Surgical Metrics: Technique classification during simulated surgical bowel repair

1 Introduction

The collision of computer vision and object detection within the surgical field has brought about an explosion of ontological research to understand which techniques and decisions are correlated with surgical success and improved patient outcomes. Can an expert surgeon be differentiated from a novice resident just by noticing how and when certain tools are used? How do the tool choices reflect their technique, and how do these differences correlate to outcomes?

Using annotations of surgical tool useage during a simulated surgical bowel repair, I will aim to differentiate between two different suturing techniques which are known to correlate to a difference in outcome. During bowel repair, the surgeons may use one of two suturing techniques: a running suturing style where the surgeon places many interlaced sutures before cutting the suture threads, or interrupted suturing style where a surgeon cuts the suturing thread after each suture placement.

2 Dataset:

2.1 Dataset Collection

I work with the Technology Enabled Clinical Innovation Lab at Stanford who collected the data in 2019 at the American College of Surgeons Conference. They simulated a surgical setting for over 200 participants (i.e. medical students, surgical residents, practitioners, and retirees) where they simulated a surgical setting of a suture repair for two injuries (a small and large hole) in a porcine intestine. During the procedure, the surgeons wore motion tracking sensors on their hands, and two POV cameras recorded the surgical field. The simulation setup is shown below in Figure 1. After the procedure is complete, the intestine is pumped full of water to a given pressure to determine if the procedure was successful (no leak) or unsuccessful (leakage of water.

Post data collection, my lab identified key decisions that the surgeons made when repairing the bowel. The decision I focused on was the difference in suture technique, interrupted or running which are visualized in figure 2A. An interrupted suturing technique means that the surgeons will insert a suture and immediately tie it off and cut the suture thread. The running technique employs a more interconnected, interlaced approach where the suregon will place several sutures before tying it off and cutting the suture. As shown figure 2B, the lower leak rate for those who chose the running suture technique was shown to be statistically significant with respect to outcome. The intention with this project was to see if we could identify which of the techniques the surgeons were employing.

2.2 Dataset Annotation

There are four tools used during the procedure: hemostat (a clamping tool), forceps (large tweezers), scissors, and needle-driver (clamp to hold the suturing needle.) We defined 5 possible states the tool could be in: 1) actively being used by the surgeon 2) Held by the surgeon but not in a position that it's usable (e.g. the surgeon is holding the scissors but currently tying a knot instead) 3) actively being used by the assistant 4) Held by the assistant but not in a position that it's usable 5) Lying inactive in

CS230: Deep Learning, Winter 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

Figure 1: Set-up of the simulated surgical repair of the bowel. Tools are held in the blue tray to the left of the field of view. The leak test to determine the outcome is currently being conducted to check if the suture is robust to water at a given pressure.

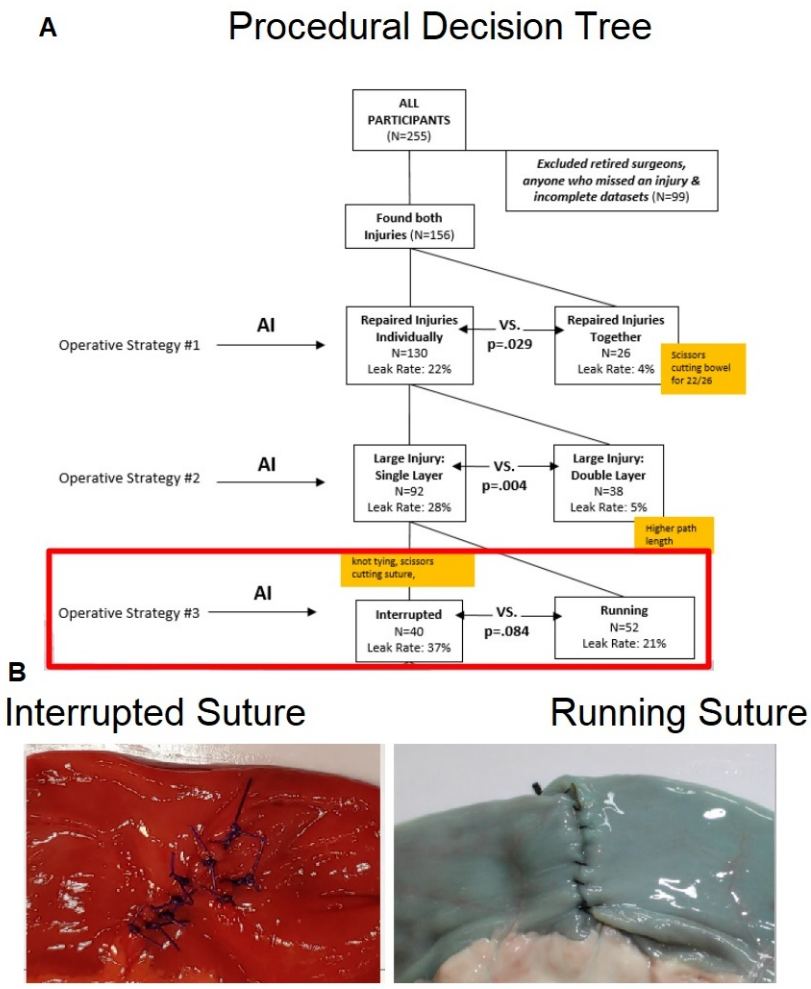


Figure 2: A) Decision Tree. These operative decision were identified post data collection and identified as being correlated to success of the procedure as determined by the leak test. B) Visualization of the difference in suturing technique

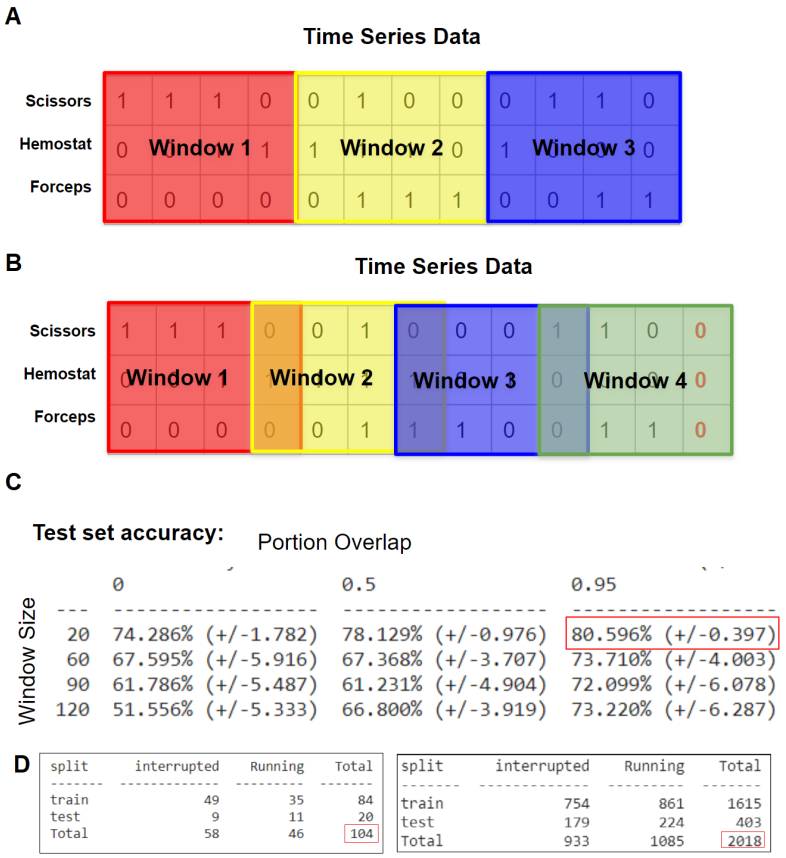


Figure 3: A) Visualization of the windowing technique that demonstrates the curation of 3 samples from a single time series video. B) Demonstration of the overlap technique that generates 4 samples with this overlap of 25%. In the case that the last window is not long enough it is padded with zeros. C) The psuedo hyperparameter sweep to show the optimal pairing determined by test accuracy. D) THe increase of number of samples from 20 second slicing with 95% overlap.

the surgical field of view. Thus, each of the four tools has 5 states (surgeon active/inactive, assistant active/inactive, inactive in surgical field), yielding 20 features.

Post data collection, my lab and I annotated these videos to denote when and which tools the surgeons use throughout their procedures. suture technique used by the surgeon. In some cases, the surgeon decided to switch techniques between repairing the small and large injury holes. Therefore, each participant has two labels of how they repaired the large and small enterotomy.

2.3 Dataset Augmentation:

One of the biggest barriers to learning on this data is the small size of the dataset. Thus, I tried to two related techniques to augment the number of videos we had. We only had 52 participants, each who repaired 2 injuries in the bowel, yeidling 104 videos (Figure 3d left). However, knowing that the suturing technique can be identified from about 20-30 seconds of a procedure video, I decided to experiment with splitting each of the suture repair time-series into smaller portions (Figure 3A). The shortest repair was 20 seconds long, so chose that as the lower edge case and 2 minutes the upper edge case. In addition, I experimented with overlapping these windows splices of the data (similar to adjusting stride figure 3B) to further generate samples. I ran a psuedo hyper-parameter sweep of window size on 20, 60, 90, 120 second slices against overlap percentages from none (0% overlap) to a stride of 1 on the 20 second window (95% overlap). The results of this experiment are shown in figure 3C. Result: A window size of 20 seconds and overlap of 95% showed best testing accuracy yielding 2018 samples (figure 3D right).

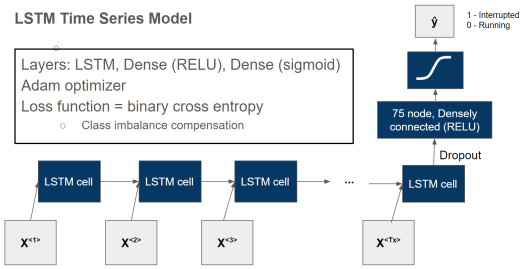


Figure 4: Block diagram of the LSTM model, dense RELU activation to sigmoid activation giving prediction of suturing technique.

Testing Accuracy:

Hidden State

Learning Rate	25	50	75	100
0.0001	79.454% (+/-1.168)	80.546% (+/-1.024)	79.156% (+/-2.932)	80.199% (+/-1.517)
0.001	81.191% (+/-1.046)	80.248% (+/-1.477)	80.050% (+/-1.582)	80.496% (+/-0.639)
0.01	80.199% (+/-2.377)	82.581% (+/-1.103)	81.985% (+/-1.024)	81.737% (+/-0.923)
0.1	58.362% (+/-7.626)	64.913% (+/-3.887)	72.903% (+/-6.911)	57.370% (+/-16.026)

Testing Accuracy:

Decay

Learning Rate	0.0001	0.001	0.01
0.0001	82.382% (+/-0.000)	79.653% (+/-0.000)	65.012% (+/-0.000)
0.001	79.156% (+/-0.000)	81.390% (+/-0.000)	81.390% (+/-0.000)
0.01	80.149% (+/-0.000)	81.638% (+/-0.000)	78.412% (+/-0.000)

Figure 5: Hyperparameter sweeps between Hidden State Size and Learning Rate in addition to exploration with learning rate decay. Values boxed in red were chosen

3 Approach

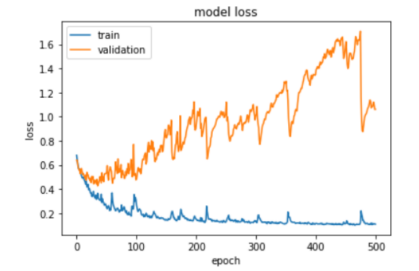
3.1 Model Choices:

Given the time-series nature of the data, I decided I would use an LSTM architecture to capitalize on the temporal patterns that distinguish the two classes. I started with the example code from Jason Browlee's "LSTMs for Human Activity Recognition Time Series Classification." [1]

The LSTM model code base uses a single LSTM hidden layer followed by a fully-connected layer (RELU activation), fully connected to a signle sigmoid activation node which is used to make predictions (1 - interrupted, 0 - running). Diagram of the model is shown in Figure 4. I use the categorical cross entropy loss function, Adam optimizer to speed up training, and accuracy as the evaluation metric as the example suggests. In addition, to compensate for the slight imbalance between classes, create a compensating multiplier for the loss function to penalize proportional to the reciprocal of the class's representation in the data set.

3.2 Model Hyper-parameters:

The two hyper-parameters I decided to tune for learning were the hidden state size and learning rate. I tried a range of [25, 50, 75, 100] and [0.1, 0.01, 0.001, 0.0001]. Results shown in figure 5A. Since the learning rate of 0.01 was best but seemed large, I implemented learning rate decay in hopes that it would help the model converge to the optimal weights. I did a sweep between the same learning rates and the same magnitudes of learning rate decay to come up with the optimal pairing of 0.01 learning rate and 0.0001 decay.



Test Accuracy : 82.630%

	precision	recall	f1-score
Running	0.84	0.85	0.84
Interrupted	0.81	0.80	0.80

Figure 6: Plot showing the loss from the training of the model and resulting test accuracy. Precision and Recall are also reported with respect to each class.

3.3 Regularization:

In addition, I experimented with implementing dropout between my final LSTM hidden state and the densely connected RELU activation layer. I tried dropout rates [0.5, 0.6, 0.7, 0.8, 0.9] with 0.8 yielding highest accuracy on the test set.

4 Results and discussion:

Results of final tuned model are reported in Figure 6. While I was able to achieve reasonable precision of 82.64%, the validation loss remained high. When examining the errors in prediction, I noticed that over 58% (43/73) incorrect classifications in the final test set were extremely confident guesses from the model (>0.999, <1e-6). This leads me to believe that these samples are either mislabeled or outliers. This would explain the high validation loss, as a very confident incorrect classification would get highly penalized. Thus, there may be merit in investigating the correctness of the labels since these errors account for such a high percentage of the classifications.

5 Future Work:

As we work with our collaborators, we each are exploring different pathways to bring meaning and understanding to the surgical space through interpretation of video and motion data. With clear evidence that technique detection is possible given just tool annotations, this motivates work to use object detection on surgical videos to then feed into higher-level models (such as this) to yield an cascade, inputting raw surgical footage and yielding identification of techniques. As these pieces of the puzzle are put together, each of these models can begin to bring understanding of the ontologies of surgical decision making and techniques.

6 Acknowledgements:

Stanford TECI Center lab:Each of these members helped with data collection and instrument data annotations:Anna Witt, Brett Wise, Su Yang, Hossein mohamadipanah

Johns Hopkins Computational Interactions and Robotics Laboratory Team Each of these team members helped with video annotation: Brett Wolfinger. Mike Peven. Mike also helped in advising the troubleshooting and training of the model