO | C | CAM |

GAP: Global Average Pooling, GMP: Global Mixed Pooling, CAM: Class Activation Function, AM: Attention Maps, KT: Knot Tying, NP: Needle Passing, ST: Suturing.

**Note:** Bold-written models under Data Processing achieve the highest score.

*The authors applied an additional split on the training set and trained their models respectively. However, the details for the split, i.e., ratio, were not disclosed.

### Table 2(a). Summary of objective and automated surgical skill assessment on kinematic data.

| Author [year] | Surgical dataset / procedure | Type of surgery | Data processing | CV | Form. |
|---------------|-----------------------------|-----------------|-----------------|----|-------|
| Ahmad et al. \(^{52}\) [2013] | PT and Grasping Transvaginal assessment diagnosis | Virtual Laparoscopic | Elman DNN LSTM, Simple RNN | N/A | R |
| Cifuentes et al. \(^{55}\) [2019] | RARP | Robotic Laparoscopic | DeepSurv, DNN | 5-fold | C |
| Kowalewski et al. \(^{14}\) [2019] | ST and KT | Robotic Laparoscopic | DNN | 10-fold | R and C |
| Pérez Escamisosa et al. \(^{55}\) [2019] | PC, ST, and KT | Virtual | RBFNet (feed forward DNN) | LOUO, Holdout | C |
| Uemura et al. \(^{56}\) [2019] | ST | Laparoscopic | 9 statistical features + Chaos Neural Network LSTM, CNN, FCN (converted SNN) | N/A | C |
| Getty et al. \(^{57}\) [2020] | P&P, PT, and TOR | Virtual | LSTM, CNN, CNN with final discriminative layer, Multi-channel CNN | N/A | C |
| Vaughan et al. \(^{58}\) [2020] | Epidural Training | Virtual | ResNet, FCN (with GAP), CNN | LOOOCv.5-fold | C |

SNN: Spiking Neural Network, RBFNet: Radial basis function networks. RARP: Robot-assisted radical prostatectomy. PT: Peg Transfer, ST: Suturing, KT: Knot Tying, PC: Pattern Cutting, P&P: Pick and Place, TOR: Thread the Rings R: Regression, C: Classification.

**Note:** Bold-written models under Data Processing achieve the highest score.
for the tools so that the dataset would be suitable for tracking tool motion and the technical assessment of surgical skills. They were also the first ones to use the dataset. Their results matched with OSATS classification for four experts. Besides public datasets, studies reported assessment on private datasets collected using the dVSS, simulators, sensors, and cameras.54–56,58,60

Following our analysis, we found 5 studies out of 25 that acquired data virtually using simulators. In addition, 15 papers, including the JIGSAWS studies, were robotic, and 4 were of laparoscopic surgery. One paper studied Capsulorhexis; hence, this is labeled as general surgery. Based on works besides JIGSAWS, half of the papers utilizing kinematic data acquired them virtually. In contrast, only a quarter of the papers obtained data from virtual simulations for video-based studies. Additionally, our survey showed that only tool motion (23 studies) and hand motion (2 studies) were utilized. We could not find papers employing other methods, including eye motion tracking and muscle contractions used in earlier ML techniques.62

Table 2(b). Summary of objective and automated surgical skill assessment on video-based data.

| Author [year]                      | Surgical dataset / procedure       | Type of surgery | Data acquisition | Data processing     | CV          | Form. |
|------------------------------------|------------------------------------|-----------------|-----------------|---------------------|-------------|-------|
| Law et al.39 [2017]                | Vesicourethral Anastomosis         | Robotic         | Hourglass       | SVM                 | LOOCV       | C     |
| Jin et al.40 [2018]                | m2cai16-tool + m2cai                | Laparoscopic    | Faster RCNN     | SFR                 | Splitting   | C     |
| Mazomenos et al.59 [2018]          | Transesophageal Echocardiography    | Virtual         | AlexNet and VGG | SFR                 | N/A         | R and C|
| Kim et al.22 [2019]                | Capsulorhexis                      | Regular         | TCN             | SFR                 | 5-fold      | C     |

TCN: Temporal Convolutional Neural Network, RCNN: Region-Based Convolutional Neural Network, SFR: Statistical Feature Representation.

Note: Bold-written models under Data Processing achieve the highest score.

Figure 3. DNNs employed for surgical skill assessment. Circles other than Visual Feedback represent the architectures, namely convolutional neural networks (CNNs), recurrent neural networks (RNNs), fully-connected networks (FCNs), and CNN-RNN variates, whereas the circle labeled ‘Visual Feedback’ is for visualization. The boxes illustrate the specific frameworks.
3.2. DNN models and data processing

There are three main DNN types used for surgical skill assessment. These are CNNs, RNNs, and Fully-Connected Neural Networks (FCNs). Figure 3 illustrates the DNN models that have been utilized in the reviewed papers:

In the figure, the DNNs mentioned above and also combinations of CNNs and RNNs are shown. In addition, well-known employed structures such as VGG, ResNet, and LSTM are also mentioned. Class Activation Maps (CAMs) are utilized to provide visual feedback to the trainee by highlighting the parts of the performance that contributed significantly to the outcome. They are mentioned under Visual Feedback in Figure 3. CAMs are enabled through Global Pooling. Thus, they are also placed under Visual Feedback. Besides allowing visualization, Global Pooling also eliminates the duration difference problem that is crucial for time-series data.\(^{36,37}\)

Two main data processing approaches were observed. For the kinematic inputs, the data can be directly fed to the DNN, and the model outputs either the regressed score or the computed class. As for the video-based inputs, data processing is usually a two-step process. A DL model first localizes either the tool or the hand motion, which a second DNN subsequently utilizes to assess skill. Notwithstanding, there is one study where the authors developed a Temporal Segment Network (TSN) as a Siamese architecture.\(^{41,62}\) Multiple CNNs working on different segments of a video were aggregated in the consensus layer for classification. To the best of our knowledge, this is the first and only end-to-end model for video-based surgical skill assessment using DNN.

Additionally, out of the 25 papers, there were 12 CNN, 6 RNN, 3 CNN–RNN combinations, 3 Temporal-CNN, and 5 FCN models. Hence, CNN and CNN-X models are the most popular choice, being used in 62% of the studies. Some papers had more than one network developed for comparison. We also observed that the highest accuracies were obtained using CNN models. The results for JIGSAWS are given in Table 3. In this table, the first two rows are studies utilizing ML models provided for comparison. The results by Zia et al. are the current state-of-the-art (accuracy 99.96%) for both ML and DNN, with the best overall accuracy. When the results are compared exclusively for DNN, the model introduced by Castro et al.\(^{37}\) has the highest accuracy overall (98.7%) and also for Knot Tying (98.9%) whereas, for Suturing Funke et al.\(^{41}\) (100%) and Fawaz et al.\(^{36,38}\) (100%) and Needle Passing Fawaz et al.\(^{36,38}\) report the highest accuracies.

3.3. Cross-validation

In ML and DL, cross-validation (CV) is a well-known technique for training a network using as many observations as possible and simultaneously validating the model performance.\(^{65,66,67}\) The validation set does not contain observations that were used for model training. This is achieved by segmenting the data sets into k-folds and applying a holdout strategy, where one of the segments is removed. The network is trained using the remaining (k–1) folds, and the model performance on the fold left out is then recorded. The CV procedure methodically removes one fold at a time until each fold has served as a validation set. The 25 articles surveyed here apply three distinct CV strategies, which are as follows:

- **Leave-one-super-trial-out** (LOSO): the \(i^{th}\) trial from each participant is removed and used to validate the performance of the DNN model while the remaining data are used for training. The performance of the DNN model is then evaluated after k-folds of super trials have been left out for model validation. \(k = 5\) for JIGSAWS.\(^ {12}\)
- **Leave-one-user-out** (LOUO): removing the trials conducted by a single user to validate the model performance until each user has been left out once.
- **k-fold cross-validation**, the trials are concatenated in a randomized fashion, and the combined dataset is divided into \(k\) folds. A randomized arrangement of trials aims to avoid biasing the model validation to particular users.

LOSO is employed for studies utilizing JIGSAWS and those involving robotic surgery. A total of 9 out of 25 papers have utilized LOSO to cross-validate their models. This is because JIGSAWS lacks sufficient data, and LOSO allows DNN to observe all but one data point for each participant during training. Moreover, LOSO achieved higher accuracy for skewed datasets such as JIGSAWS. LOUO, another CV technique, has only been used to compare it with the other CV schemes. In addition, 7 out of 25 papers used k-folds (\(k = 4, 5, \text{and } 10\)) to cross-validate their models. Here, two out of seven were LOOCV, a special version of k-fold where \(k\) equals the data size. Furthermore, k-fold is the most popular CV method simultaneously amongst non-JIGSAWS and video-based skill assessment studies. Other than those above, four studies utilized Train/Test/Validation Set Splits, and one paper did not provide validation during the training. Also, 6 out of 25
did not report any CV method. The models that were not cross-validated were trained and tested once; hence, there is no guarantee that the fixed set covers all possible variations of movements.

3.4. Skills assessment techniques

We found that the two skills assessment techniques used in this field are classification and regression. Out of the 25 papers analyzed, we found three papers that use regression in addition to classification, and one used regression alone. For instance, in a study based on the JIGSAWS kinematic data, Fawaz et al. 36,38 used a fully connected neural network (FCN) to predict the OSATS score for each trial. They used Spearman’s correlation coefficient to assess the model performance and computed 0.6, 0.57, and 0.65 for Suturing, Needle Passing, and Knot Tying, respectively.

Secondly, Kowalewski et al. 14 also developed an FCN to predict OSATS score for Suturing and Knot Tying and reported an overall mean squared error (MSE) of 3.71 ± 0.64 and R² of 0.03 ± 0.81. They also did binary classification analysis and computed 70% accuracy.

For the classification studies, the papers based on JIGSAWS favored classification into three categories: Novice, Intermediate, and Expert. The Intermediate class suffers the most from misclassification. For example, Funke et al. 41 misclassified all the intermediate surgeons into either expert or novice using optical flow data with 3D-CNN for Knot Tying. Their model was cross-validated by LOUO, and they claimed that the uneven distribution of the dataset leads the DNN to fail to generalize the features. The lack of sufficient expert surgeons might also explain why we did not find any study using LOUO alone. In another study, Wang et al. 35 computed their lowermost Intermediate class score as 0.57 for Knot Tying. Furthermore, the Intermediate class score was consistently lower for the other two JIGSAWS tasks. Anh et al. 21 also had similar performance issues for the Intermediate class, especially for Knot Tying.

In addition to these studies, Getty et al. 57 took another approach and classified four expert surgeons with similar expertise. They achieved 83.4% accuracy using LSTM. Also, three studies trained DNN only with expert data and ranked the non-expert surgeons based on the deviation. 45,48,52 The most common metrics to formulate regression were R² and Root Mean Squared Error (RMSE), whereas, it was accuracy, precision, recall, and F1 score for classification.

4. Discussion

Automation of technical skill assessment promises objective evaluation, reduction of resource requirements, and personalized training with real-time feedback. In this study, we analyzed the current literature on the use of DNN for objective surgical skill assessment. Our first finding was that only tool and hand motion studies were available in the literature that fit our criteria. This may be attributed to the lack of public datasets or resources available for other data.

We also observed that kinematic and video-based data are the only two data types utilized in the studies, with kinematic data being used in two-thirds of the published papers. We believe this is because the sensors can record the kinematic data of the tool and hand motion directly, as opposed to video-based data. Moreover, it can be directly fed to the DNN without any intermediate steps. Authors either annotated a certain portion of the dataset spatially or relied on crowd-sourcing to perform the same task. Amazon Mechanical Turk was the method of crowd-sourcing for all the surveyed studies that utilized crowd-sourcing. Then, DNNs were trained to detect and localize the tools for the

| Author          | Method          | Suturing | Needle Passing | Knot Tying | Overall |
|-----------------|-----------------|----------|----------------|------------|---------|
| Tao et al. 64   | S-HMM           | 97.4     | 96.2           | 94.4       | 96.0    |
| Zia et al. 26   | ApEn            | 100      | 100            | 99.9       | 99.96   |
| Zhang et al. 51 | CNN             | 98.4     | 86.8           | 80.8       | 88.7    |
| Anh et al. 35   | LSTM            | 95.1     | 91.5           | 89.6       | 92.1    |
| Wang et al. 21  | CNN             | 92.3     | 95.4           | 91.3       | 93.1    |
| Anh et al. 21   | CNN-LSTM        | 96.4     | 93.4           | 91.0       | 93.6    |
| Wang et al. 47  | CNN-GRU         | N/A      | N/A            | N/A        | 96.6    |
| Nguyen et al. 50| CNN-LSTM        | 98.4     | 98.4           | 94.8       | 97.2    |
| Fawaz et al. 36 | CNN             | 100      | 100            | 92.1       | 97.4    |
| Funke et al. 41 | 3DCNN + TSN     | 100      | 95.8           | 95.8       | 97.4    |
| Castro et al. 37| CNN             | 98.4     | 98.9           | 98.9       | 98.7    |
rest of the dataset. However, collecting kinematic data is more accessible in robotic surgery or virtual reality-based simulators and not in open or laparoscopic surgery, where tools will have to be instrumented during the operation. In addition, the features extracted from the kinematic data, such as tooltip location or rotation, might not cover the information that the collected data possess; hence, it might not necessarily be optimal for the given surgical task. On the other hand, video data are easy to collect and unobtrusive, and collected during laparoscopic and robotic surgeries. Although video-based data require significant annotation and preprocessing and the localization DNNs require high computational power, they can assess skills solemnly based on performance recordings without any explicit preparation once the network is trained and cross-validated.

Among the different techniques employed, CNNs were superior to the other DNNs such as LSTM and FCN. This might be because RNNS are optimized for data sequences, and the public datasets are not rich with long sequences of surgical data. In addition, most CNNs take statistical features as input, eliminating the need for LSTM since the features of choice represent the temporal information. Among the studies sharing the same dataset, JIGSAWS, we found that a model proposed by Castro et al.\(^{37}\) achieved the highest overall score. They achieved 98.7% accuracy using a CNN with SELU activation, GMP, and quaternion convolutional layers. However, the current state-of-the-art for JIGSAWS is due to Zia et al.\(^{26}\) in which the predictability of the tool motion is encoded using approximate entropy [ApEn].\(^{25}\) They reported 99.96% overall accuracy in differentiating surgeons into categories. Even though there is a 1.26% difference in accuracy between these two models, we believe we should also consider two main advantages that Castro et al.\(^{39}\) have over Zia et al.\(^{26}\) First, the ML model needs to extract information from the kinematics before feeding them into the nearest neighbor classifier. However, in Castro et al.\(^{37}\) the kinematics are fed directly into CNN without an explicit step; therefore, it is an end-to-end model. Second, Castro et al.\(^{37}\) embed the CAM to the DNN based on Fawaz et al.\(^{36,38}\) It fits the time-series data into a segregated color diagram with respect to the contribution of each timestamp. This visual feedback helps the participant on which parts of the performance he/she should improve, a quality missing in Zia et al.\(^{26}\) We believe these advantages are more than enough to favor the work of Castro et al.\(^{37}\) over Zia et al.,\(^{26}\) especially considering the increasing number of studies published over the years, as seen in Figure 3.

Our analysis suggested that LOSO is the most commonly used CV method for surgical skill assessment. Results obtained by LOSO are reported to be higher in comparison to LOUO. We believe this is because of the skewed nature of the datasets. For instance, JIGSAWS has only two experts and two intermediates, whereas there are four novices present. LOSO models trained by seeing data from all the surgeons are therefore exposed to all different styles. Nevertheless, the same is not valid for LOUO because the DNN does not see data from one surgeon, and in a case, the surgeon is an expert or an intermediate; the DNN is left with only one expert or intermediate data for training. This hurts the generalization power of the DNN. In addition, models cross-validated by LOSO have already achieved accuracies of up to 98.7%. Therefore, the field for LOSO CV is almost saturated. We hypothesized that, even though LOUO reports poorer scores, we will see more papers published to improve LOUO to compensate for the gap in the literature. We were surprised to find that k-fold CV is favored by the video-based studies over LOSO and LOUO. We expected the opposite because video-based data are computationally expensive to work with, and the CV techniques require multiple runs of the same network. We assume k-fold CV is utilized to prevent bias in the dataset, hence, overfitting regardless of the expensive computational requirements.

Finally, we noted that the majority of the papers formulate skill assessment as a classification problem. There are only four papers that predict the performance score directly. However, the model developed by Kowalewski et al.\(^{14}\) suggests R\(^2\) of 0.03 with an SD of 0.81; hence, we assume their model has limited or no generalization power. This leads us to believe that there is still room for improvement in regression accuracy. A reason for limited regression studies is that, unlike classification, regression requires a ground-truth scoring system. However, such scoring systems are limited and not necessarily publicly available, e.g., the FLS scores used for board certification are IP protected. In contrast, classification can be done based on their operative expertise.

We observed that intermediate surgeons are the most difficult to classify, which may have multiple reasons. First, the chosen categorization method may not reflect actual surgical experience. For JIGSAWS, we hypothesize that there is a substantial performance improvement between 10 and 100 hours of expertise, making classification harder. Secondly, the DNN may be biased toward novice data. This is because expert behavior may not adequately be represented in unbalanced datasets containing a significantly larger amount of novice data. This is particularly prevalent in the JIGSAWS studies. Lastly, a surgeon may be an expert or a novice in reality. Hence, any DNN classification for such cases would be considered correct by default.

Additionally, we observed that the intermediate surgeons were classified poorly, primarily for the Knot Tying task of JIGSAWS. This might be because the ground-truth performance of the experts and intermediates was reported to be similar for the Knot Tying task.
5. Concluding remarks

In this paper, we examined papers related to the automation of surgical skill assessment using DNN. We used PRISMA guidelines to process papers and ended up analyzing 25 papers. As a result of the systematic review, we observed that DNN is useful to assess the surgical skills and has already reported higher accuracies comparing top ML models, with the work of Zia et al. being the only exception as tabulated in Table 3. Tool and hand motion data obtained from sensors and videos have been used for DNNs, though other sensor-based metrics are possible. In recent work, we have shown that non-invasive brain imaging can predict FLS scores in pattern cutting using a CNN framework named Brain-NET. In our analysis CNN outperformed various machine learning techniques, namely, Kernel-Partial Least Squares, Support Vector Regression, and Random Forests. This survey yielded that the best-performing models rely on convolutional neural networks, including the current state-of-the-art model in DL, achieving overall 98.7% accuracy on JIGSAWS. Finally, we found that the most preferred method to report technical skills is to classify surgeons into distinct categories, namely Expert, Intermediate, and Novice.

Several limitations of the current studies are evident from the analysis, which points to future research directions. The most critical limitation is, of course, the lack of publicly available datasets of significant size. Surgical data are expensive to collect, and participation is voluntary. Having sufficient numbers of expert surgeons dedicate time to generate datasets is challenging due to their clinical obligations. Moreover, there is no consensus on the datasets regarding the effects of gender and handedness on the outcome. Thus, we expect to see future datasets focusing on diversity as much as the quantity.

Moreover, we noticed a lack of rigor in establishing ground-truth performance scores. Most of the studies suggest they had an expert on the team who did the annotation, which might adversely affect inter-rater reliability. Moreover, some studies did not even cite the source of their supervision, meaning it is unclear whether they consulted an expert or not to annotate the ground-truth performance. Additionally, most papers utilized kinematic data over video-based data, possibly due to limited video information in existing surgical datasets. Moreover, there is only one end-to-end DNN model based on video analysis. With the advances in data augmentation to dilute the effect of sample size, we believe there will be more studies focusing on end-to-end frameworks via video sequences.

We were only able to find four regression studies in the literature. Regression differs from classification as the aim is to predict scores for each performance based on GRS. This makes regression advantageous as it is continuous, and the progress can be tracked by the trainee even when the assigned class for the performances does not change. Unlike regression, classification puts surgeons into discrete categories that leave gray areas. For example, for JIGSAWS, the self-proclaimed range of expertise was extensive for intermediates (10–100 hours), with novices up to 10 hours. Experts were 100+ hours, but we do not know the actual hours required to classify surgical skills. However, regression requires more data to pinpoint the performance. Also, there is no agreement in the literature on how the different regression models should be evaluated. We see this as a limitation for regressive models and expect more studies to be published in this field.

Finally, we found only one study based in the OR setting, while the rest of the studies used either physical or virtual simulators for skill assessment. While the training and certification of surgeons can be done on simulators, it is critical to quantify their skills in clinical situations objectively. There are multiple challenges to doing this due to the inherent messiness of the operative environment where more than one surgeon may be involved in performing a procedure, annotation of the operative videos to provide high-quality training datasets, and privacy issues related to the acquisition and use of the datasets.

In conclusion, we believe that DNNs possess strong potential for assessing surgical skills rapidly and objectively, providing real-time feedback, and certifying and credentialing surgeons in the near future. The availability of large, high-quality, publicly available annotated datasets will accelerate the achievement of that goal.

ORCID iD

Erim Yanik https://orcid.org/0000-0001-8290-2144

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Nataraja RM, Webb N and Lopez PJ. Simulation in paediatric urology and surgery, part 2: An overview of simulation modalities and their applications. J Pediatr Urol 2018; 14: 125–131.
2. Dutta S and Krummel TM. Simulation: a new frontier in surgical education. Adv Surg 2006; 40: 249–263.
3. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Brit J Surg 1997; 84: 273–278.
4. Reiley CE, Lin HC, Yuh DD, et al. Review of methods for objective surgical skill evaluation. Surg Endosc 2011; 25: 356–366.
5. DaRosa DA, Zwischenberger JB, Meyerson SL, et al. A theory-based model for teaching and assessing residents in the operating room. J Surg Educ 2013; 70: 24–30.
6. Darzi A, Smith S and Taffinder N. Assessing operative skill: needs to become more objective. BMJ.
7. Fried GM. FLS assessment of competency using simulated laparoscopic tasks. J Gastrointest Surg 2008; 12: 210–212.
8. Scott DJ, Ritter EM, Tesfay ST, et al. Certification pass rate of 100% for fundamentals of laparoscopic surgery skills after proficiency-based training. Surg Endos 2008; 22: 1887–1893.
9. Poulose BK, Vassiliou MC, Dunkin BJ, et al. Fundamentals of endoscopic surgery cognitive examination: development and validity evidence. Surg Endos 2014; 28: 631–638.
10. Smith R, Patel V and Satava R. Fundamentals of robotic surgery: a course of basic robotic surgery skills based upon a 14-society consensus template of outcomes measures and curriculum development. Int J Med Robot Comp 2014; 10: 379–384.
11. Rupprecht C, Lea C, Tombari F, et al. Sensor substitution for video-based action recognition. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9 October 2016, pp.5230–5237.
12. Gao Y, Vedula SS, Reiley CE, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: Muccai workshop: M2cang, Boston, MA, September 2014, vol. 3, p.3.
13. Wanzel KR, Hamstra SJ, Caminiti MF, et al. Visual-spatial ability correlates with efficiency of hand motion and successful surgical performance. Surgery 2003; 134: 750–757.
14. Kowalewski KF, Garrow CR, Schmidt MW, et al. Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying. Surg Endosc 2019; 33: 3732–3740.
15. Bouget D, Allan M, Stoyanov D, et al. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Med Image Anal 2017; 35: 633–654.
16. Pérez-Escamisosa F, Alarcón-Paredes A, Alonso-Silverio GA, et al. Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches. Int J Comput Ass Rad 2020; 15: 27–40.
17. Kirsch S, Comstock B, Warren J, et al. Crowd sourced assessment of technical skills (CSATS): A scalable assessment tool for the nursing workforce. J Invest Med 2015; 63: 92.
18. Hofstad EF, Välenstad C, Chmarra MK, et al. A study of psychomotor skills in minimally invasive surgery: what differentiates expert and non-expert performance. Surg Endos 2013; 27: 854–863.
19. Bajka M, Tuchschmid S, Fink D, et al. Establishing construct validity of a virtual-reality training simulator for hysteroscopy via a multimetric scoring system. Surg Endos 2010; 24: 79.
20. Pellen M, Horgan L, Barton JR, et al. Laparoscopic surgical skills assessment: can simulators replace experts? World J Surg 2009; 33: 440–447.
21. Anh NX, Nataraja RM and Chauhan S. Towards near real-time assessment of surgical skills: A comparison of feature extraction techniques. Comput Meth Prog Bio 2020; 187: 105234.
22. Kim TS, O’Brien M, Zafar S, et al. Objective assessment of intraoperative technical skill in Capsulorhexis using videos of cataract surgery. Int J Comput Ass Rad 2019; 14: 1097–1105.
23. Zia A, Sharma Y, Bettadapura V, et al. Video and accelerometer-based motion analysis for automated surgical skills assessment. Int J Comput Ass Rad 2018; 13: 443–455.
24. Winkler-Schwartz A, Bissonnette V, Mirechi N, et al. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. J Surg Educ 2019; 76: 1681–1690.
25. Winkler-Schwartz A, Yilmaz R, Mirchi N, et al. Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. JAMA 2019; 2; e198363.
26. Zia A and Essa I. Automated surgical skill assessment in RMIS training. Int J Comput Ass Rad 2018; 13: 731–739.
27. Ahmidi N, Tao L, Sefati S, et al. A dataset and benchmarks for human motion modeling. In: Proceedings of the 2016 International Conference on Robotics and Automation (ICRA), Daejeon, Korea, 9 October 2016, pp.5230–5237.
28. Fard MJ, Ameri S, Chinnam RB, et al. Machine learning approach for skill evaluation in robotic-assisted surgery. arXiv preprint. November 2016. ArXiv:1611.05136.
29. Zia A, Sharma Y, Bettadapura V, et al. Automated video-based assessment of surgical skills for training and evaluation in medical schools. Int J Comput Ass Rad 2016; 11: 1623–1636.
44. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting statement.

43. Sarikaya D, Corso JJ and Guru KA. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. *IEEE T Med Imaging* 2017; 36: 1542–1549.

42. Twinanda AP, Shehata S, Mutter D, et al. Endonet: a deep neural network architecture for recognition tasks on laparoscopic videos. *IEEE T Med Imaging* 2016; 36: 86–97.

41. Funke I, Mees ST, Weitz J, et al. Video-based surgical skill assessment using 3D convolutional neural networks. *Int J Comput Ass Rad* 2019; 14: 1217–1225.

40. Jin A, Yeung S, Jopling J, et al. Tool detection and operative assessment using computer vision based analysis. In: *Machine learning for healthcare conference*, Boston, MA, November 2017, pp.88–99.

39. Law H, Ghani K and Deng J. Surgeon technical skill assessment using computer vision based analysis. In: *Machine learning for healthcare conference*, Boston, MA, November 2017, pp.88–99.

38. Vaughan N and Gabrys B. Scoring and assessment in medical VR training simulators with dynamic time series classification. *Eng Appl Artif Intel* 2020; 94: 103760.

37. Gottry N, Zhou Z, Grussner S, et al. A deep-learning model using automated performance metrics and clinical features to predict urinary continence recovery after robot-assisted radical prostatectomy. *BJU Int* 2019; 124: 487–495.

36. Perez-Escamirosa F, Alarcon-Paredes A, Alonso-Silvera GA, et al. Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches. *Int J Comput Ass Rad* 2020; 15: 27–40.

35. Uemura M, Tomikawa M, Miao T, et al. Feasibility of an AI-based measure of the hand motions of expert and novice surgeons. *Comput Math Method M* 2018: 6.

34. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Plos Med* 2009; 6: e1000097.

33. Benmansour M, Handouzi W and Malti A. A neural network architecture for automatic and objective surgical skill assessment. In: *2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)*, Algiers, Algeria, 28 October 2018, pp.1–5.

32. Doughty H, Damen D and Mayol-Cuevas W. Who’s better, who’s best: Skill determination in video using deep ranking. *arXiv preprint*. May 2020. arXiv:2005.05868.

31. Uemura M, Tomikawa M, Miao T, et al. Feasibility of an AI-based measure of the hand motions of expert and novice surgeons. *Comput Math Method M* 2018: 6.

30. Mohammed E, Bansal K, Martin B, et al. Automated performance assessment in transossephageal echocardiography with convolutional neural networks. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Granada, Spain, September 2018, pp.256–264. Cham: Springer.

29. Alshirbaji TA, Jalal NA and Möller K. Surgical tool classification in laparoscopic videos using convolutional neural network. *Curr Direct Biom Eng* 2019; 4: 407–410.

28. Stauder R, Ostler D, Kranzfelder M, et al. The TUM LapChole dataset for the M2CAI 2016 Workflow challenge. *arXiv preprint*. October 2016. arXiv:1610.09278.

27. Ahmadi N, Ishii M, Fichtinger G, et al. An objective and efficient recurrent neural network to evaluate classified surgery tasks. In: *5th International Conference on Automation, Control Engineering and Computer Science-ACECS*, 2018.

26. Ahmadi N, Ishii M, Fichtinger G, et al. An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data. In: *International forum of allergy & rhinology*, November 2012, vol. 2(6), pp.507–515. Hoboken: Wiley Subscription Services, Inc., A Wiley Company.

25. Bromley J, Guyon I, LeCun Y, et al. Signature verification using a “siamese” time delay neural network. *Adv Neu In 1994; 6: 737–744.

24. Tao L, Elhamifar E, Khudanpur S, et al. Sparse hidden Markov models for surgical gesture classification and skill evaluation. In: *International conference on information processing in computer-assisted interventions*, Pisa, Italy, 27 June 2012, pp.167–177. Berlin, Heidelberg: Springer.

23. Schaffer C. Selecting a classification method by cross-validation. *Mac Learn* 1993; 13(1): 135–143.

22. Arlot S and Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv* 2010; 4: 40–79.
67. Rodriguez JD, Perez A and Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE T Pattern Anal* 2009; 32: 569–575.

68. Gao Y, Yan P, Kruger U, et al. Functional brain imaging reliably predicts bimanual motor skill performance in a standardized surgical task. *IEEE Trans Biomed Eng* 2020; 7: 2058–2066.

Author biographies

**Erim Yanik**, is a Ph.D. student in the Department of Mechanical, Aerospace and Nuclear Engineering. He received his B.S. in Mechanical Engineering from Istanbul Technical University and obtained his M.S. from Southern Illinois University Edwardsville. His research interests are Computer Vision, Data Analysis, and Deep Neural Networks for surgical training and skill assessment.

**Xavier Intes**, is a professor in the Department of Biomedical Engineering, codirector of the Center for Modeling, Simulation, & Imaging in Medicine, and an AIMBE/SPIE/OSA fellow. He acted as the chief scientist of Advanced Research Technologies Inc., Montreal, Canada. His research interests are on the application of diffuse functional and molecular optical techniques for biomedical imaging in preclinical and clinical settings.

**Uwe Kruger**, is a Professor of Practice in Biomedical Engineering at RPI and engaged in research related to biomedical data science. Dr. Kruger is also a member of the Center for Modeling, Simulation and Imaging in Medicine, where he provides oversight to projects involving statistical and machine learning problems. His research interests focus on developing and applying innovative solution methods to solve nonlinear regression and classification problems.

**Pingkun Yan**, is an Assistant Professor at the Department of Biomedical Engineering at Rensselaer Polytechnic Institute. He obtained his Ph.D. from National University of Singapore. His research interests include medical imaging informatics and image-guided intervention using machine learning and computer vision techniques.

**David Diller**, is a Senior Scientist at Raytheon BBN Technologies. He received his Ph.D. from Indiana University Bloomington. His research interests are AI, computer vision, and human performance modeling.

**Brian Van Voorst**, is a Lead Scientist at Raytheon BBN Technologies. He obtained his M.S. from Michigan Technological University. His research interests are AI and computer vision.

**Basiel Makled**, is a Science and Technology Manager and Engineer at the United States Army Combat Capabilities Development Command. His work is focused on Medical Simulation and Human Performance Training out of Orlando, FL. Basiel completed his Master of Science in Biomedical Engineering at the University of Central Florida and completed his undergraduate schooling at Florida State University, where he received a Bachelor of Science in Engineering with Honors Distinction. Mr. Makled’s primary research interests lie at the cross-roads of Medical Simulation, Training, Human Performance, Artificial Intelligence, and Virtual Reality.

**Jack Norfleet**, is a Chief Engineer at the United States Army Combat Capabilities Development Command. He received his degree from the University of Central Florida. His interests are medical simulations and human tissue properties.

**Suvranu De**, currently serves as the J. Erik Jonsson ’22 Distinguished Professor of Engineering, Director of the Center for Modeling, Simulation and Imaging in Medicine, and Head of the Department of Mechanical, Aerospace and Nuclear Engineering. His research interests are medical simulations, virtual reality, surgical training, machine learning.