Generalization of Deep Learning Gesture Classification in Robotic-Assisted Surgical Data: From Dry Lab to Clinical-Like Data

Danit Itzkovich, Yarden Sharon, Anthony Jarc, Yael Refaely, and Ilana Nisky, *Senior Member, IEEE*

Abstract*—*Objective: **Robotic-assisted minimally invasive surgery (RAMIS) became a common practice in modern medicine and is widely studied. Surgical procedures require prolonged and complex movements; therefore, classifying surgical gestures could be helpful to characterize surgeon performance. The public release of the JIGSAWS dataset facilitates the development of classification algorithms; however, it is not known how algorithms trained on dry-lab data generalize to real surgical situations. *Methods:* We trained a Long Short-Term Memory (LSTM) network for the classification of dry lab and clinical-like data into gestures. *Results:* We show that a network that was trained on the JIGSAWS data does not generalize well to other dry-lab data and to clinical-like data. Using rotation augmentation improves performance on dry-lab tasks, but fails to improve the performance on clinical-like data. However, using the same network architecture, adding the six joint angles of the patient-side manipulators (PSMs) features, and training the network on the clinical-like data together lead to notable improvement in the classification of the clinical-like data. *Discussion:* Using the JIGSAWS dataset alone is insufficient for training a gesture classification network for clinical data. However, it can be very informative for determining the architecture of the network, and with training on a small sample of clinical data, can lead to acceptable classification performance. *Significance:* Developing efficient algorithms for gesture classification in clinical surgical data is expected to advance understanding of surgeon sensorimotor control in RAMIS, the automation of surgical skill evaluation, and the automation of surgery.**

***Index Terms*—Augmentation, clinical data, machine learning, medical robotics, supervised learning, surgical robotics.**

Manuscript received January 20, 2021; revised May 19, 2021 and July 6, 2021; accepted September 26, 2021. Date of publication October 6, 2021; date of current version March 7, 2022. This work was supported in part by the Israeli Science Foundation under Grant 327/20, in part by the Helmsley Charitable Trust through the Agricultural, Biological, and Cognitive Robotics Initiative, in part by the Marcus Endowment Fund (both at Ben-Gurion University of the Negev), and in part by the Besor Fellowship to YS, and the multidisciplinary fellowship to DI. *(Corresponding author: Ilana Nisky.)*

Danit Itzkovich, Yarden Sharon, and Ilana Nisky are with the Biomedical Engineering Deaprtment, Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (e-mail: danititz@post.bgu.ac.il; shayar@post.bgu.ac.il; nisky@bgu.ac.il).

Anthony Jarc is with the Medical Research, Intuitive Surgical Inc., Norcross GA 30092 USA (e-mail: anthony.jarc@intusurg.com).

Yael Refaely is with the Thoracic Surgery Unit, Soroka Medical Center, Beer-Sheva 84101, Israel (e-mail: yaelrefaely@clalit.org.il).

Digital Object Identifier 10.1109/JBHI.2021.3117784

I. Introduction

ROBOTIC-ASSISTED minimally invasive surgery (RAMIS) is a rapidly developing field, and each year the number of RAMIS procedures increases. Compared to open surgery, RAMIS offers the patient many advantages, such as reduced hospital stay and postoperative complications [1]. Compared to standard minimally-invasive surgery, the surgeon enjoys enhanced dexterity and improved vision [2], [3]. An important advantage of RAMIS is that it enables capturing kinematic and video data in real-time [4], opening opportunities for surgical data science research [5]. For example, developing computational models of surgical work-flow and movements [6]–[10], and automated skill assessment [11]–[14]. However, surgical procedures are long, and therefore, require segmentation as part of the data analysis to focus on the relevant surgical context.

Surgical tasks are composed of gestures – short meaningful actions, such as "reaching for needle" and "pulling a suture" [15]. Decomposing tasks into gestures has been proven to be an efficacious preliminary step in the analysis of surgical data, e.g. for skill assessment [15]. Moreover, segmentation allows for optimizing each segment separately in autonomous systems [16]–[19]. The terms gestures *segmentation* [20], [21] *classification* [22], [23], and *recognition* [24]–[26] are often used interchangeably in surgical data science. We will adopt the term *classification* to refer to the assignment of a gesture class to each sampled data point. Manual classification demands watching the procedure, identifying the segments, and deciding which gesture to assign to each segment. Automated classification can be more efficient than manual, less prone to human error, and open the way to online classification for applications such as automated training, detection of complications, or shared control. Therefore, automated gesture classification is widely studied [24], [27]–[29].

To study gesture classification and skill assessment, a surgical dataset, the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [4] was published. It is a dataset of kinematics and video that were recorded during the performance of surgical tasks, and its publication enabled many researchers to develop and test automated classification algorithms. Supervised methods have been used in many of them, such as Hidden Markov Models (HMM) [30], Support Vector Machines (SVM) [31], and logistic regression [31]. These methods rely on having a labeled

dataset. Unsupervised methods were also used [32]–[34], but they tend to be less effective in the case of data as varied as surgical data.

Deep learning from video [20], [35] or kinematic data [25], [36] is also gaining increasing attention in surgical data science, Convolution (CNNs) [20], [35], and recurrent neural networks (RNNs) [25], [36] were employed to classify surgical tasks into gestures. Long-term Recurrent Convolutional Networks (LRCNs) [37], were used on video data [38], [39], and a Time Delay Neural Network (TDNN) on kinematic data [27]. Recently, methods using CNN, and LSTM combined with Temporal Convolutional Networks (TCN) on both video and kinematic data have been presented [40]–[42]. Those methods provide information from both temporal and spatial domains, but their use of video data may be too slow for online gesture classification. Therefore, in this study, we chose to focus on a simple architecture from our previous work [21] – a 2 layer LSTM network on kinematic data.

In contrast to other fields of data science, there are no large public robotic surgery datasets available. The release of surgical data is impeded by factors such as the privacy of the patients and the surgeons, and the need for labor-intensive data annotation [5]. As a result, the only publicly available dataset (JIGSAWS) is a small dataset from performance of a structured dry-lab task, and it is not known how to generalize algorithms that were trained on this data to real surgical data.

Models that are trained with small datasets tend to suffer from overfitting the training data. A number of solutions to overfitting have been proposed, including batch normalization of the layer output activation [43], and dropout of chosen neurons from the network during training [44]. One more technique is transfer learning – the knowledge that was accrued while training with a similar dataset is stored in a pre-trained network, and used when training later with the desired dataset [45]. In [28], the authors pre-trained a network on the suturing task and used it when training on the knot-tying and needle-passing tasks.

Another solution is data augmentation – creating new data from transformations of the existing dataset while preserving the original labels. For images, scaling, cropping, and rotating can be effective [46]. Data augmentation is common in medical applications such as mammogram [47], and liver lesion classification [48]. For time series data, such as the kinematic RAMIS data of this paper, data augmentation by rotation, window slicing, and time warping can be used [21], [49]–[51].

In this paper, we examine the classification of suturing and knot-tying into gestures in a clinical-like surgical dataset using motion kinematics. The contribution of our work is a guideline of how to generalize algorithms that were developed, trained, and tested on structured dry-lab datasets to real surgical data. We start with a simple LSTM architecture from our previous study [21] that is designed, trained, and tested with the JIGSAWS dataset, and demonstrate its poor generalization when attempting to classify the gestures of clinical-like surgical training tasks on porcine models. We unsuccessfully attempted to improve performance with several common techniques such as data augmentation and transfer learning. In contrast, we demonstrate that using the same network architecture together with additional features and training on even a small clinical-like dataset leads to promising classification performance.

II. Methods

A. Data

In this paper, we worked with three different datasets. The first was the JIGSAWS dataset [4]. It contains the data of eight surgeons performing three different tasks five times using the da Vinci Surgical System (Intuitive Surgical Inc., Sunnyvale, California). In this work, we used only two tasks: *suturing* (39 trials), and *knot-tying* (36 trials). The dataset consists of kinematic data sampled at $30Hz$ and manual annotations of gesture labels. Fig. 1(a)-(c) show examples of the endpoint position of the left and right patient side manipulator (PSMs) and a snapshot from the gesture *pushing the needle through tissue* from the dataset.

The second dataset was our internal BGU Biomedical Robotics (BBR) dataset. We reproduced the suturing and knot-tying tasks of JIGSAWS using our dVRK [52] (Fig. 2). The protocol was approved by the Human Subjects Research Committee of Ben Gurion University of the Negev, Be'er-Sheva, Israel, approval number 1283-1, dated July 6, 2015. Two participants (an expert surgeon and an engineering graduate student with many hours of experience working with dVRK) performed the task 3 times, and additional 6 times in which we rotated the task board by 30 and 60 degrees. The surgeon did not perform the task in the 60 ° direction due to technical issues. The dataset consists of kinematic data sampled at $50Hz$ and manual annotations of gesture labels. Fig. 1(d)-(f) show examples of the endpoint position of the left and right PSMs and a snapshot from this dataset.

The third dataset was the clinical-like dataset. Data from surgical training on a porcine model was provided to us under a collaboration agreement with Intuitive Surgical Inc. The protocol, titled "Computer Enhanced Minimally Invasive Surgery-Surgeon and Staff Training" was approved on July 31, 2019. It contains the performance of interrupted suturing by six surgeons. The dataset consists of the endpoint as well as joints' angles kinematics (the latter was not available in the other datasets), sampled at 50 Hz, and video data. We used the video to manually annotate each datapoint into gestures, but we only used kinematic data in the network training and testing. Fig. 1(g)-(i) show examples of the endpoint position of the left and right PSMs and a snapshot from this dataset.

We designed several networks for gesture classification, and tested them using these datasets. First, we trained a network from our previous study [21] with the JIGSAWS data, and tested it with: (1) a surgeon from the JIGSAWS dataset, (2) the BBR dataset to test the generalization to a new but similar dataset, and (3) the clinical-like dataset to test generalization to a more realistic dataset. Second, we designed a new network that was trained and tested on the clinical-like data. Importantly, unlike the networks in prior studies [24], [25], [36] (including our own [21]), we trained a single network on both the suturing and the knot tying tasks together rather than training a separate network for each. This is because in clinical data these tasks are

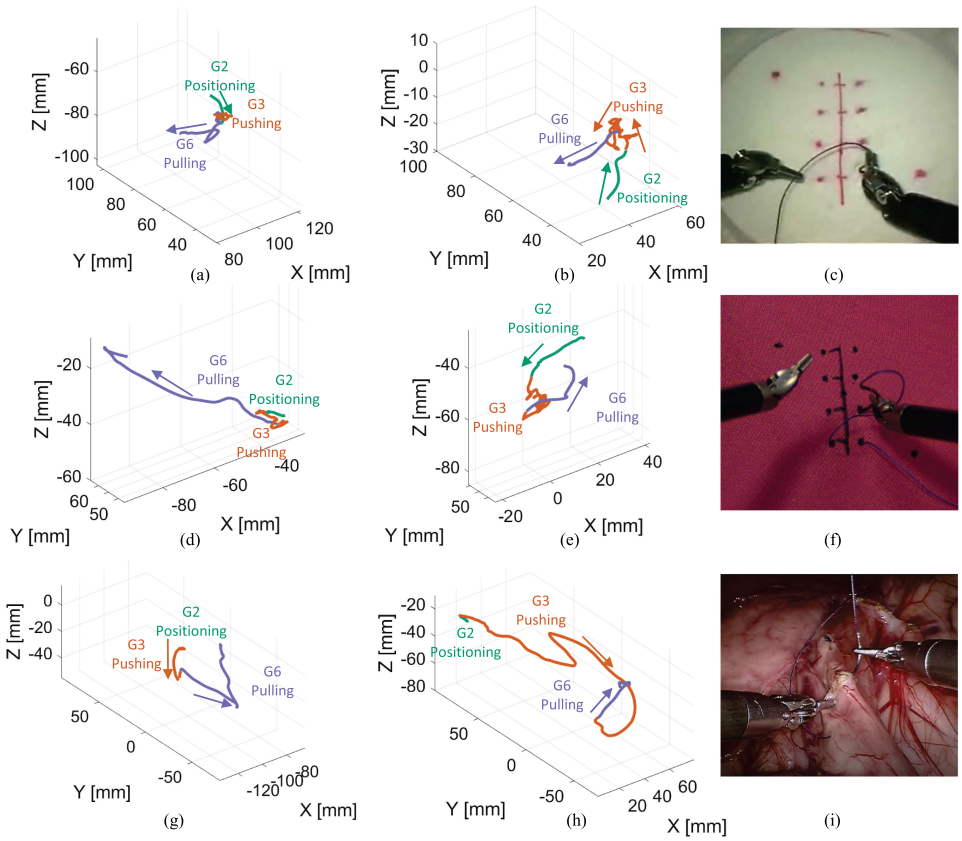

Fig. 1. Three datasets of suturing tasks: JIGSWAS dataset - endpoint position of the left (a) and right (b) PSM and a snapshot of a gesture (c). Similarly, (d-f) for the BBR dataset and (g-i) for the clinical-like dataset.

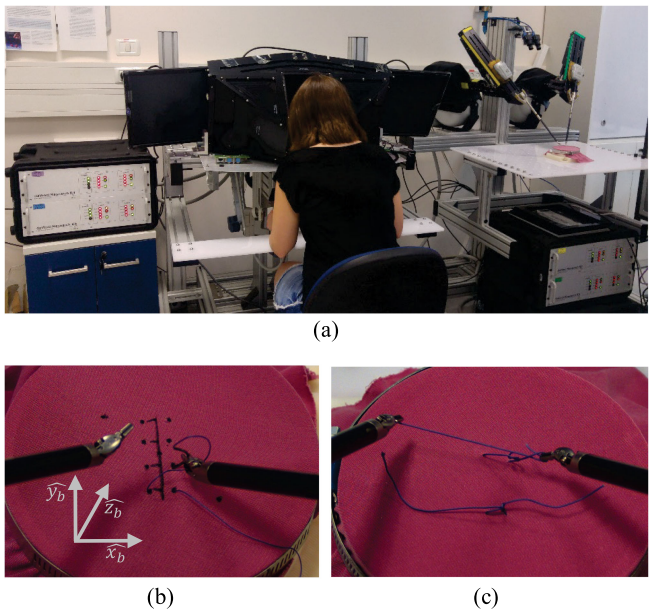

Fig. 2. Collection of data for BBR dataset: (a) A participant in front of the dVRK, (b) A suturing task, (c) A knot-tying task.

never separated a-priori, and we wished to develop networks that would be suitable for analysis of clinical data without preprocessing to separate suturing from knot tying.

B. Gestures Vocabulary

Our gestures vocabulary is summarized in Table I. We started with the JIGSAWS vocabulary of the suturing and knot tying tasks [4]. However, in the clinical-like dataset, there were movements that did not fit into this vocabulary. For example, when a surgeon did an ambiguous movement with both hands. Thus, we created a new "non-specific movement" gesture. In addition, there were gestures that were very similar but not identical to the JIGSAWS vocabulary. For example, transferring the needle from the left to the right hand instead of the right to the left hand. Therefore, we modified the definition of 6 of the gestures (marked with an asterisk in the table) – e.g. a gesture of transferring the needle from one hand to the other. This resulted in 12 gestures, of which 5 were used in the suturing task only – $G_1 - G_5$, 3 in the knot-tying task only – $G_6 - G_8$, and 4 in both tasks – $G_9 - G_{12}$.

C. Preprocessing

1) Features: In all the networks, we used the following 14 features: three endpoint positions in Cartesian coordinates (x, y, z), three linear velocity components ($\dot{x}$, $\dot{y}$, $\dot{z}$), and one gripper angle θ_g, all from both the left and the right PSMs. In the networks that were trained on the clinical-like dataset, we also used the six joint angles of the PSMs q_i; $i = 1...6$, resulting in 26 features overall, We used the labels of the manual annotation into gestures as the output.

TABLE I
GESTURE VOCABULARY

Gesture index	Gesture description	Task	Probability in clinical-like dataset
G1	Positioning needle	Suturing	0.04
G2	Pushing needle through tissue	Suturing	0.08
G3	Transferring needle from one hand to the other*	Suturing	0.02
G4	Pulling suture*	Suturing	0.03
G5	Using one hand to help tighten suture*	Suturing	0.03
G6	Making C loop around one hand*	Knot-tying	0.11
G7	Reaching for suture with one hand while the other holding the needle*	Knot-tying	0.14
G8	Pulling suture with both hands	Knot-tying	0.25
G9	Reaching for needle*	Suturing and knot-tying	0.03
G10	Moving to center with needle in grip	Suturing and knot-tying	0.11
G11	Orienting needle	Suturing and knot-tying	0.14
G12	"Non specific movement"	Suturing and knot-tying	0.02

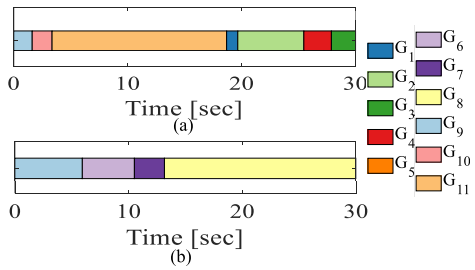

Fig. 3. Segmentation of the tasks: (a) Segmentation of suturing trial, (b) Segmentation of a knot-tying trial.

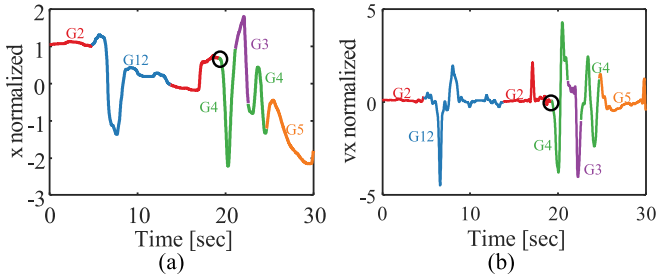

Fig. 4. Decomposed trial after discarding a camera movement segment: (a) Normalized position, (b) Normalized velocity. The circles mark the concatenated segments after discarding a camera movements segment.

In the clinical-like data we calculated the velocity numerically. We filtered the position using a 4^{th} order zero-lag Butterworth filter with a 6 Hz cutoff (Matlab filtfilt()), differentiated the position with respect to time, and filtered the velocity again using a 2^{nd} order zero-lag Butterworth filter with a 10 Hz cutoff.

2) Decomposed Trials: In real surgeries and in our datasets there are two different types of sutures. Running sutures are several consecutive throws (JIGSAWS and BBR datasets). Interrupted sutures are sutures with a knot after each throw (clinical-like dataset). For the network to generalize well to the clinical-like data, we trained on the suturing and the knot-tying tasks together, and decomposed the trajectories into individual throws regardless of the type of sutures.

In the JIGSAWS dataset, to perform the first throw, the surgeons began with reaching for the needle with one hand, then moving to the center with the needle in grip, and orienting the needle. Following these preparation steps, the first and each following throw consisted of positioning the needle, pushing the needle through the tissue, pulling suture with the other hand, and transferring the needle back to the first hand, as depicted in Fig. 3(a). We decomposed the four consecutive throws from each original trial into four trials (one suture per trial), and with five repetitions of the task by each surgeon, we had 156 decomposed suture trials.

Similarly, in the knot-tying task, the surgeons tied two consecutive knots. Each knot consisted of reaching for the needle, making a C loop around one hand, reaching for the suture with one hand while the other hand is holding the needle, and pulling the suture with both hands, as depicted in Fig. 3(b). We decomposed the two consecutive knot-ties into two knot-tying trials, and with five repetitions of the task by each surgeon, we had 72 decomposed knot-tying trials.

We decomposed the BBR and clinical-like datasets in the same way. For the BBR dataset this procedure resulted in 56 and 30 decomposed trials of suturing knot-tying respectively. For the clinical-like dataset, this resulted in 26 and 97 decomposed trials of suturing and knot-tying, respectively.

3) Discarded Segments: In all three datasets, we discarded some gestures. In the JIGSAWS and the BBR datasets we discarded a synthetic gesture that marked the end of the trial and is not performed in real surgical procedures, "dropping suture at the end and move to endpoints". After consulting with an expert surgeon, we also discarded the rare "loosening more suture" gesture. These segments appeared either at the end or at the beginning of our decomposed trials, so cutting them out did not require any further adjustment to the datasets.

In the clinical-like data, we discarded camera movement segments. We also discarded segments of cutting the string that did not occur in the JIGSAWS dataset. After cutting these segments, we re-concatenated the decomposed trial. To assure continuity, we re-centered the second gesture such that the position of the beginning of the second gesture begins at the end of the first gesture. For the gripper and six joint angle features, we concatenated the two segments, discarded three samples from each end, and created new samples using Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) [53]. Fig. 4 shows an example

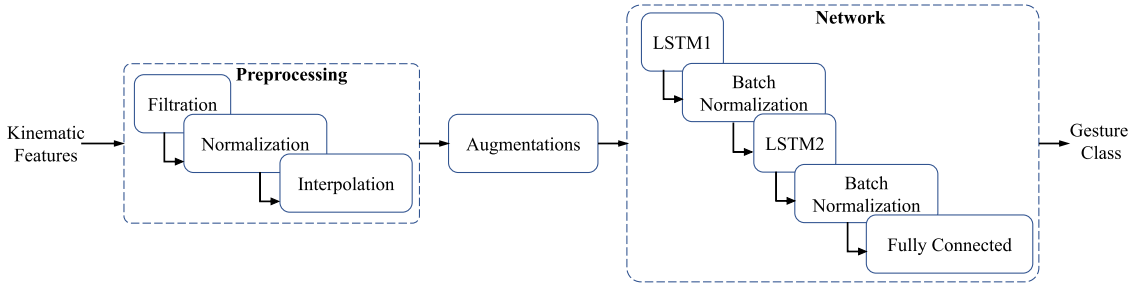

Fig. 5. Framework diagram.

of a re-concatenated trial after discarding a segment of camera motion.

4) Interpolation and Normalization: The execution of each trial takes a different time. Therefore, to ensure that all trials had the same number of samples, we used the Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation [53], which produces smooth trajectories while preserving their shape. Moreover, different features have different scaling and centering, and hence, we z-normalized the data:

$$f_z = \frac{f - \mu_f}{\sigma_f} \tag{1}$$

where f is the feature, μ_f is the mean of the feature (sample average), and σ_f is the standard deviation of the feature.

D. Network Architecture, Training, and Testing

We treated the segmentation problem as a classification problem of the entire time series of the tasks into one of the 12 gestures. We followed our previous work [21] and constructed an RNN, depicted in Fig. 5. The input layer received the time samples with the kinematic features. It was connected to two hidden layers of bidirectional LSTM, with 1024 and 512 neurons, respectively. We chose these hyper-parameters using the validation set. In the output layer, we used softmax logistic regression to produce a distribution of probability over the gestures and classify each sample. To avoid over-fitting the training data, we had layers of batch normalization between the hidden layers [43] and we used dropout.

To take advantage of the JIGSAWS dataset, we used three different approaches: (1) to train a network using the JIGSAWS dataset and test the generalization, (2) to use transfer learning, and (3) to use the architecture of the JIGSAWS network but to train it on the clinical-like data. With this rationale, we trained 10 networks as summarized in Table II:

1) JIGSAWS Network: trained with $N_{surgeons} = 7$ of the JIGSAWS dataset using the 14 kinematic features.

2) Augmented JIGSAWS Networks: five networks that were identical to the JIGSAWS network, but were trained with augmented data as follows.

Data with rotation: Based on our results in [21], we expected that augmentation with rotated data would improve generalization. We concatenated the training set with 12 rotated versions of it. We rotated the training set by left multiplying the position and velocity data with a rotation matrix $R_{\hat{k}_t}(\theta_t)$, where $\hat{k}_t$, θ_t are the axis and angle of the rotation. The gripper angle was left unaltered. We rotated about each of the three principal axes, such that $\hat{k}_{a1} = \hat{x}$, $\hat{k}_{a2} = \hat{y}$, and $\hat{k}_{a3} = \hat{z}$. We chose the angles to be equally spaced in the (0,360) degrees rane, i.e. $72, 144, 216, 288$ degrees. Fig. 6 shows examples of a rotation of a gesture about the $\hat{x}$ axis.

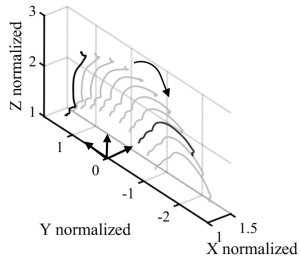

Fig. 6. Augmentation example endpoint positions of a gesture after different rotation about the $x-axis$, the black curve is the original segment, the light grey are example for rotations and the dark grey is a rotation that we used in our training set.

Data exchanged between right and left hands: In the JIGSAWS dataset the surgeons were told which hand to use for each action. In contrast, in the clinical-like dataset, some surgeons performed the suturing using the same hand, while others performed the suturing using one hand to insert the needle and the other hand to pull out the needle. Some used their left hand and others their right hand. Thus, to create a varied dataset, we augmented the JIGSAWS dataset with copies in which we switched the features of left and right hands and preserved the labels.

Data exchanged between right and left hands and rotation: To test if the combination of two augmentations can improve the overall performance compared to using just one, we combined the previous two augmentations.

Time warping: Because different surgeons can perform different gestures faster or slower, we expected that disturbing the temporal location of the samples could improve generalization. We used [54] to create 200 additional versions of the data an augmented version for each trial.

Window slicing: In images, cropping is a widely used augmentation, and in time series, window slicing is similar to cropping. We followed [54] and randomly extract segments from our data of 0.9 of the length of a segment.

TABLE II
MODELS

Model		training data	test data	features	architecture	hyper parameters
Simple JIGSAWS network		JIGSAWS: $N_{surgeons} = 7$, $N_{trials} = 200$	JIGSAWS: $N_{surgeons} = 1$, $N_{trials} = 28$, BBR: $N_{surgeons} = 2$, $N_{trials} = 86$, Clinical-like: $N_{surgeons} = 6$, $N_{trials} = 123$	(x, y, z), $(\dot{x}, \dot{y}, \dot{z})$ θ_g from both the left and the right PSMs	2 layers of bidirectional LSTM	learning rate – 0.01, batch size – 5, dropout with a probability – 0.5
Augmented JIGSAWS network	Rotated data	JIGSAWS: $N_{surgeons} = 7$, $N_{trials} = 2600$	JIGSAWS: $N_{surgeons} = 1$, $N_{trials} = 28$, BBR: $N_{surgeons} = 2$, $N_{trials} = 86$, Clinical-like: $N_{surgeons} = 6$, $N_{trials} = 123$	(x, y, z), $(\dot{x}, \dot{y}, \dot{z})$ θ_g from both the left and the right PSMs	2 layers of bidirectional LSTM	learning rate – 0.01, batch size – 5, dropout with a probability – 0.5
	Data exchanged between hands	JIGSAWS: $N_{surgeons} = 7$, $N_{trials} = 266$				
	Rotated data and data exchanged between hands	JIGSAWS: $N_{surgeons} = 7$, $N_{trials} = 2666$				
	Data with time warping	JIGSAWS: $N_{surgeons} = 7$, $N_{trials} = 400$				
	Data with window slice	JIGSAWS: $N_{surgeons} = 7$, $N_{trials} = 400$				
Transfer network		pre-trained: JIGSAWS: $N_{surgeons} = 7$, $N_{trials} = 2600$ train: Clinical-like: $N_{surgeons} = 5$, $N_{trials} = 107$	Clinical-like: $N_{surgeons} = 1$, $N_{trials} = 16$	(x, y, z), $(\dot{x}, \dot{y}, \dot{z})$ θ_g from both the left and the right PSMs	Pre-trianed: 2 layers of bidirectional LSTM train: 1 layers of bidirectional LSTM	learning rate – 0.05, batch size – 2, dropout with a probability – 0.4
Simple clinical-like network	JIGSAWS features	Clinical-like: $N_{surgeons} = 5$, $N_{trials} = 107$	Clinical-like: $N_{surgeons} = 1$, $N_{trials} = 16$	(x, y, z), $(\dot{x}, \dot{y}, \dot{z})$ θ_g from both the left and the right PSMs	2 layers of bidirectional LSTM	learning rate – 0.05, batch size – 10, dropout with a probability – 0.5
	Clinical-like features	Clinical-like: $N_{surgeons} = 5$, $N_{trials} = 107$		(x, y, z), $(\dot{x}, \dot{y}, \dot{z})$ θ_g q_i ; $i = 1...6$ from both the left and the right PSMs		
Augmented clinical-like network		Clinical-like: $N_{surgeons} = 5$, $N_{trials} = 1391$	Clinical-like: $N_{surgeons} = 1$, $N_{trials} = 16$	(x, y, z), $(\dot{x}, \dot{y}, \dot{z})$ θ_g q_i ; $i = 1...6$ from both the left and the right PSMs	2 layers of bidirectional LSTM	learning rate – 0.05, batch size – 10, dropout with a probability – 0.5

In the different augmentations, we added the augmented trials to the original dataset and used this larger dataset as the training data which resulted in different sizes of augmented training sets as can be seen in Table II.

3) Transfer Network: We used the Augmented JIGSAWS network with rotation as the pre-trained network with its weights. Then we removed its last layer and added a new layer of bidirectional LSTM with 512 neurons, one layer of batch normalization and a output layer with softmax for the classification. We then trained this transfer network with $N_{surgeons} = 5$ from the clinical-like dataset.

4) Simple Clinical-Like Network: We trained two such networks. For both of them we used the architecture of two layers of bidirectional LSTM as explained in II-D, and trained those networks with $N_{surgeons} = 5$ from the clinical-like dataset. For one network we used 14 kinematic features, the same features that were used in the networks that were trained with the JIGSAWS dataset. For the second network we used the 26 kinematic features of the clinical-like dataset.

5) Augmented Clinical-Like Network: We used an identical network to the simple clinical-like network, but here we trained it with a larger augmented dataset. We used $N_{surgeons} = 5$, and we rotated this dataset to create rotated versions of it as explained in II-D2.

E. Evaluation of Classification Performance

We tested the simple and augmented networks with one left out test surgeon from the JIGSAWS dataset, two participants from the BBR dataset, and six surgeons from the clinical-like dataset. We tested the transfer, simple clinical-like, and augmented clinical-like networks with one left out test surgeon from the clinical-like dataset (see Table II).

We tested the performance of the network using 3 metrics: Accuracy, F_1 score, and AUC. Accuracy measures the sample-by-sample agreement between the true gestures and the network-predicted gestures. The F_1 score combines the Recall, the proportion of correctly identified samples out of all the

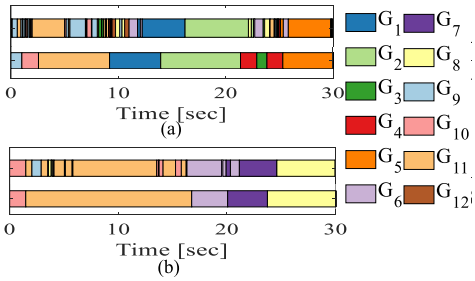

Fig. 7. Example of the time series classification of gestures in comparison to the true lable for a single trial of a test surgon from the clinical-like dataset: (a) An example of a time series classification of gestures of a single suturing trial of the test surgeon (b) An example of a time series classification of gestures of a single knot-tying trial of the test surgeon.

true samples of each gesture, and Precision, the proportion of the correctly identified samples out of all the samples that the network classified as each gesture:

$$F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2}$$

The F_1 score is an important performance measure in cases of uneven class distributions.

The AUC measures the area under the Receiver Operating Characteristic (ROC) curve. We calculated the ROC curve for each gesture versus all, the AUC of each of these curves, and the micro-average AUC, which is known as a good measurement for classification of imbalanced data.

To assess the repeatability of our results, we repeated the training and testing of the two best performing networks five times, and report the mean and standard deviation of the measures.

We also compared the performance of our best performing JIGSAWS network that was augmented with rotation to previously published approaches that were tested on the JIGSAWS dataset [24], [25], [36]. None of these studies trained their network on both suturing and knot-tying together, and they only reported accuracy and $F1$ score measures. Hence, We computed the accuracy and $F1$ score of our network on gestures $G1 - G5$ and $G9 - G11$ – the suturing gestures alone.

III. Results

In Fig. 1, we can see the three different datasets, both the endpoint position of the left and right PSMs and an image from each dataset. When examining the kinematics, it is hard to notice a substantial difference between the three datasets; however, they are not identical. The clinical-like dataset is more complex; watching the video of this dataset (not included with paper) taught us that the tissue is not planar, and it moves due to breathing. In contrast, the elastic cloth in the JIGSAWS and BBR datasets is planar and does not move. The use of an elastic cloth creates an additional difference: the surgeon has to insert the needle into the cloth and pull it out, but between those two actions the surgeon moves in air. In contrast, when suturing real tissue, the surgeon moves inside the tissue, which is harder than moving in air.

Fig. 7 shows an example of classification into gestures of a single suturing and knot-tying trial from the clinical-like dataset with the Augmented clinical-like network. This network succeeded in classifying the gestures correctly for the two tasks, but it achieved better results for knot-tying. The clinical-like dataset was unbalanced – the task demanded one suture followed by three or four knot-ties. Therefore, the network had more trials of knot-tying than suturing, which lead to better performance for the knot-tying task.

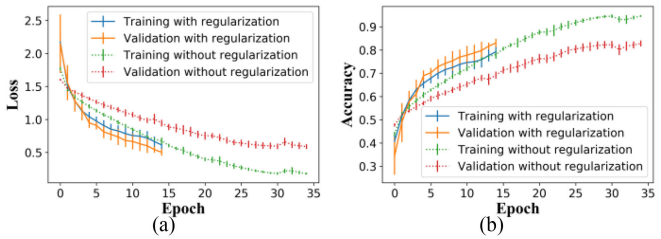

Fig. 8. Model training vs. validation with and without regularization: (a) Loss, (b) Accuracy.

Fig. 8 shows the average and the standard deviation of the loss and the accuracy of the training and validation sets with and without regularization over five repetitions of this network. Training without regularization shows overfitting the training data and worse results for the validation set, which does not happen with regularization. With regularization, the performance of the validation set is better than that of the training set. This could stem from several reasons; for example, that the regularization is applied on the training set only.

A summary of the performance of all the networks is presented in Table III. Next, we discuss the performance of the different networks:

A. Simple JIGSAWS Network

Fig. 9(a) depicts the confusion matrix over one surgeon from the JIGSAWS dataset. This network classified well gestures $G_1 - G_4$, $G_7 - G_{10}$, but not gestures G_5, G_{11}. The network classified well most of the suturing gestures except for G_5 (square 1), and most of the knot-tying gestures except for G_6 (square 5). It classifies worse gestures that appear in both tasks (square 9). This is likely due to the latter being preparation gestures, and as mentioned in II-C2, these gestures do not appear in all decomposed trials, which leads to few examples that the network can learn from and worse performance.

Fig. 9(b), (c) depict the confusion matrices of the same network over the two participants of the BBR and the clinical-like datasets, respectively. These matrices are not diagonal and depict poor classification. This network could not learn the kinematic patterns for both generalization datasets, and the classification into gestures reflected mostly the frequency of the gestures in the training dataset see Table I. Therefore, we conclude that the Simple JIGSAWS network can not generalize to new or more complex datasets.

B. Augmented JIGSAWS Networks

Fig. 9(d) depicts the confusion matrix of the network that was trained with rotation augmentation, tested with the JIGSAWS

TABLE III
AUGMENTATION RESULTS

Test set		JIGSAWS			BBR			Clinical-like dataset		
Method		Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Simple JIGSAWS network		0.56	0.57	0.89	0.09	0.08	0.50	0.13	0.12	0.58
Augmented JIGSAWS network	Rotation	**0.64 ± 0.017**	**0.67 ± 0.026**	**0.93 ± 0.009**	**0.52 ± 0.02**	**0.63 ± 0.017**	**0.90 ± 0.017**	**0.22 ± 0.017**	**0.24 ± 0.024**	**0.67 ± 0.012**
	hand exchange	0.44	0.45	0.87	0.12	0.12	0.80	0.12	0.11	0.65
	Rotation, hand exchange	0.57	0.60	0.87	0.39	0.46	0.8	0.15	0.15	0.65
	Time warping	0.46	0.45	0.86	0.13	0.13	0.54	0.14	0.15	0.60
	Window slice	0.52	0.52	0.88	0.08	0.08	0.50	0.17	0.18	0.63
Transfer network		-	-	-	-	-	-	0.45	0.52	0.81
Simple clinical-like network	JIGSAWS features	-	-	-	-	-	-	0.45	0.49	0.80
	clinical-like features	-	-	-	-	-	-	059	0.6	0.84
Augmented clinical-like network		-	-	-	-	-	-	**0.68 ± 0.02**	**0.69 ± 0.02**	**0.96 ± 0.005**

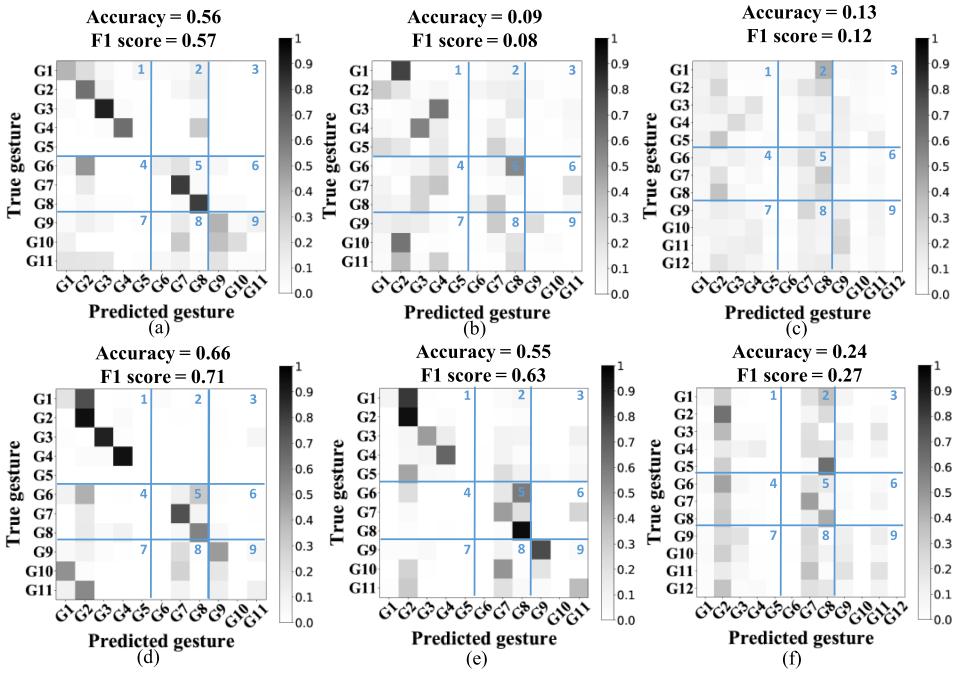

Fig. 9. Confusion matrices for networks that were trained with the JIGSAWS dataset (with and without augmentation) and tested on the JIGSAWS, BBR and clinical-like test sets: (a) a confusion matrix for the network that was trained without augmentations and tested with the JIGSAWS dataset, (b) a confusion matrix for the network that was trained without augmentations and tested with BBR dataset, (c) a confusion matrix for the network that was trained without augmentations and tested with clinical-like dataset, (d) a confusion matrix for the network that was trained with rotation and tested with the JIGSAWS dataset, (e) a confusion matrix for the network that was trained with rotation and tested with BBR dataset, (f) a confusion matrix for the network that was trained with rotation and tested with clinical-like dataset.

dataset, and reached the best results. Comparing to the network that was trained without rotation (Fig. 9(a)) suggests that the matrix of the network that was trained with rotation is closer to being diagonal, implying better classification of gestures. Specifically, the classification of gestures $G_2 - G_4, G_9, G_{11}$ improved when rotation was added. Moreover, squares 2, 3 in Fig. 9(d) show that the network confuses less between the suturing gestures to the knot-tying gestures and gestures which appeared in both tasks. However, in spite of the general superiority of the rotation augmented network, the classification of gestures G_1, G_8, G_{10} declined following the addition of this augmentation set.

Fig. 9(e) depicts the confusion matrix of the same network tested with the BBR dataset. Comparing to the network that was trained without rotation (Fig. 9(b)) suggest that adding rotation augmentation achieved noticeably better results, which are comparable with the results obtained for the JIGSAWS test set. This network classified well gestures $G_2 - G_4, G_7 - G_9, G_{11}$, and in some trials, also gesture G_1. This corroborates our previous results from the suturing task alone [21] that adding rotation augmentation improves generalization to a new dataset and even improves the performance on the original test set.

In contrast, adding rotated data did not improve the classification of the clinical-like dataset (Fig. 9(f)). These results are noticeably worse than the results on the two other test sets, and below chance level. We conclude that the Augmented JIGSAWS network that was trained with rotation generalizes well to similar but not to more complex data.

Adding to the rotation augmentation sets with switching between the hands did not further improve the results for all three test sets, and in fact, the performance declined. Similarly,

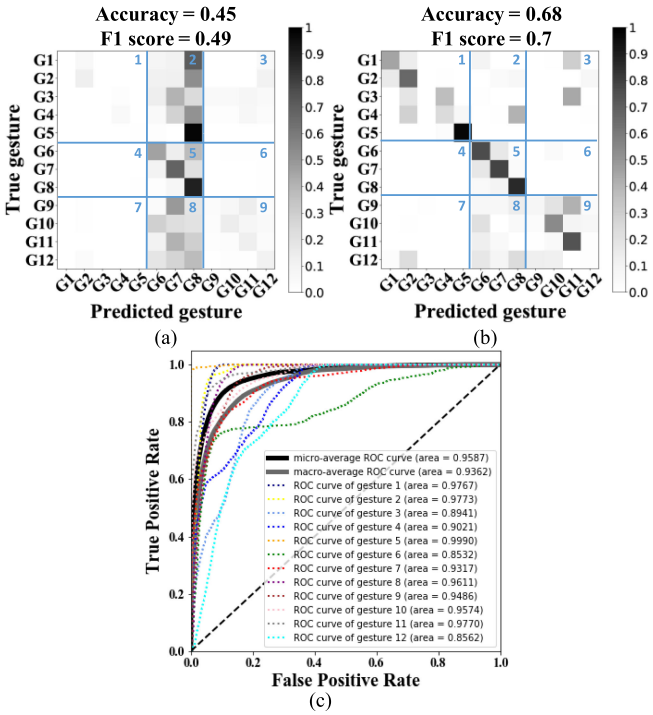

Fig. 10. Confusion matrices for networks that were trained with the clinical-like dataset and tested on the clinical-like test sets: (a) a confusion matrix for the Transfer network; (b) a confusion matrix for one repetition of the network that was trained with rotation; and (c) the ROC curves for one repetition of the network that was trained with rotation.

augmentation by switching between the hands, time warping, or window slicing alone did not improve classification.

We compared the classification of suturing gestures by the Augmented JIGSAWS network that was trained with rotation to other surgical gesture classification studies [24], [25], [36]. Our model achieved an accuracy of 0.75 and an $F1$ score of 0.78 on the suturing gestures. These results are better than [24] (accuracy 0.73), but not as good as [25] (accuracy 0.83) and [36] (accuracy 0.85, $F1$ score 0.88). Note that these studies trained their models on the suturing task only. Our network that was trained on both suturing and knot tying had to learn a more complex model and was in a disadvantage, but this allows for working with more complex data such as the clinical-like dataset that has both tasks together.

C. Transfer Network

In this network, we used the clinical-like dataset in the training phase. Since our goal is to create a network that will work for clinical data, we did not test its performance for the JIGSAWS and BBR datasets. Adding a layer that was trained on the clinical-like data exposed the network to complex data and forced it to learn features that are related to these tasks, and improved the classification. However, the improved result did not reach the levels of classification of the other two datasets. Fig. 10(a) depicts the confusion matrix of this network and shows that this network learned the complex clinical-like patterns, but the network could classify well mostly gestures from the knot-tying task - $G_6 - G_8$ and classified worse gestures from the suturing task, or gestures from both tasks.

D. Simple Clinical-Like Network

To further improve the classification, and when we used only the clinical-like dataset which has more features, we added the six joint angles of the PSM's to the training features. Comparing the results of this network to a network that was trained with the clinical-like dataset but without the addition of the joint angles shows that the joint angles added relevant information which improved the classification.

E. Augmented Clinical-Like Network

Fig. 10(b) depicts the confusion matrix of the Augmented clinical-like network that was trained with the clinical-like data with rotation and reached the best results, which are comparable with the results achieved for the JIGSAWS dataset. This network classified well gestures $G_1, G_2, G_5 - G_8, G_{10}, G_{11}$, but not gestures G_3, G_{12} (see Discussion). This network classified well both suturing and knot-tying; however, it classified the knot-tying gestures better, consistently with the example in Fig. 7. Examination of squares 4,6 in Fig. 10(b) shows that the network did not classify knot-tying gestures as suturing gestures or as gestures which appear in both tasks. Squares 2,3 shows almost similar results for the suturing gestures.

Fig. 10(c) depicts the ROC curves for the different gestures and the micro and macro average ROC curves. Similarly to Fig. 10(b), it shows that the network classifies well gestures $G_1, G_2, G_5, G_8, G_{10}, G_{11}$, but not gestures G_3, G_{12}. Gesture G_6 achieved low AUC, although the network succeeded in classifying it as seen in Fig. 10(b) (see Discussion).

IV. Discussion

Towards developing deep neural networks for efficient segmentation of kinematic data from real surgical procedures, we explored how to generalize algorithms that were developed, trained, and tested on structured dry-lab datasets to real surgical data. We started with a previously reported LSTM network for classification of gestures [21], [25], and experimented with different approaches, including augmentation, transfer learning, training on clinical-like data, and adding features, to improve its classification performance on three different structured suturing and knot-tying datasets.

The availability of real surgical data is a bottleneck in surgical data science [5], and even data of training on pigs such as the clinical-like dataset of this paper is not accessible. In contrast, the JIGSAWS dataset is publicly available, reproducible (as we did in the BBR dataset) and has been well studied for gesture classification [20], [21], [25], [33]. Hence, it is important to understand what is the best way to utilize this dataset for developing algorithms that generalize well to clinical data, but to the best of our knowledge, such generalization to clinical-like data using kinematic data only has never been presented.

As mentioned earlier, contrary to real surgical cases, the JIGSAWS dataset is structured: the surgeons were asked to perform

a sequence of gestures, and were not allowed to move the camera or to use the clutch. This creates similarities between all surgeons and trials. The clinical-like dataset also has a structured task, but the surgeons performed this task as they preferred: with their right hand or their left hand and using different techniques. They also moved the camera and used the clutch. Real surgical data can be completely unstructured; in addition to the differences in the implementation between different surgeons, the tasks may be different as well. Although the clinical-like dataset is less complex than real surgical data, we expect that a gestures classification algorithm that will achieve good performance for it may generalize well also to real surgical data.

It is important to note that the clinical-like data was not labeled, so a manual data annotation was required before we could use it. While we were annotating the data according to the criteria of the JIGSAWS annotation, we noticed that the known gesture vocabulary of the JIGSAWS dataset might be less suitable for other datasets. During the annotation, we encountered segments that needed slightly different gestures than to the gestures from the vocabulary, and segments that could not be labeled using the known gesture vocabulary. Furthermore, as reported in [29], they found that re-segmentation of the JIGSAWS dataset itself according to some other criteria leads to improved performance. Therefore, adding more criteria to the manual segmentation and creating a slightly different gesture vocabulary should be considered while working with the JIGSAWS dataset, and with a more complex dataset. Moreover, noise at the boundaries between segments can influence classification performance and is impossible to detect automatically.

We assessed the performance of the different networks in this paper with common metrics - accuracy, F_1, and AUC, as well as by examining the confusion matrices and the ROC curves. In general, the gestures that the network succeeded in classifying as seen in the confusion matrix also result in better ROC curves and higher AUC values. However, gesture G_6, which the network was able to classify, did not result in a good ROC curve. These results are caused by the fact that the network succeed in finding the correct G_6 gestures, but at the same time misclassify many other gestures as G_6, which caused a worse ROC curve with a lower AUC value. In contrast, there are gestures that the network almost could not classify with high AUC values. This is probably because to calculate the AUC in our multiclass classification problem we used one vs all classification, which is an easier task than classifying the different gestures.

Among the augmentations we tried, adding rotated versions of the data to the training set performed best. However, even though this method improved the performance when tested on the JIGSAWS and the BBR datasets, it did not achieve satisfactory results when tested on the clinical-like dataset. This leads us to the conclusion that the network cannot generalize to a more complex dataset. Therefore, we used the knowledge of the adequate architecture achieved from working with the JIGSAWS dataset, and tried different methods to use the clinical-like dataset in the training stage. This exposes the network to complex data and forces it to learn features that appear in the clinical-like data and cannot be learned from a simple dataset. Our best results were achieved by adding features of six joint angles of the PSMs and training the network using the clinical-like dataset with rotation augmentation. The JIGSAWS dataset does not have the joint angles data, and therefore, we could not add them to the networks that were trained with the JIGSAWS dataset.

Adding rotation to all networks generally improves the performance; however, when testing the Augmented JIGSAWS network with the JIGSAWS dataset, we saw that the classification of gestures G_1, G_8, G_{10} slightly declined. We notice a similar behavior when testing the Augmented clinical-like network with the clinical-like dataset, where we saw that the classification of gestures $G_1, G_3.G_9$ slightly declined as well. A possible reason for this is that these gestures are less common in these datasets. Hence, when we used rotation augmentation which increases the training dataset, it causes the probability of these gestures to decrease and therefore, they were less common in the datasets which can cause the network performance to decline.

We expected that adding the exchanging between the hands augmentation would improve the classification of the clinical-like dataset and not alter the classification of the JIGSAWS and the BBR test sets. However, this augmentation did not improve performance, and the classification of the JIGSAWS test set declined. This is probably because contrary to our expectation, exchanging the features of the right and left hands did not result in trajectories that resemble performing the task with the opposite hand. As a result this augmentation just cluttered the network with irrelevant information.

Time warping and window slice augmentations also failed to improve classification. We think that this is related to the nature of the added diversity by these augmentations. Time warping and window slice augmentations add variations in the trajectories that resemble variations due to personal style, expertise, fatigue, and others. The JIGSAWS dataset is small, but it consists of tasks produced by different surgeons with different expertise. Our results suggest that possibly the variation in the trajectories could already be well represented in this dataset. In contrast, rotation adds variations to the paths that resemble performing the suturing and knot tying in different orientations. In the JIGSAWS dataset, all the surgeons performed the task at the same orientation, and we believe that the improvement and generalization to other datasets is due to this added diversity. Another less plausible reason for the advantage of rotation augmentation is due to the large size of the added data. This is less likely because when we combined two augmentations – rotation and data exchange between hands – the performance did not improve and rather got worse. Therefore, we did not test other combinations of augmentations and believe that the improvement is a result of the rotations themselves rather than the amount of data added.

The overall performance of our network for the JIGSAWS dataset is worse than that reported in [25], although we both used bidirectional LSTM. This is probably because we created one network to classify both the suturing and the knot-tying tasks and we did not create separate networks for each task. When we used different networks for different tasks, as we did in [21], we achieved better results. Another possible reason is that they train their network over a larger dataset, which we did not have access to. Moreover, the confusion matrices reveal that certain gestures

are poorly recognized (e.g. transferring needle between hands and reaching for needle), but for practical applications these are not necessarily the most important gestures. Therefore, for applications such as training or surgical decision support even our moderate performance levels are impactful.

In recent years, gesture classification algorithms that used the JIGSAWS dataset were reported which achieved high performance [26], [55], [56]; however, very few studies reported using additional datasets other than JIGSAWS [56], [57]. Two other recent studies [41], [58] developed networks for clinical data, but they used different inputs, tasks, or outputs; hence, we could not compare our networks to these studies. Importantly, our focus here was on creating a network that can classify the clinical-like dataset while using the knowledge we gathered during the work with the simple JIGSAWS dataset. Therefore, instead of looking for a network that will achieve better performance for the JIGSAWS dataset, we focused on studying different ways to generalize the classification to the clinical-like dataset. We showed that algorithms that trained with the JIGSAWS dataset do not work with clinical-like data, and hence we conclude that they are not likely to work with real surgical data, and may demand adjustments to real surgical data. Therefore, one must aim to work with data that is as close as possible to real surgical data, but may also implement the knowledge obtained from the gesture classification algorithms that are trained with simpler datasets.

We chose to use only kinematic data and LSTM network to allow online segmentation in the future. However, we use bidirectional LSTM layers, which means that the network uses information from both the past and the future. Hence, our network will need some adjustments before online segmentation. Other future extensions would be to try different networks such as LCRN [37], which can learn both the spatial information using the CNN, and temporal information using the RNN. This has the potential to enrich the network with more important information when comparing the LSTM networks that can learn only temporal information, and may improve the network performance for all three datasets. Zia *et al.* [39], compared between the performance of a network that combines CNN and RNN to networks that use solely CNN or solely RNN for surgical task identification. The LCRN network outperformed other networks, which suggests that in surgical data, both spatial and temporal information are important and that the combination of both can lead to the best performance in surgical gesture classification as well.

There are several limitations in our work. Our dataset is very small, and when using a small dataset, the performance of deep learning algorithms may be compromised compared to a large dataset. In addition to that, we also used the test set several times. Moreover, we cannot be sure that our algorithm will work when it will be tested over a real surgical data. However, even if it will not achieve high performance in data from real cases, it will still be useful for surgical training that is executed as a structured task which our network can classify, and for research.

It is important to note that there are several impeding factors to publishing clinical data; in contrast, the JIGSAWS dataset is available, and similar structured data can be recorded using a suitable robot such as the da Vinci Research Kit (dVRK) [52]. Hence, even though networks that were trained on the JIGSAWS data were not successful in segmentation of the clinical-like data, using structured data, such as the JIGSAWS dataset, is an important first step. Furthermore, generalizing these algorithms to also classify clinical data which are more complex than the structured data, can have a notable influence on the development of surgical robotics.

V. Conclusion

We created a gesture classification network for a complex dataset in RAMIS. We demonstrate a problem of poor generalization to clinical-like data of an LSTM network that was trained on a structured and not sufficiently diverse dataset. We proposed data augmentation to improve the robustness to new data, and although this improved the robustness to new structured data, it did not perform well for the clinical-like data. Similarly, transfer learning also did not sufficiently improve performance on the clinical-like data. However, we implemented the knowledge from working with the JIGSAWS dataset and proposed a network that had a similar structure and augmentation, but was trained using clinical-like data. This network outperforms all other tested networks and achieved the best performance for the clinical-like data. To conclude, we highlighted the importance of using clinical data in gesture classification algorithms in RAMIS as well as the benefit of using information that is learned from working with publically available datasets.

Acknowledgment

The author would like to thank Maya Chuchem for her advise and valuable insights during the revision of the paper.

References

[1] J. Rassweiler, M. Hruza, D. Teber, and L.-M. Su, "Laparoscopic and robotic assisted radical prostatectomy critical analysis of the results," *Eur. Urol.*, vol. 49, no. 4, pp. 612–624, Apr. 2006.

[2] K. Moorthy *et al.*, "Dexterity enhancement with robotic surgery," *Surg. Endosc.*, vol. 18, no. 5, pp. 790–795, May 2004.

[3] J. C. Byrn *et al.*, "Three-dimensional imaging improves surgical performance for both novice and experienced operators using the da Vinci Robot System," *Amer. J. Surg*, vol. 193, no. 4, pp. 519–522, Apr. 2007.

[4] Y. Gao *et al.*, "JHU-ISI gesture and skill assessment working set (JIGSAWS) : A surgical activity dataset for human motion modeling," in *Med. Image Comput. Comput. Assist. Interv.*, Boston, MA, USA, Sep. 2014, pp. 1–10.

[5] L. Maier-Hein *et al.*, "Surgical data science for next-generation interventions," *Nat. Biomed. Eng.*, vol. 1, no. 9, pp. 691–696, Sep. 2017.

[6] J. Rosen, J. D. Brown, L. Chang, M. N. Sinanan, and B. Hannaford, "Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete markov model," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 399–413, Mar. 2006.

[7] I. Nisky, A. M. Okamura, and M. H. Hsieh, "Effects of robotic manipulators on movements of novices and surgeons," *Surg. Endos.*, vol. 28, no. 7, pp. 2145–2158, Jul. 2014.

[8] I. Nisky, M. H. Hsieh, and A. M. Okamura, "Uncontrolled manifold analysis of arm joint angle variability during robotic teleoperation and freehand movement of surgeons and novices," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 12, pp. 2869–2881, Dec. 2014.

[9] I. Nisky, Y. Che, Z. F. Quek, M. Weber, M. H. Hsieh, and A. M. Okamura, "Teleoperated versus open needle driving: Kinematic analysis of experienced surgeons and novice users," in *Proc. IEEE Int. Conf. Robot. Autom.*, Seattle, Washington, 2015, pp. 5371–5377.

[10] Y. Sharon and I. Nisky, "Expertise, teleoperation, and task constraints affect the speed-curvature-torsion power law in RAMIS," *J. Med. Robot. Res.*, vol. 3, no. 3/4, Jul. 2018, Art. no. 1841008.

[11] Y. Sharon, T. S. Lendvay, A. Jarc, and I. Nisky, "Rate of orientation change as a new metric for robot-assisted and open surgical skill evaluation," *IEEE Trans. Med. Robot. Bionics*, vol. 3, no. 2, pp. 414–425, May 2021.

[12] S. Estrada, C. Duran, D. Schulz, J. Bismuth, M. D. Byrne, and M. K. O'Malley, "Smoothness of surgical tool tip motion correlates to skill in endovascular tasks," *IEEE Trans. Hum.-Mach. Syst*., vol. 46, no. 5, pp. 647–659, Oct. 2016.

[13] N. Enayati, G. Ferrigno, and E. De Momi, "Skill-based humanrobot cooperation in tele-operated path tracking," *Auton. Robots.*, vol. 42, no. 5, pp. 997–1009, Jun. 2018.

[14] B. Poursartip, M. LeBel, R. V. Patel, M. D. Naish, and A. L. Trejos, "Analysis of energy-based metrics for laparoscopic skills assessment," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 7, pp. 1532–1542, Jul. 2018.

[15] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," *Comput. Aided Surg.*, vol. 11, no. 5, pp. 220–230, Sep. 2006.

[16] M. Bonfe *et al.*, "Automated surgical task execution: The needle insertion case," in *Proc. 3rd Joint Workshop New Technol. Computer/Robot Assist. Surg.*, Verona, Italy, 2013, pp. 45–47.

[17] N. Preda *et al.*, "A cognitive robot control architecture for autonomous execution of surgical tasks," *J. Med. Robot. Res.*, vol. 1, no. 4, Aug. 2016, Art. no. 1650008.

[18] G. P. Moustris, S. C. Hiridis, K. M. Deliparaschos, and K. M. Konstantinidis, "Evolution of autonomous and semi-autonomous robotic surgical systems: A review of the literature," *Int. J. Med. Robot*., vol. 7, no. 4, pp. 375–392, Aug. 2011.

[19] E. Dolph, C. Krause, and D. Oleynikov, "Future robotic systems: Microrobotics and autonomous robots," in *Proc. Robot.-Assist. Minimally Invasive Surg*. Springer, 2019, pp. 329–335.

[20] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. Eur. Conf. Comput. Vis.,* Amsterdam, The Netherlands: Springer, 2016, pp. 47–54.

[21] D. Itzkovich, Y. Sharon, A. Jarc, Y. Refaely, and I. Nisky, "Using augmentation to improve the robustness to rotation of deep learning segmentation in robotic-assisted surgical data," in *Proc. IEEE Int. Conf. Robot. Autom*. Montreal, Canada, 2019, pp. 5068–5075.

[22] D. Liu and T. Jiang, "Deep reinforcement learning for surgical gesture segmentation and classification," 2018, *arXiv:1806.08089*.

[23] B. B. Haro, L. Zappella, and R. Vidal, "Surgical gesture classification from video data," in *Proc. Med. Image Comput. Comput. Assist. Interv.*. Nice, France: Springer, 2012, pp. 34–41.

[24] L. Tao, L. Zappella, G. D. Hager, and R. Vidal, "Surgical gesture segmentation and recognition," in *Proc. Med. Image Comput. Comput. Assist. Interv*. Nagoya, Japan: Springer, 2013, pp. 339–346.

[25] R. DiPietro *et al.*, "Recognizing surgical activities with recurrent neural networks," in *Proc. Med. Image Comput. Comput. Assist. Interv*. Athens, Greece: Springer, 2016, pp. 551–558.

[26] D. Sarikaya and P. Jannin, "Surgical gesture recognition with optical flow only," 2019, *arXiv:1904.01143*.

[27] G. Menegozzo, D. Dall'Alba, C. Zandonà, and P. Fiorini, "Surgical gesture recognition with time delay neural network based on kinematic data," in *Proc. Int. Symp. Med. Robot.*, Atlanta, GA, USA, 2019, pp. 1–7.

[28] Y.-Y. Tsai, B. Huang, Y. Guo, and G.-Z. Yang, "Transfer learning for surgical task segmentation," in *Proc. Int. Conf. Robot. Autom*. Montreal, Canada, 2019, pp. 9166–9172.

[29] B. van Amsterdam, H. Nakawala, E. De Momi, and D. Stoyanov, "Weakly supervised recognition of surgical gestures," in *Proc. Int. Conf. Robot. Autom*. Montreal, Canada, 2019, pp. 9565–9571.

[30] N. Ahmidi *et al.*, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2025–2041, Sep. 2017.

[31] M. J. Fard, S. Ameri, R. B. Chinnam, A. K. Pandya, M. D. Klein, and R. D. Ellis, "Machine learning approach for skill evaluation in robotic-assisted surgery," 2016, *arXiv:1611.05136*.

[32] A. Zia, C. Zhang, X. Xiong, and A. M. Jarc, "Temporal clustering of surgical activities in robot-assisted surgery," *Int J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 7, pp. 1171–1178, May 2017.

[33] M. J. Fard, S. Ameri, R. B. Chinnam, and R. D. Ellis, "Soft boundary approach for unsupervised gesture segmentation in robotic-assisted surgery," *IEEE Robot. Autom. Lett.*, vol. 2, no. 1, pp. 171–178, Jan. 2017.

[34] S. Krishnan *et al.*, "Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning," *Int J. Rob. Res.*, vol. 36, no. 13-14, pp. 1595–1618, Nov. 2017.

[35] C. Rupprecht, C. Lea, F. Tombari, N. Navab, and G. D. Hager, "Sensor substitution for video-based action recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* Daejeon, Korea, 2016, pp. 5230–5237.

[36] B. van Amsterdam, M. J. Clarkson, and D. Stoyanov, "Multi-task recurrent neural network for surgical gesture recognition and progress prediction," in *Proc. Int. Conf. Robot. Autom*. Paris, France, 2020, pp. 1380–1386.

[37] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, 2015, pp. 2625–2634.

[38] D. Sarikaya, K. A. Guru, and J. J. Corso, "Joint surgical gesture and task classification with multi-task and multimodal learning," 2018, *arXiv:1805.00721*.

[39] A. Zia, L. Guo, L. Zhou, I. Essa, and A. Jarc, "Novel evaluation of surgical activity recognition models using task-based efficiency metrics," in *Proc. Int J. Comput. Assist. Radiol. Surg.*, 2019, pp. 1–9.

[40] Y.-H. Long *et al.*, "Relational graph learning on visual and kinematics embeddings for accurate gesture recognition in robotic surgery," 2020, *arXiv:2011.01619*.

[41] Y. Qin *et al.*, "Temporal segmentation of surgical sub-tasks through deep learning with multiple data sources," in *Proc. IEEE Int., Conf. Robot. Autom.*, 2020, pp. 371–377.

[42] S. Ramesh *et al.*, "Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures," 2021, *arXiv:2102.12218*.

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[45] J. West, D. Ventura, and S. Warnick, "Spring research presentation: A theoretical foundation for inductive transfer," Brigham Young University, *College Phys. Math. Sci.*, vol. 1, p. 32, Aug. 2007.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[47] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 127, pp. 248–257, Apr. 2016.

[48] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.* Washington, DC, USA, 2018, pp. 289–293.

[49] T. T. Um *et al.*, "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks," 2017, *arXiv:1706.00527*.

[50] A. Le Guennec, S. Malinowski, and R. Tavenard, "Data augmentation for time series classification using convolutional neural networks," in *Proc. Workshop Adv. Analytics Learn. Temporal Data*, Riva Del Garda, Italy, 2016.

[51] Z. Wang and A. M. Fey, "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery," *Int J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 12, pp. 1959–1970, Dec. 2018.

[52] P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio, "An open-source research kit for the da vinci surgical system," in *Proc. IEEE Int Conf. Robot. Autom.*, 2014, pp. 6434–6439.

[53] F. N. Fritsch and R. E. Carlson, "Monotone piecewise cubic interpolation," *SIAM J. Numer. Anal.*, vol. 17, no. 2, pp. 238–246, 1980.

[54] B. K. Iwana and S. Uchida, "Time series data augmentation for neural networks by time warping with a discriminative teacher," 2020, *arXiv:2004.08780*.

[55] I. Gurcan and H. Van Nguyen, "Surgical activities recognition using multiscale recurrent networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 2887–2891.

[56] R. DiPietro *et al.*, "Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks," *Int. J. Comput. Assist. Radiol. Surg.*, 2019, pp. 1–16.

[57] A. M. Jarc, A. A. Stanley, T. Clifford, I. S. Gill, and A. J. Hung, "Proctors exploit three-dimensional ghost tools during clinical-like training scenarios: A preliminary study," *World J. Urol.*, vol. 35, no. 6, pp. 957–965, 2017.

[58] Y. Qin, S. Feyzabadi, M. Allan, J. W. Burdick, and M. Azizian, "daVinciNet: Joint prediction of motion and surgical state in robot-assisted surgery," 2020, *arXiv:2009.11937*.