Machine Learning as an Effective New Tool for Assessing Human Performance During Simulation-based Training

Roger Smith, Modelbenders

Danielle Julian, AdventHealth

Orlando, FL

rdsmith@Modelbenders.com, Danielle.Julian@AdventHealth.com

ABSTRACT

Objective: To apply machine learning (ML) models to analyze and score videos of simulation-based surgical training. ML has the potential to replace the burdensome task of enlisting experienced surgeons to score large numbers of videos from training and research events. This is equally applicable to videos of military training.

Methods: Researchers collected 254 videos of two different simulation-based exercises. The quality of performance in each video was scored both by a simulator and by an experienced instructor, creating two different objective labels for ML models. Both numeric scores were converted into the class labels – expert, intermediate, and novice - consistent with accepted surgical evaluation practices. The videos were cut into 10 second clips for analysis by ML algorithms. 2,227 video clips were processed using the Google Cloud Platform AutoML service. 80% of the clips were assigned to the training set, 20% to the test set. Multiple ML models were created with different combinations of the videos.

Results: Datasets which included videos of the Ring & Rail exercise achieved accuracy on simulator scores of 85.9/77.3/77.3% (average/precision/recall) while the Suture Sponge exercise achieved 81.0/76.3/70.8%. For the instructor-assigned scores the model's performance was 83.1/76.1/67.7% and 80.8/72.8/67.7% respectively. A model combining all 2,227 videos from both exercises was able to achieve: 78.3/72.7/65.2% for simulator generated scores; and 75.3/72.9/59.3% for instructor generated scores.

Conclusions: ML models for individual exercises deliver very good results (80+% accuracy) in matching the scores assigned by both simulators and human instructors. The accuracy of these models is impacted by the number of samples available for training, the balance between the classes, and the clarity of differentiable skills in the video clips. Combining the videos into a single dataset results in a unified model that can be used for multiple exercises. The accuracy of this model declined (78% simulator, 75% instructor) because of the heterogeneity of the exercises, but was aided by the larger training dataset. Collecting larger datasets should improve the accuracy of single and combined exercise models. ML models created from services like Google's AutoML can potentially relieve humans from the burden of evaluating videos of training and research projects. The methods used here can be applied to scoring videos of military training events.

ABOUT THE AUTHORS

Roger Smith, Ph.D., has spent 25 years creating leading-edge simulators for the Department of Defense and Intelligence Community. He has served as the Chief Technology Officer for the AdventHealth Nicholson Center; the CTO for the U.S. Army PEO-STRI; VP and CTO for training systems at Titan Corp; and VP of Technology at BTG Inc. He holds a Ph.D. in Computer Science, a Doctorate in Management, an M.S. in Statistics, and a B.S. in Applied Mathematics. He has published 3 professional textbooks on simulation, 17 book chapters, and over 100 journal and conference papers. His most recent book is *Thinking About Innovation*. He has served on the editorial boards of *Transactions on Modeling and Computer Simulation* and *Research Technology Management* journals.

Danielle Julian, M.S., is the Director of Education at AdventHealth Nicholson Center. She manages and directs the creation and delivery of multiple surgical training courses and events. Her research focuses on robotic surgery simulation and effective surgeon training using intelligent tutoring systems, rapid prototyping of surgical education devices, and the evaluation of simulation systems. Her background includes research in Human Factors and learning and training to enhance the higher-order cognitive skills of military personnel. She is currently a Ph.D. student in Modeling and Simulation at the University of Central Florida where she previously earned an M.S. in Modeling and Simulation, Graduate Simulation Certificate in Instructional Design, and a B.S. in Psychology.

Machine Learning as an Effective New Tool for Assessing Human Performance During Simulation-based Training

Roger Smith, Modelbenders

Danielle Julian, AdventHealth

Orlando, FL

rdsmith@Modelbenders.com, Danielle.Julian@AdventHealth.com

BACKGROUND

Surgeons are constantly learning new techniques and the use of new tools by attending short courses at specialized facilities around the world. Their individual skill levels are traditionally assessed by an experienced mentor or instructor who observes a set of defined exercises and assigns a score to the observed performance. In recent decades, structured assessment tools have been created and validated for the major categories of surgery – open, minimally invasive, and robotic-assisted minimally invasive (Martin et al, 1997; van Hove et al, 2010; Goh et al 2012). These tools are generally subjective Likert scale metrics for observed behaviors while the students performed a difficult exercise. These traditional methods create a recurring burden on mentors and instructors to provide sufficient time, attention, and objective standardized reasoning to assign a reliable score to the demonstrated skill. Recently, computer-based surgical simulators have been created with embedded scoring metrics. Simulators are able to provide consistent and objective scores to every user without requiring the attention of a human mentor or instructor. However, there has been a persistent question about whether the metrics collected by a computer program are actually assessing the most important skills of the subjects they are evaluating (Lui & Curet, 2015).

Given the limitations of the methods used for assessment of surgical skills, we propose that modern, advanced machine learning methods, specifically deep learning neural networks, could be trained to provide a score that captures the consistency of computer simulation algorithms and the wholistic evaluation of a human evaluator.

Deep learning neural networks require highly advanced computer programming skills to create from scratch. The algorithms also require significant computing resources to build and configure. Both of these have limited the application of the algorithms in industries that do not possess these specialized resources. Recently, organizations like Google, Microsoft, Amazon, and IBM have made their most advanced deep learning algorithms and their computing infrastructures available as cloud-based services. These services open the door for industries like healthcare and defense to apply these advanced tools to their domain-specific problems.

OBJECTIVE

In an effort to relieve the burden of scoring surgical performance of residents, fellows, and students in short courses or research projects, we sought to leverage available cloud services for computing, storage, and machine learning algorithms to perform the task. Our goal was to create a machine learning model that could assign the same scores that would be expected from a human mentor or instructor. Models were also created which could match the scoring of a simulator device, which would contribute to the validation of the methodology.

Surgical training exercises are typically performed in three different environments (Figure 1). Wet-lab exercises are conducted with cadavers, animals, or excised biological tissue. Dry-lab exercises are performed with plastic and silicon models that recreate the appearance of human tissue, or are designed as “skills tasks” specifically to measure a skill like knot-tying or suturing. Simulator exercises are performed with a computer system and may have the appearance of a wet-lab or a dry-lab. These devices usually include automated metrics of performance.

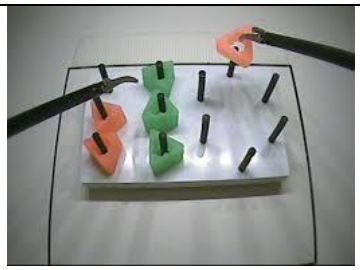
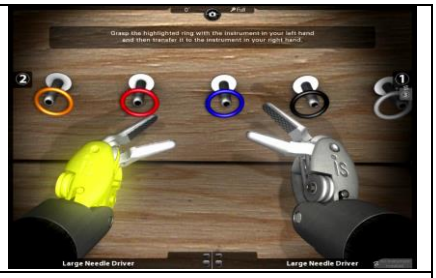
		
Wet lab evaluated by expert observation.	Dry lab evaluated by expert observation.	Simulated lab evaluated by computer algorithm.

Figure 1. Surgical exercise modalities and assessment methods.

For this research we used a library of videos of simulator exercises that had been collected in a series of robotic surgery training courses offered over a period of a year. These videos had already been labeled with multiple simulator-assigned scores. They had also been viewed by human instructors and assigned Likert scores of performances using the GEARS tool (Goh et al, 2012). The video images from a simulator are very consistent every time the exercise is attempted and they present simpler and more consistent images than either wet-labs or dry-labs. The objects are always identical, as are the textures, dimensions, positions, and lighting. These constraints make simulation-based video an ideal environment in which to test an ML models capability. However, the real value of a working ML model will be in scoring wet-lab or dry-lab exercises that currently require significant time from human evaluators. Creating a successful ML model for simulator video is a first step toward those more visually challenging tasks.

This research was conducted in two distinct steps. Phase 1 applied the computer and AI services to still images taken from the videos (Smith, Julian, 2019). The models were trained to identify objects that appear in the images. This allowed our team to develop proficiency with the cloud-based tools. Phase 2, reported in this paper, applied ML tools to scoring the quality of performance that is demonstrated in video.

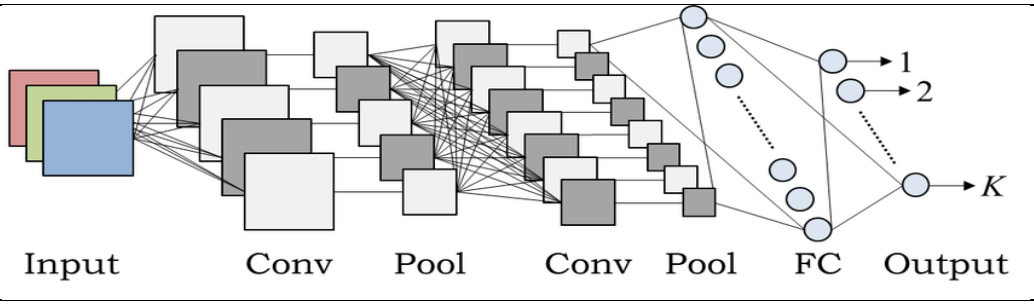
MATERIALS AND METHODS

Deep neural networks (DNNs) use an extensive, directed graph of nodes, links, weights, mathematical transformations, and decision gates (Figure 2a) to create a model that can identify patterns in data of any form (LeCun, 2015). This pattern recognition ability is learned or trained by supplying the network with hundreds or thousands of data items for which the desired or "correct" answer is already known. The network uses these cases to make brute force attempts to adjust its internal settings to generate the matching "correct" answer as its output. The network "learns" by recognizing the degree to which its initial answers missed the correct answer. It then back-propagates that error though its settings, raising and lowering thousands or millions of internal values until the network produces the "correct" answer for a high percentage of cases. The goal is to achieve a high percentage of correct answers, but never to reach 100% accuracy, for reasons that are beyond the scope of this paper.

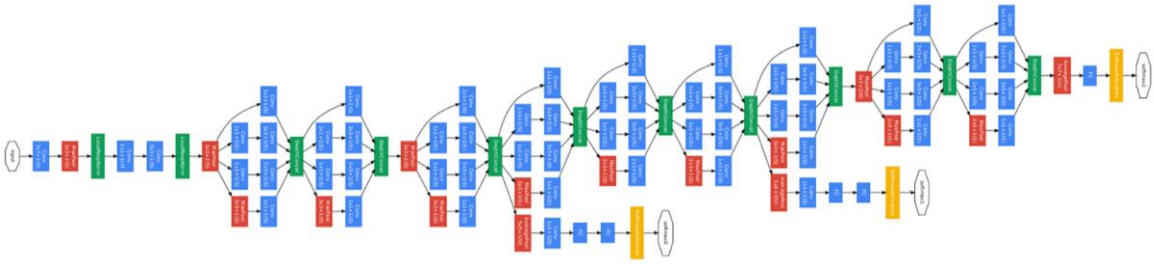
The mathematics and network topologies that are used for DNNs are very extensive and explanations can be found in thousands of published resources on the topic. Interested readers can begin with Lecun's original paper in *Nature* (2015). Figures 2(a) and (b) offer a simplified graphical representation of the architectural structure of DNNs. The first is typically used in educational materials to convey the principles by which these networks work. But, complex problems like video evaluation require many more layers, such as those shown in the second diagram.

The videos that served as the training set for this experiment consisted of 254 sessions which included two different simulated exercises on simulators of the da Vinci robot – the Ring & Rail (RR) exercise and the Suture Sponge (SS) exercise illustrated in Figure 3. These videos were collected during multiple surgical training courses over a period of one year. The performance scores for every video were assigned using both the simulator scores and the human assessed GEARS scores as part of previous experiments (Dubin et al, 2017 & 2018).

Both the simulator scores and the human assigned GEARS scores follow a similar pattern in which specific metrics are gathered for multiple actions being performed. These metrics are then combined into a single "overall score" using a weighted sum (Table 1). This experiment used the overall score from each method as the objective, or correct answer, for the DNN model to match. The training toward the overall score was conducted independently for each scoring method – simulator and human-scored.



(a). Traditional architecture diagram of a deep learning neural network - 7 layers.



(b). Architectural diagram of a neural network sufficient for video processing – 95 layers.

Figure 2. Deep learning convolutional neural network architectures

Ring & Rail. Objective is to thread the ring along the bent rail to a finishing point at the end. Action is performed a single time.

Suture Sponge. Objective is to insert a needle and extract it through the specified points. Action is performed multiple times.

Figure 3. Images from the Ring & Rail and Suture Sponge exercises.

Table 1. da Vinci robotic surgical performance metrics from GEARS and the Simulator

GEARS Metrics (range)	Simulator Metrics (units)
Depth Perception (1 – 5)	Time to Complete (seconds)
Bimanual Dexterity (1 – 5)	Excessive Instrument Force (seconds)
Efficiency (1 – 5)	Master Workspace Range (centimeters)
Force Sensitivity (1 – 5)	Instruments Out of View (centimeters)
Robotic Control (1 – 5)	Instrument Collision (count)
GEARS Overall Score (5 – 25) (weighted sum of above scores)	Overall Score (0 - 100) (weighted sum of above scores)

Video Clipping

Current DNN algorithms do not have the ability to process long segments of video before classifying them. In human terms, we would say that the algorithms have a limited attention span or a limited ability to remember what they have seen beyond a few seconds. For most video classification problems, a segment of one to three seconds is sufficient to determine the action that is occurring (e.g. jumping, throwing, walking). A longer clip is required to assess the quality of the action exhibited (e.g. novice, intermediate, expert).

Human evaluators have a similar limitation, though their attention span is usually longer. For this experiment, all videos were cut into 10-second clips to accommodate the algorithms while retaining enough activity that a decision on quality of performance could be made. The Ring & Rail exercise videos average 30 seconds each and the Suture

Sponge exercises 150 seconds. Therefore, on average, the former videos became 3 clips and the latter became 15 clips. These clips formed a database of 2,227 video clips (RR=494, SS=1,733) that served as the training set for the DNN models. Since there are more Suture Sponge video clips, it should be possible to create a more accurate scoring model from those videos. But the Suture Sponge exercise is also more complex, which makes it more difficult to score reliably.

Data Transformation

Experts on the process of creating machine learning systems emphasize the amount of time that will be required to transform and standardize the data that is collected before any model development can begin. These tasks are very tedious, time consuming, and potentially error prone. But the quality and standardization of the data is essential if the resulting model is to be credible.

Cutting the videos into 10-second clips was one of the simpler forms of data transformation that was required. The file names for the videos had been previously assigned on the assumption that they would be read by human researchers. As a result, the names did not follow a consistent naming pattern that could be easily parsed by a computer. Neither did they use the file naming convention required by the DNN cloud services that were used for this project. Therefore, scripts had to be created to rename several thousand videos, converting multiple inconsistent naming conventions into a single standard. The videos were also collected at a very high resolution (1080p or higher) which is neither necessary nor beneficial to the DNN algorithms. Therefore, the videos were down sampled to 720p, making each video into significantly smaller file size. Note, that even 720p is much higher resolution than is required for DNN analysis.

The simulator and GEARS scores applied to the videos were stored in a spreadsheet with multiple internal worksheets. Most of those records were similarly intended for reading by humans who would mentally adjust for variations in naming conventions, the separation of data items, and collection from multiple simulator devices. Spreadsheet macros were created to transform the data into the formats required by the DNN cloud services. Table 2 provides a list of the key transformation steps that were used. This table illustrates the extensive, and often miniscule, adjustments that must be made to thousands of data records when preparing for any type of machine analysis.

Table 2. Key data transformation steps.

Video Files	Data Records
Cut into desired video clip lengths (5 steps) Resample to lower resolution (3 steps) Rename to standard convention (11 steps)	Rename Subject + Exercise to standard convention Merge cells to match video file name Append Google cloud directory location Normalize simulator overall score (0-100 range) Convert numeric Overall Score to Classes Insert video start and stop times Generate separate files for each exercise and metric
Num Steps =19	Num steps = 7

Discrete Thresholds

Both the GEARS and simulation scoring methods generate a continuous numerical "Overall Score" as their final assessment. But the DNNs that are used for this research are "classifying machines." They assign each video to a discrete class of performance, such as Novice, Intermediate, or Expert. These three classes are the most widely used and recognized in the surgical education and training community. Therefore, the numeric Overall Score for both methods was converted into one of these discrete classes. Validation studies of both the GEARS and simulator metrics have led to scoring thresholds that are consistent with measured performance by groups of surgeons known to be at each of these levels. These are often characterized as follows. Novices are medical students and residents. Intermediates are senior residents (e.g. 5th year) and fellows (those in advanced study programs). Experts are attending surgeons with privileges to perform surgery without supervision. For this research project, three different sets of thresholds were explored to define these class levels. First, we applied the thresholds that had been previously derived by validation studies. Second, we used scatter plots of the overall scores to visually identify clusters and to set the thresholds. Third, we used the machine learning technique K-means clustering to identify naturally occurring clusters of scores in the data set. K-means has no prior knowledge about where each skill level should reside on the scoring scale. It identifies the scores that clustered together and were separated from other clustered scores. The selection of threshold has a definite impact on the achievable accuracy of the resulting models. The results reported in this paper were achieved using the K-means derived thresholds shown in Table 3.

Table 3. "Overall Score" thresholds for classifying performance.

Metric	Novice	Intermediate	Expert
GEARS	0 - 16	17 - 19	20 – 25
Simulator	0 - 45	46 - 80	81 - 100

Cloud-based Resources

This project was only feasible because of the cloud-based services that have recently become available from the vendors listed earlier. Without high performance computer nodes and expertly generated DNN algorithms that can be purchased by-the-hour, this project would have required more financial and intellectual resources than are available in all but the most elite organizations. As reported in the 2019 paper, we experimented with cloud-based services from Google, IBM, Microsoft, and Amazon for the first phase of the project. We found those of Google and IBM to be the most effective for the first phase of the project. For the second phase reported here, we used only the Google services because those were the first to be available for video classification (Google, 2020; Hosseini et al, 2017).

For this work, we required four different cloud-based services. First, data storage in the cloud for 2,227 video files occupying 3.5 GB of disk space. Cloud storage is similar to online file systems like Google Drive, Microsoft One Drive, and Dropbox which people use for the safe, shareable storage of working documents. It is different in that the storage is closely tied to compute and networking resources that will be used to analyze the data. Second, high performance computer clusters. DNNs that are used for image and video recognition require a huge amount of computation to run in a reasonable amount of time. These computers must have high end graphics processors (GPUs) or tensor processors (TPUs) to handle millions of vector calculations. For our experiment, training one model required 13 computer nodes, each with 4 CPU's, 16GB RAM, and one Nvidia Tesla V100 card containing 5,120 GPU's and 640 TPU's. Third, network connectivity in the cloud and outward to customer computers. Efficient networks are needed to connect all of the computers and storage locations in the cloud, as well as to support web-based user interfaces and data delivery to a scientist's personal computer. Fourth, cloud-based DNN algorithms. These are world-class algorithms created by teams of leading scientists at the cloud services companies.

We also took advantage of two additional advanced capabilities that were only available from Google's video processing services at the time of the research – transfer learning and AutoML. The DNN model used for this experiment was not built from scratch by our research team, it was an existing Google-created model that had been pretrained by analyzing millions of videos online (e.g. YouTube videos) so it could recognize millions of different patterns, objects, movements, activities, and behaviors. Leveraging these pretrained models is known as "transfer learning" (illustrated in Figure 4). The initial DNN model contains approximately 100 layers (as shown in Figure 2b) and millions of parameters that were learned on other data sets (step 1 in Figure 4). Then our specific surgical simulation data added only a few layers on top of that much larger network (step 2 in Figure 4). This model can be trained much faster (i.e. hours) than the entire underlying model (i.e. weeks) (step 3 in Figure 4). Google also offers the AutoML service which trains and retrains on the provided data set using hundreds of different variations of the model architecture shown in Figure 2b. It compares the performance of all of these variations and returns the results of the best performing version of the model. The AutoML service automates processes that are otherwise extremely human time intensive and which require significant AI expertise.

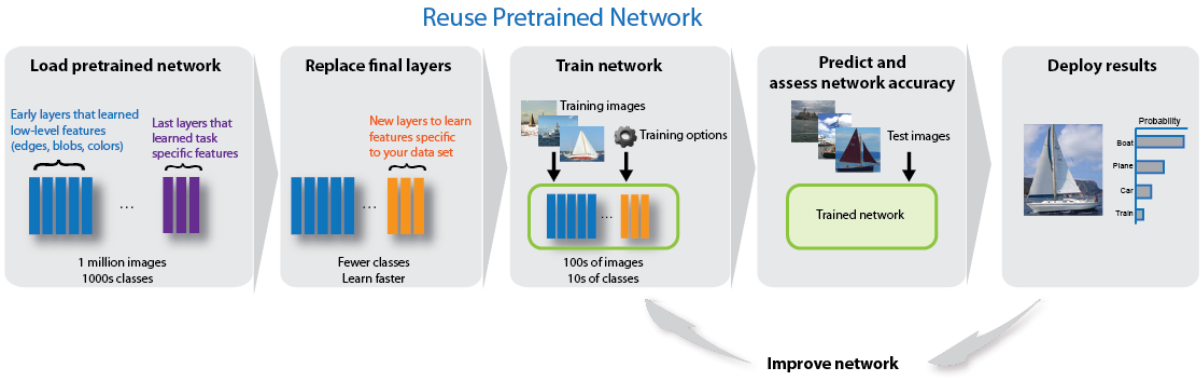


Figure 4. How transfer learning works for a pretrained network in Google AutoML.

Pros and Cons of Cloud Services

Using these commercial cloud services presents a number of pros and cons to the scientist/customer. The most significant pro is that the cost of purchasing and configuring the necessary computers is reduced to an hourly rental rate. For this experiment, the purchase price for the computers used for the work would have been approximately $46,000 for the hardware, excluding the labor hours to configure them. We were able to rent these for a cumulative time of 36 compute hours for $1,300.

Similarly, the AutoML service condenses the expertise of hundreds of Google AI scientists into an accessible software package that is usable by scientists in other domains. It is not necessary for the customer to master the extensive mathematics and computer programming required for creating DNN's from scratch. The cost to use this software is included in the computer expenses given above.

The cons are that, though customers avoid the necessity of mastering mathematics and computer programming, they also lack insight into how the model functions. As a result, they are not able to deeply investigate results that might appear to be too good or too poor. This also limits their ability to propose changes that could improve performance.

Second, the models created on the Google computers cannot be exported and used elsewhere. They can only operate on Google billed services. The details of the exact structure of the models and the settings of the hyperparameters is not available. If a customer knew these details, they could recreate the network on a different computer or service. For this reason, some researchers use Google AutoML to explore the feasibility of a problem, and when they learn that a solution does exist, they can then expend more resources to create the solution on their own computers where they have full control and ownership of the model.

RESULTS

Multiple models were created for different combinations of the video data set. These initially focused on isolating one exercise and one metric, and progressed toward combining the exercises. For the training process, 80% of the videos were used to train the DNN and 20% were used as the test set to measure the accuracy of the model that was created. The validation stage that is typically conducted after test, is embedded in the AutoML algorithm optimization process.

The DNN model seeking to match the scoring of the Ring and Rail exercise by the simulator achieved an average accuracy of 85.9% (top left data cell of Table 4) when given 481 video clips for training and testing. This means that for 85.9% of similar videos the model can be expected to return the same label as was assigned by the simulator – i.e. Expert, Intermediate, or Novice. The best balance of Precision (Type I error) and Recall (Type II error) is a Precision of 77.3% and Recall of 77.3%. The Ring & Rail exercise is the simplest of the two exercises, so one would expect the DNN algorithm to be able to score it more accurately than the Suture Sponge exercise.

Table 4. ML model performance metrics in matching GEARS and Simulator scores. Results in each cell are Average Accuracy / Precision / Recall.

Metric	Ring & Rail	Suture Sponge	Combined Exercises
Simulator	85.9/77.3/77.3%	81.0/76.3/70.8%	78.3/72.7/65.2%
GEARS	83.1/76.1/67.7% (+)	80.8/72.8/67.7%	75.3/72.9/59.3%

+ Ring & Rail exercises with the GEARS metric include only two classes - expert and intermediate.

For the Suture Sponge exercise with the Simulator metrics, the DNN model achieved 81.0% average accuracy, 76.3% Precision, and 70.8% Recall when training and testing on a dataset of 1,733 video clips. These are slightly lower than the performance of the Ring & Rail exercise. This suggests that it is more difficult for the algorithm to differentiate the actions in the Suture Sponge videos in spite of the fact that it had more samples to train on.

DNN models attempting to match the GEARS metric that was assigned by humans viewing the videos achieved slightly lower accuracy in all areas. Ring & Rail = 83.1/76.1/67.7%, Suture Sponge = 80.8/72.8/67.7%. Note that the human-assigned GEARS scores for the Ring & Rail exercises in this study were all classified as either an Expert or an Intermediate. The human evaluators did not believe that any of the subjects were performing at Novice levels in this exercise. This is probably due to the simple nature of the exercise. Figure 5 shows the formatted output of Google's cloud services for one model as it sought out the optimal settings for reducing both Type I (Precision) and Type II (Recall) errors in the Suture Sponge exercise using the GEARS metric.

After creating a dedicated model for classifying each exercise we explored the potential for creating a single model that could score videos from multiple exercises. For this project we only had access to a large set of videos from two exercises. Models were created with the combined set of video clips from both exercises, 2,227 total video clips. The accuracy of these combined models was 78.3/72.7/65.2% (average/precision/recall) for simulator scores and 75.3/72.9/59.3% for GEARS scores. These are several percentage points lower in each area, primarily due to the fact that predicting the class for two different exercises simultaneously with the same model is a difficult job. This requires that nodes and layers in the neural network either learn a more general solution to both exercises (less precise for each), or that they create dedicated pathways for each exercise (fewer network nodes for each). Both of these solutions would result in less algorithmic accuracy for a combined dataset.

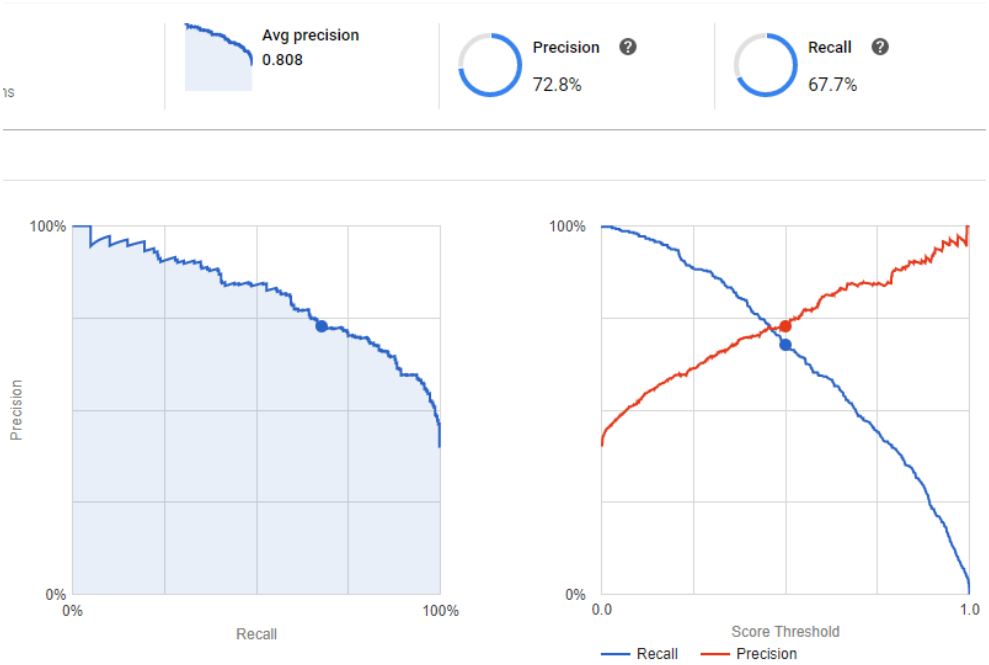


Figure 5. Plots of model precision for Suture Sponge with GEARS.

The levels of accuracy achieved from this research experiment are very encouraging. They suggest that, with further work, the DNN models might match human and simulator performance scores in the 90% accuracy range. This could be accomplished through additional fine-tuning of the classification thresholds or pruning some outliers from the existing data set. But a more powerful approach would be to simply provide many more video clips for the algorithms to learn from. 2,227 video clips are a much larger dataset than would be required for a human to learn how to score these videos, but it is relatively small when trying to teach that skill to a DNN model.

For the purposes of an educational course, a model that can reliably identify those students who exhibit Novice level skills at the conclusion would be valuable. These courses do not usually attempt to move a student to Expert level, but rather focus on improving them from Novice to a basic Intermediate level. Students who remain at the Novice level may be recommended for additional training before receiving certification.

Confusion Matrix

A Confusion Matrix is often used to assist in understanding where an ML model is making mistakes in its predictions. Table 5 is an example of this matrix for the Suture Sponge exercise with the GEARS metric. It shows that the model correctly labeled Novice performance videos 83.59% of the time. But, 14.06% of the samples it tested on were incorrectly labeled as Intermediate, and 2.34% were incorrectly labeled as Experts. It is logical that the most common mistakes would be made across the adjacent boundary between Novice and Intermediate where samples would presumably be the most similar. Performance at the Intermediate and Expert levels is less accurate. Most notable is that 31.65% of the samples that were supposed to be labeled as Intermediate were labeled as Novice by the model.

Table 5. Confusion Matrix for Suture Sponge GEARS score model.

		Actual		
Predicted		Novice	Intermediate	Expert
	Novice	**83.59%**	31.65%	7.50%
	Intermediate	14.06%	**60.43%**	25.00%
	Expert	2.34%	7.91%	**67.50%**

Performance Improvement Techniques

Understanding potential causes of the errors shown in the confusion matrix can suggest techniques that could be used to improve performance and reduce errors. Recall that the Suture Sponge exercise video was cut from one 150-second video into 15 or more 10-second video clips. Perhaps the key feature that differentiates an Intermediate from a Novice appears in just a few of the shorter clips and the majority of the clips look exactly the same for both levels. If this were the case, the training dataset would have to become much larger to improve model performance. Alternatively, performance may be improved by removing video clips that do not demonstrate differentiable skills. It is also possible that the threshold between the two levels is set at the wrong place. Recall that three methods for setting these thresholds were explored. For the purposes of creating a DNN model, different methods may be more effective. Notice that a similar effect seems to exist for Expert level videos, where 67.5% were labeled correctly and 25% were misidentified as Intermediate level. These may also be suffering from a non-optimal setting of the thresholds.

This line of questioning demonstrates why a Confusion Matrix is such a valuable tool in understanding the performance of a model. Iteratively working with the data set to address identified mistakes is a normal part of developing a production-ready ML service.

FUTURE WORK

As we analyzed the results achieved by this research experiment, we identified several opportunities for future work which could improve the results. The most powerful tool for improving the performance of any DNN is to provide it with more data to learn from. We were able to create a dataset of over 2,000 video clips to learn three classification levels (Expert, Intermediate, Novice) for both exercises (Ring & Rail, Suture Sponge). This dataset was large enough to provide encouraging results, but the algorithms could be much better if they had 10,000 video clips to work with.

We chose to cut the exercise videos into 10-second clips. These original videos could be cut into 5-second or 3-second clips to create a dataset that is 2X or 3X larger. However, as we suggested earlier, it is important that these shorter clips contain an action that truly demonstrates differentiable levels of skill. If these shorter videos provide more instances of activity that looks the same at every skill level, then they would not significantly improve the performance of the algorithm.

We could use the results from the Confusion Matrix to cull out video clips that are always misclassified. This kind of manual dataset pruning can be very helpful, but is labor intensive. It also threatens to over-train the network on specific features, which leads to underperformance when it is used as a commercial product. Experienced ML scientists prefer to address the misclassification problem by adjusting the thresholds or by providing larger datasets for training.

As expected, creating a model that can generalize across multiple exercises will require significantly more videos for training. While a generalized model holds some attraction, if this work were deployed as a commercial service, the customers would not need to know whether the answers they received came from a single general model or several specialized models. Therefore, a generalized model may not justify the expenses necessary to collect enough data to create it.

For this project we classified the videos using the standard surgical skills levels of Expert, Intermediate, and Novice. However, if the model is only being used to determine whether a student has improved beyond a defined skill level, such as from Novice to Intermediate level, the classifications could be simplified to a binary pair such as Above/Below threshold, or Pass/Fail. Training to fewer classes can achieve higher accuracy for the same data set size.

Finally, even after a deep learning model has been deployed as a product, it is possible to continue stress testing it by presenting it with videos for which the desired classification is known. This allows those who are not convinced

of the model's ability to probe its performance for weaknesses. These stress tests may identify special conditions that were not included in the training and testing process for the model (e.g. different lighting, colors, object shapes, movement patterns), and which may call for retraining with sufficient data instances containing those special conditions.

CONCLUSIONS

This project was a research experiment to determine whether the currently available DNN algorithms, with transfer learning and AutoML, could be used to replace the scoring that is normally performed by human experts. The results are encouraging that ML techniques are very close to being able to perform this job. We believe that with larger training datasets and fine-tuning of the models and parameters, a model capable for 90+% classification accuracy is possible. From our experience it appears that a dataset of 10,000 video clips for each exercise, and where those video clips demonstrate real skill differences may be sufficient to create models that can be used to replace human evaluators.

This experiment was performed with surgical training videos. But the principles, methods, and cloud services used could be applied equally to scoring the quality of performance of any video recorded action. There are currently teams working to apply these methods to the scoring of sports and dance performances. They could also be used in military applications to score the performance of a team in a shoot-house scenario or other live tactical training events.

REFERENCES

Checcucci E, Autorino R, Cacciamani GE, Amparore D, De Cillis S, Piana A, Piazzolla P, Vezzetti E, Fiori C, Veneziano D, Tewari A, Dasgupta P, Hung A, Gill I, Porpiglia F. (Feb 2020). "Artificial intelligence and neural networks in urology: current clinical applications." *Minerva Urology and Nephrology Journal*, 72(1):49-57. doi: 10.23736/S0393-2249.19.03613-0. Epub 2019 Dec 12.

Dubin, A., Smith, R., Julian, D., Tanaka, A., Mattingly, P. (August 2017). "A Comparison of Robotic Simulation Performance on Basic Virtual Reality Skills: Simulator Subjective vs. Objective Assessment Tools", *Journal of Minimally Invasive Gynecology*, August, 2017.

Dubin, A., Smith, R., Julian, D., Tanaka, A., Mattingly, P. (February 2018). "A Model for Predicting the GEARS Score from Virtual Reality Surgical Simulator Metrics", *Surgical Endoscopy*, February 2018.

Goh, A., Goldfarb, D., Sander, J., & Dunkin, B. (Jan 2012). "Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic surgical skills". *Journal of Urology*, 187(1), 247-252.

Google Cloud Team. (2018). "Machine Learning with TensorFlow on Google Cloud Platform", online training course. Available at: https://www.coursera.org/specializations/machine-learning-tensorflow-gcp.

Google Video Intelligence Service. (2020). https://cloud.google.com/video-intelligence/

Hosseini, H., Xiao, B., & Poovendran, R. (2017). "Deceiving Google's Cloud Video Intelligence API Built for Summarizing Videos." *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1305-1309.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep Learning". *Nature*, 521(7553), 436-445.

Liu, M. & Curet, M. (2015). "A Review of training research and virtual reality simulators for the da Vinci surgical system". *Teaching and Learning in Medicine*, 27(1).

Martin J.A., Regehr G., & Reznick R. (1997). "Objective structured assessment of technical skill (OSATS) for surgical residents." *British Journal of Surgery*, 84(2), 273–278.

Smith, R. & Julian, D. (2019). "Psychomotor Skills Assessment via Human Experts, Simulators, and Artificial Intelligence". 2019 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).

van Hove P.D., Tuijthof G.J., Verdaasdonk E.G., Stassen L.P., & Dankelman J. (2010). "Objective assessment of technical surgical skills." *British Journal of Surgery*, 97(7), 972–987.