Objective Surgical Skill Assessment of Computer-Aided Hysterectomy Procedures

Elif Bilgin[1]

[1]Department of Computer Science, Johns Hopkins University

Abstract

The aim of this project is to automatically assess skill in robot assisted hysterectomy procedures, particularly in the colpotomy step using machine learning (ML) and deep learning (DL) methods. In this paper, we present the development process of this project, as well as a detailed discussion of results collected and what is next in this field.

Keywords — classification; hysterectomy; robot-assisted surgery; machine learning; skill assessment

1 Introduction

Hysterectomy is the process of the removal of the uterus. Colpotomy is a step in the hysterectomy procedure where the connective tissue attaching the uterus to the vaginal opening is removed to release the uterus before removal.

The aim of this project is to automatically assess skill in robot assisted hysterectomy procedures. With the increase in use of surgical robotic systems, there is more opportunity for capture and analysis of complex surgical data. This allows for the objective computer-aided technical skill evaluation for scalable, accurate assessment; individualized feedback, and automated coaching.

Currently, there does not exist a uniform benchmark of assessing surgical skill. Research has been done mainly on VR simulations using various features such as task completion time, path length, moving time, velocity, idle time and energy activation. There is only a handful of research papers published on using data directly from the operating room, which distinguishes this project.

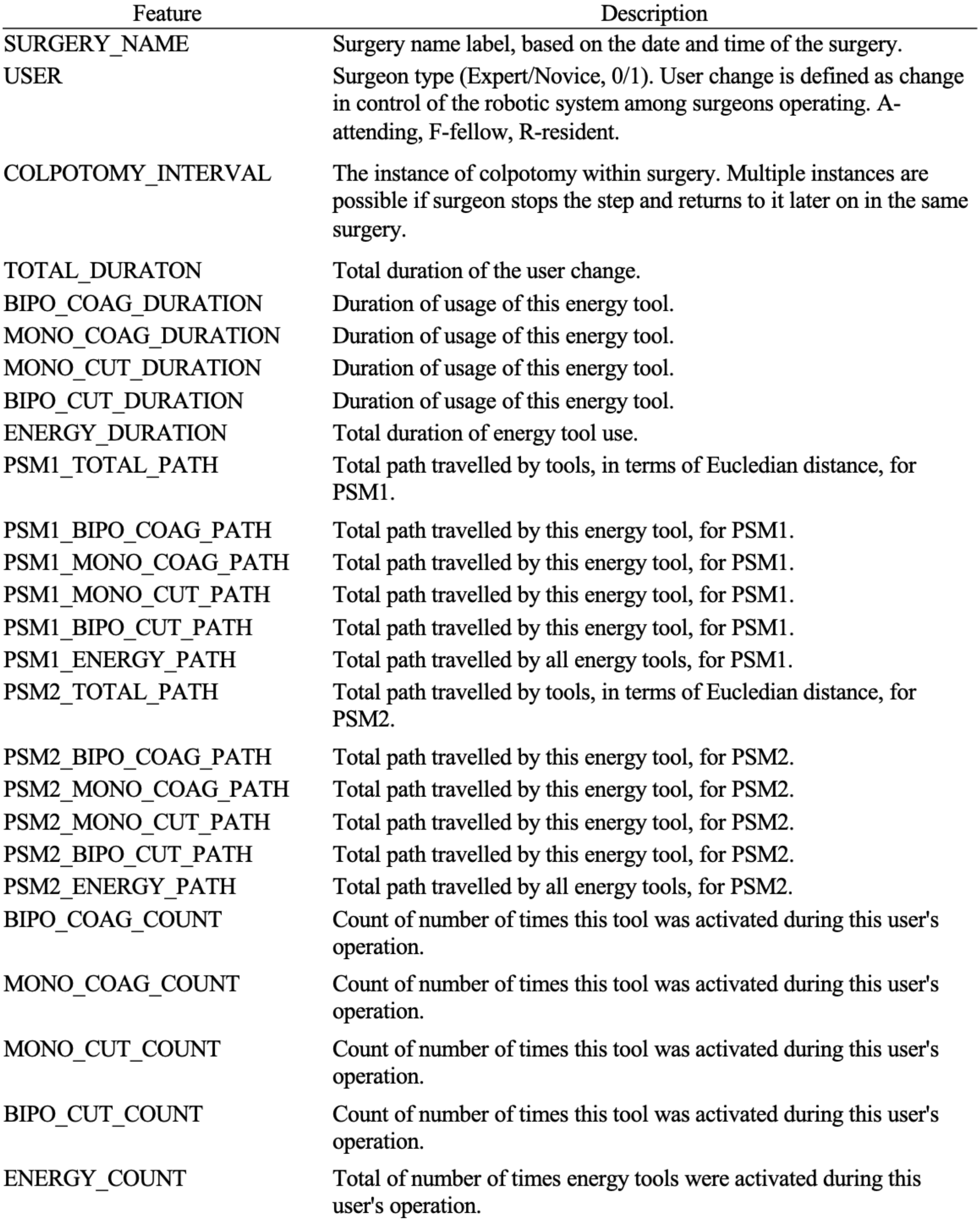
The ability to objectively assess the skill level of surgeons is critical for training future surgeons and revolutionizing this process.

2 Problem and Approach

For this project, we decided to build a dataset to be used for a feature-based classification problem of Expert or Trainee. This dataset comprises of the surgery date/time, surgeon expertise level (0 or 1, for expert or trainee), for each interval of data that is a part of the colpotomy step in the hysterectomy procedure.

We then ran multiple classification algorithms on this dataset such as Random Forrest, Logistic Regression, K-Nearest Classifier. We also did a parameter sweep to improve accuracy using these models.

Table 1: List of features in the dataset and descriptions.



Feature	Description
SURGERY_NAME	Surgery name label, based on the date and time of the surgery.
USER	Surgeon type (Expert/Novice, 0/1). User change is defined as change in control of the robotic system among surgeons operating. A-attending, F-fellow, R-resident.
COLPOTOMY_INTERVAL	The instance of colpotomy within surgery. Multiple instances are possible if surgeon stops the step and returns to it later on in the same surgery.
TOTAL_DURATON	Total duration of the user change.
BIPO_COAG_DURATION	Duration of usage of this energy tool.
MONO_COAG_DURATION	Duration of usage of this energy tool.
MONO_CUT_DURATION	Duration of usage of this energy tool.
BIPO_CUT_DURATION	Duration of usage of this energy tool.
ENERGY_DURATION	Total duration of energy tool use.
PSM1_TOTAL_PATH	Total path travelled by tools, in terms of Eucledian distance, for PSM1.
PSM1_BIPO_COAG_PATH	Total path travelled by this energy tool, for PSM1.
PSM1_MONO_COAG_PATH	Total path travelled by this energy tool, for PSM1.
PSM1_MONO_CUT_PATH	Total path travelled by this energy tool, for PSM1.
PSM1_BIPO_CUT_PATH	Total path travelled by this energy tool, for PSM1.
PSM1_ENERGY_PATH	Total path travelled by all energy tools, for PSM1.
PSM2_TOTAL_PATH	Total path travelled by tools, in terms of Eucledian distance, for PSM2.
PSM2_BIPO_COAG_PATH	Total path travelled by this energy tool, for PSM2.
PSM2_MONO_COAG_PATH	Total path travelled by this energy tool, for PSM2.
PSM2_MONO_CUT_PATH	Total path travelled by this energy tool, for PSM2.
PSM2_BIPO_CUT_PATH	Total path travelled by this energy tool, for PSM2.
PSM2_ENERGY_PATH	Total path travelled by all energy tools, for PSM2.
BIPO_COAG_COUNT	Count of number of times this tool was activated during this user's operation.
MONO_COAG_COUNT	Count of number of times this tool was activated during this user's operation.
MONO_CUT_COUNT	Count of number of times this tool was activated during this user's operation.
BIPO_CUT_COUNT	Count of number of times this tool was activated during this user's operation.
ENERGY_COUNT	Total of number of times energy tools were activated during this user's operation.

3 Dataset

The dataset was built by following these steps:

1. Go through each surgery file to find the exact intervals of colpotomy (there are multiple).
2. Locate these intervals in user change files, and extract user change files for only the colpotomy step. Do the same for energy usage files.
3. Using these user change files, extract colpotomy motion data (per PSM) and energy usage data (per instrument), for each user change, in each surgery, during colpotomy.
4. Calculate duration, count and path length for each feature.

The dataset has a total of 18 features (6 of which are repeated for PSM 1 and PSM2), collected on a total of 29 surgeries. Each of these can be seen in Table 1. Due to the user changes as well as multiple instances of the colpotomy procedure in a single surgery (the surgeon pausing at this step to do another procedure, and returning to the colpotomy), we ended up with 49 data points. Expert and Novice distribution in the dataset was near 50%, which was considered a fair distribution that avoids any bias towards either skill level, within the dataset.

4 Results

In this section, we present the various types of results we got along the progress of this project. These results will be referred back to in Section 5, Discussion and Conclusions. Tables 1 and 2 show the accuracy of the 5 feature-based models before and after the parameter sweep. Figure 1 is a plot of the data of Table 2. Table 3 is a breakdown of accuracies of these models at using different number of folds for cross validation.

Table 2: Model test accuracy values before parameter sweep.

Model	% Acc
Logistic Regression	39.13
K-Nearest Neighbor	50.00
Decision Tree	60.87
Adaboost	56.52
Linear SVM	47.83

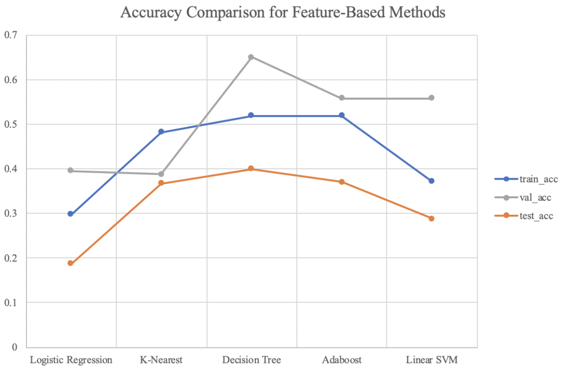
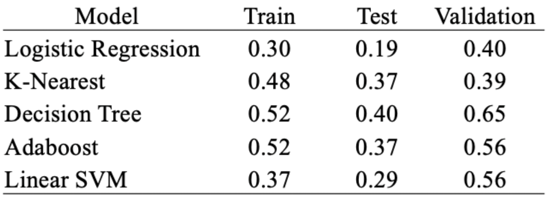


Figure 1: Average Train/Test/Validation Accuracy different feature-based models.

Table 3: Model performance after parameter sweep, for 8-fold cross validation.

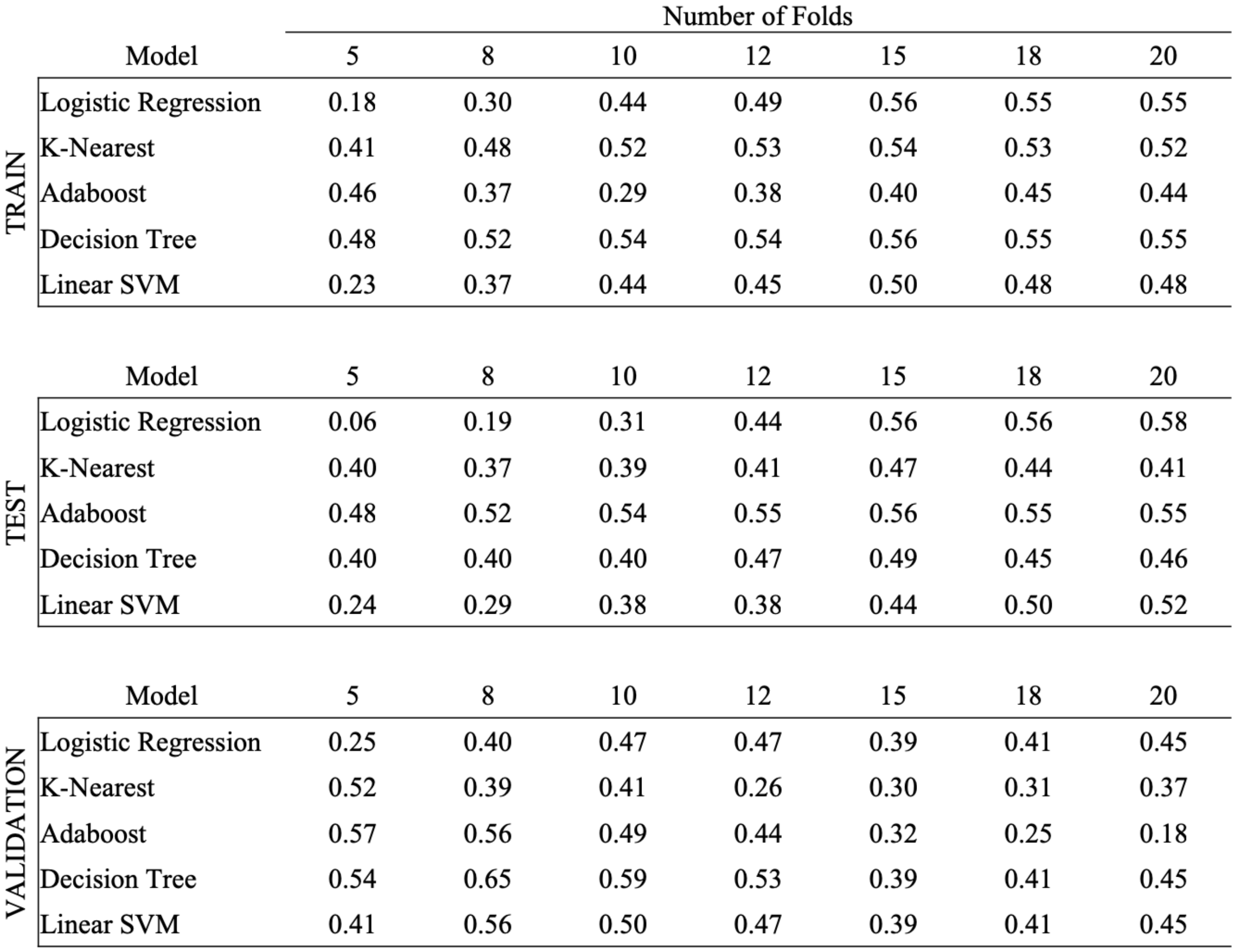


Model	Train	Test	Validation
Logistic Regression	0.30	0.19	0.40
K-Nearest	0.48	0.37	0.39
Decision Tree	0.52	0.40	0.65
Adaboost	0.52	0.37	0.56
Linear SVM	0.37	0.29	0.56

5 Discussion and Conclusions

There is a wide range of results that we have accumulated during the process of this project. First and foremost, the initial results we collected were the statistical information we gathered on our data. These were determined by the features we selected from the background research done prior to the projects. We collected statistics such as

Table 4: Average Train, Test and Validation Accuracies for different folds, for each feature-based model tested.



		Number of Folds						
	Model	5	8	10	12	15	18	20
TRAIN	Logistic Regression	0.18	0.30	0.44	0.49	0.56	0.55	0.55
	K-Nearest	0.41	0.48	0.52	0.53	0.54	0.53	0.52
	Adaboost	0.46	0.37	0.29	0.38	0.40	0.45	0.44
	Decision Tree	0.48	0.52	0.54	0.54	0.56	0.55	0.55
	Linear SVM	0.23	0.37	0.44	0.45	0.50	0.48	0.48
	Model	5	8	10	12	15	18	20
TEST	Logistic Regression	0.06	0.19	0.31	0.44	0.56	0.56	0.58
	K-Nearest	0.40	0.37	0.39	0.41	0.47	0.44	0.41
	Adaboost	0.48	0.52	0.54	0.55	0.56	0.55	0.55
	Decision Tree	0.40	0.40	0.40	0.47	0.49	0.45	0.46
	Linear SVM	0.24	0.29	0.38	0.38	0.44	0.50	0.52
	Model	5	8	10	12	15	18	20
VALIDATION	Logistic Regression	0.25	0.40	0.47	0.47	0.39	0.41	0.45
	K-Nearest	0.52	0.39	0.41	0.26	0.30	0.31	0.37
	Adaboost	0.57	0.56	0.49	0.44	0.32	0.25	0.18
	Decision Tree	0.54	0.65	0.59	0.53	0.39	0.41	0.45
	Linear SVM	0.41	0.56	0.50	0.47	0.39	0.41	0.45

average colpotomy duration, which varied from a min of 183.82 seconds to a max of 1028.17 seconds. The average values for each feature we selected for expert and novice categories are presented in Table 1.

It was noted that in the case of expert surgeons, the average total path travelled, energy usage duration and energy path travelled were longer compared to novice surgeons. This can be perhaps explained by the fact that the expert surgeons take their time to make sure that the surgery is done as precisely as possible while the novice surgeons may be more impulsive in their process, although this discussion is beyond the scope of this paper.

Motivated by these statistics, we were able to build the dataset. The second set of results we have are the initial model test accuracy values for the 5 different feature-based classifier models we ran on the dataset: Logistic Regression, K-Nearest Neighbor, Decision Tree, Adaboost and Linear SVM. The results of this initial classification using a 10-fold cross validator did not yield desirable values, as the highest accuracy achieved was about 60%. Taking these values into consideration, we chose to explore two areas: 1) How is the number of folds affecting the accuracy of our models? 2) Can we do a parameter sweep to improve the accuracy of the models? These explorations yielded the last results we acquired.

Table 3 displays the different training, testing and validation accuracies that we got for each model with different fold sizes. We decided to use 8 folds, as the model that performed the best among the 5, decision trees, performed the best at this fold count. The values in Table 2 have been calculated using 8 folds.

Table 2 presents the training, testing and validation accuracies that we got for these 5 models after the parameter sweep. The validation accuracy of the decision tree reached around 65%. In Figure 1, these values are shown in the form of a plot. As expected, the training and validation accuracies are higher than the testing accuracy, which means the models are working correctly. However, these accuracy values are quite low. This can be explained by the fact that the dataset is sparse and very complex for a machine learning problem. One way to remedy this is to use more features that we had not opted to use for this project in the future, such as velocity and idle time. However, our expectation is that this won't change the fact that our dataset is possibly too complex for a feature-based method to work well with.

For the scope of this project, it was assumed that the expert surgeons were experts at using the robotic system, while the novice surgeons are in training and are not as capable. However, this is a big assumption to be made - perhaps the newer generation of doctors may have more experience in training with the robot and possibly may have better skills than the Attending or Fellow they are operating alongside.

This has been a point of discussion time and time again in the papers we have reviewed. It is a crucial problem that there does not exist a ground truth in skill assessment. Most papers we reviewed had various different versions and definitions of skill, and this causes discrepancies among different research projects. This causes research to fail to build on one another. Therefore, a consensus being established on what the best approaches and benchmarks are necessary.

6 Management Summary

Since this was a single student group, all work was completed by Elif Bilgin, under the mentorship of Anand Malpani. The deliverables that were established were the following:

Deliverable	Description
Min	Statistics on video, kinematics and motion data, create dataset.
Expected	Apply feature-based ML methods to get an Expert/ Novice classification.
Max	Apply Deep Learning methods to time series data such as DNNs and RNNs to get a Expert/Novice classification at an accuracy level of 80%.

Table 5: Deliverables planned.

The minimum and expected deliverables were met as planned. The maximum deliverable was not met due to difficulties getting neural network code provided to work with the data format we had, as well as dependencies. This is quite unfortunate, as we expected the deep learning methods to have a higher level of accuracy than the machine learning algorithms.

7 Future Steps

The next step in this project timeline would be to work more on getting the neural net code to work with the data we have, and perhaps add more data in the meantime. We expect that the neural network will be able to produce better accuracy values in determining if the surgeon is a Novice or Expert. Molly's code is a convolutional neural network that is made up of many layers including multiple convolution layers, linearization and normalization layers. This code has been used in septoplasty OR skill assessment procedures before, and is expected to perform desirable results. However, Molly's dataset was significantly larger than the dataset we have currently, which means more data is crucial in this neural net to perform well, as the performance of NNs is directly correlated to the amount of data that is fed in.

For the next steps of this project, we also believe that it is important to collect skill scores from surgeons based on video footage corresponding to the colpotomy step motion data. We hope to then use the feature-based machine learning algorithm results and the neural network results along with the scoring from surgeons to regress a skill score for each user. Another goal for the future in the scope of this project is to collect more data from the OR, for a better predictive model accuracy.

In a more general scope, it is important to note that for advancements to occur in this area, a consensus being established on what the best approaches and benchmarks are necessary. Currently, there does not exist a ground truth in skill assessment, which means each project done in this field has a different notion of ”skill.” For research to build on each other, they must be comparable. Hence, this is essential for advancement in the field of automated surgical skill assessment.

8 Acknowledgments

Thank you to Anand Malpani, PhD for his mentorship.

Thank you to Saranga Arora, JHU BME and Chi Chiung Grace Chen, MD MHS, JHU Gynecology Obstetrics for the data collection.

Thank you to Molly OBrien for providing

the code for the neural network.

References

[1] Malpani, A Martinez, N Vedula, S Hager, G Chen, C. (2018). Automated skill classification using time and motion efficiency metrics in vaginal cuff closure. American Journal of Obstetrics and Gynecology. 218. S891-S892.

[2] Azari, David P. et al. Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. Annals of surgery (2017)

[3] Ershad, Marzieh Rege, R Majewicz Fey, A. (2018). Meaningful Assessment of Robotic Surgical Style using the Wisdom of Crowds. International Journal of Computer Assisted Radiology and Surgery. 13. 10.1007/s11548-018-1738-2.

[4] Swaroop Vedula, S O Malpani, Anand Tao, Lingling Chen, George Gao, Yixin Poddar, Piyush Ahmidi, Narges Paxton, Christopher Vidal, Ren Khudanpur, Sanjeev Hager, Gregory Chiung Grace Chen, Chi. (2016). Analysis of the Structure of Surgical Activity for a Suturing and Knot-Tying Task.

[5] Soto, Enrique Lo, Yungtai Friedman, Kathryn Soto, Carlos Nezhat, Farr Chuang, Linus Gretz, Herbert. (2011). Total laparoscopic hysterectomy versus da Vinci robotic hysterectomy: Is using the robot beneficial?. Journal of gynecologic oncology. 22. 253- 9. 10.3802/jgo.2011.22.4.253.

[6] Poddar, Piyush Ahmidi, Narges Swaroop Vedula, S Ishii, Lisa Hager, Gregory Ishii, Masaru. (2014). Automated Objective Surgical Skill Assessment in the Operating Room Using Unstructured Tool Motion. International journal of computer assisted radiology and surgery.

[7] Zhang, Yetong Law, Hei Kim, Tae-Kyung Miller, David Montie, James Deng, Jia Ghani, Khurshid the Michigan Urological Surgery Improvement Collaborative, for. (2018). Surgeon Technical Skill Assessment Using Computer Vision-Based Analysis. The Journal of Urology.

[8] Malpani, Anand Swaroop Vedula, S Chiung Grace Chen, Chi Hager, Gregory. (2015). A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. International journal of computer assisted radiology and surgery.

[9] Zia, Aneeq Essa, Irfan. (2017). Automated Surgical Skill Assessment in RMIS Training. International Journal of Computer Assisted Radiology and Surgery. 13.

[10] Vedula, Satyanarayana S et al. Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. Annual review of biomedical engineering 19 (2017): 301-325 .

[11] Krishnan, Sanjay Garg, Animesh Patil, Sachin Lea, Colin Hager, Gregory Abbeel, Pieter Goldberg, Kenneth. (2017). Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning. The International Journal of Robotics Research.