

Assessment of Training Progression on a Surgical Simulator Using Machine Learning and Explainable Artificial Intelligence Techniques

Constantinos Loukas^a and Konstantina Prevezanou^b

Medical Physics Lab, Medical School, National and Kapodistrian University of Athens,
Mikras Asias 75 str., Athens, Greece

Keywords: Surgery, Virtual Reality, Skills Assessment, Machine Learning, Explainable AI (XAI).


Abstract: Surgical training on VR simulators provides an efficient education paradigm in laparoscopic surgery. Most methods for skills assessment focus on the analysis of video and kinematic data for self-proclaimed skill classification and technical score prediction. In this paper we evaluate a machine learning (ML) framework for classifying the trainee's performance with respect to the phase of training progression (beginning vs. end of training and beginning vs. middle vs. end of training). In addition, we leverage techniques from the field of Explainable Artificial Intelligence (XAI) to obtain interpretations on the employed black-box ML classifiers. Three surgical training tasks with significant educational value were selected from a training curriculum followed by 23 medical students. Five machine learning algorithms and two model-agnostic XAI methods were evaluated using performance metrics generated by the simulator during task performance. For all surgical tasks, the accuracy was >84% and >86% in the 2- and 3-class classification experiments, respectively. The XAI methods seem to agree on the relative impact of each performance metric. Features related to hand-eye coordination and bimanual dexterity (e.g. economy of movements, instrument pathlength and number of movements), play the most important role in explaining the classification results.


1 INTRODUCTION

Virtual reality (VR) simulators have been increasingly recognized as valuable tools for training and assessment of surgical skills. Especially for laparoscopic surgery, where surgeons are faced with additional challenges compared to open surgery (such as reduced depth perception, working with elongated instruments and minimal force feedback), VR simulation offers an efficient education paradigm compared to traditional training on bench top models and inanimate video trainer boxes (Guedes *et al.*, 2019). Specifically, VR systems include a plethora of photorealistic scenarios ranging from basic to procedural skills (Matzke *et al.*, 2017) and advanced surgical scenarios (Ikonen *et al.*, 2012). With the aid of dummy instruments trainees are able not only to realistically interact with primitive virtual objects (such as pegboard, suture, needle, etc.), but also to perform demanding surgical tasks (e.g. bowel suturing, gallbladder dissection), and entire surgical

procedures (e.g. cholecystectomy, appendectomy, hernia repair surgery, etc.).

In addition to the safe and flexible training environment, another significant advantage of VR simulators lies on their ability to capture the hand kinematics and interaction events with the virtual world via motion tracking sensors embedded into the mechanical interface of the dummy surgical tools (Dosis *et al.*, 2005). Upon task completion the simulator generates an assessment report that includes key metrics of task performance with respect to time (e.g. task and activity completion time), technical competency (instrument pathlength, number of movements, etc.), safety (involuntary errors such as tissue injuries, misplaced clips, suture damage, etc.), and dexterity (% of adhesions removed, number of knots locked, alternating throws, etc.). However, the underlying relation and educational interpretation of these parameters with respect to the level of surgical competency that

^a  <https://orcid.org/0000-0001-7879-7329>

^b  <https://orcid.org/0000-0002-3091-7413>

trainees aim to achieve is still under investigation (Varras *et al.*, 2020).

Objective computer-aided technical skill evaluation (OCASE-T) has received an increasing amount of attention over the past few years for several reasons (Vedula, Ishii and Hager, 2017). In addition to saving time and money, it allows novice surgeons to train effectively and with greater flexibility until they reach an adequate level of competency by receiving constructive feedback in the absence of human supervision. Moreover, the assessment output includes quantitative measures of performance that allow trainees to evaluate their dexterity level with respect to that achieved by expert surgeons. Over the last decade, several studies supported the effectiveness of VR simulators by demonstrating their construct validity for laparoscopic skills assessment (Larsen *et al.*, 2006), comparing the learning curves after training on a VR curriculum with traditionally trained groups (Aggarwal *et al.*, 2007), proposing assessment methodologies based on quantitative analysis of key laparoscopic skills (Loukas *et al.*, 2011), and highlighting skill retention following laparoscopic simulator training (Stefanidis *et al.*, 2005).

Early approaches to OCASE-T focused on hidden Markov models (HMMs), which consider the multidimensional hand motion signal as an unobserved state sequence relating to a set of primitive gestures (Rosen *et al.*, 2006). After training a model for a respective skill-level based on signals from the same class (e.g. novice, intermediate, expert), a statistical distance is employed to compare the likelihood of a new performance to those in the training set and hence return the corresponding class.

Later works adopted approaches that extract features, or descriptive metrics, from kinematic signals in order to determine the skill level. For example, data from an armband device (e.g. acceleration, orientation, etc.), was employed in (Kowalewski *et al.*, 2019) for gesture detection and skill assessment in laparoscopic suturing. In (Fard *et al.*, 2018) features such as instrument pathlength and smoothness are extracted to train various machine learning (ML) algorithms to classify experts vs. novices. Other works employ entropy, texture and frequency features, for self-proclaimed skill classification and performance score prediction using ML regression models (Zia *et al.*, 2018). Alternative sources such as electroencephalogram (EEG) and electromyogram (EMG) have also been proposed for laparoscopic expertise evaluation, but the reported accuracy and applicability in a real-surgical environment is limited compared to the kinematic

signals (Shafiei *et al.*, 2021), (Fogelson *et al.*, 2004), (Soto Rodriguez *et al.*, 2023).

Recently, deep learning techniques, for example 1D convolutional neural networks (CNN) and time-series models (such as long short-term memory (LSTM) and temporally convolutional networks), have been employed to capture and process sequential information from the kinematic signals. In (Wang and Majewicz Fey, 2018) various CNN models are proposed to assess surgical performance by extracting patterns in the surgeon's maneuvers in robotic surgery tasks. In (Benmansour, Malti and Jannin, 2023) a CNN+BiLSTM architecture that takes advantage of both temporal and spatial features of kinematic data was proposed for performance score prediction in robotic surgery tasks. Moglia *et al.* utilized data from a robotic surgery VR simulator to develop an ensemble deep neural network (DNN) for predicting the number of attempts and training time required to attain proficiency (Moglia *et al.*, 2022).

Most methods for surgical skills assessment focus on the analysis of kinematic data. However, obtaining this data requires access to the application programming interface of the VR device, which is not always feasible due to permission constraints from the owner company. Another approach is to employ a separate tracking system with motion sensors attached to the surgeon's hand, or the dummy laparoscopic tools, which introduces additional complexity and data management issues in the overall training process. Moreover, most works provide limited information about the trainee's progress while training on the VR simulator and they focus on classification in predefined skill classes (e.g. novices, intermediates, experts). Being able to provide immediate constructive feedback to the trainees about their training progress using the performance metrics generated by the simulator can alleviate many of these constraints.

In this paper we propose an ML approach for classifying the trainee's performance with respect to the phase of training progression on a laparoscopic VR simulator (beginning vs. end of training and beginning vs. middle vs. end of training). In addition, we leverage techniques from the field of Explainable Artificial Intelligence (XAI) to obtain not only interpretations on the employed black-box ML classifiers, but also better understanding about the most valuable metrics of surgical performance. Specifically, we utilize two well-known XAI techniques: *Permutation feature importance (PFI)* (Fisher, Rudin and Dominici, 2019) and a more advanced one based on *SHapley Additive exPlanation (SHAP)* (Lundberg, Allen and Lee, 2017). Both

methods are applied to derive model-agnostic, post-hoc interpretations on five ML classifiers: Support Vector Machine, Linear Discriminant Analysis, Random Forest, Linear Regression and Gaussian Naïve Bayes. In addition to providing results on skills classification in three surgical tasks performed by trainees following a structured VR simulation training curriculum, we compare XAI techniques based on the ranking of feature importances. Additionally, we utilize visualization tools (summary plots) to rank the performance metrics and investigate their effect on model decisions.

2 METHODS

2.1 Dataset

The study included 23 medical students with no experience in laparoscopy. The participants followed a structured training curriculum on a laparoscopic VR simulator (LapMentor™, Surgical Science Sweden AB). In particular, the participants performed 9 sequential laparoscopic tasks selected from the ‘Basic Skills’ module of the simulator. The training goal was to reach for 3 consecutive times the performance of an expert, defined by quantitative thresholds on predefined performance metrics, separately for each surgical task. Upon reaching these thresholds the student was allowed to advance to the next task. In this study we focused on three training tasks with significant educational value according to the evaluation of our surgical education board (Figure 1): *Clipping and Grasping* (Task 5), *Two-Handed Maneuvers* (Task 6) and *Cutting* (Task 7).

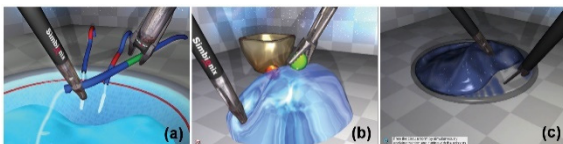


Figure 1: Screen shots of the three surgical training tasks performed on the VR simulator: (a) Clipping and Grasping (Task 5), Two-Handed Maneuvers (Task 6) and Cutting (Task 7).

For Task 5 the aim was to grasp and clip leaking ducts within specific segments. Red segments appear on the ducts at the beginning of the trial whereas the segment turns green only when grasped properly. After grasping the leaking duct, the trainee had to use the clipper to place a clip within the green segment only to stop the leakage. The task needs to be completed before the pool overflows. For Task 6 the

goal was to use 2 grasping tools to locate a jelly mass and move part of the jelly aside to expose a ball. While holding the jelly aside, the trainee had to use the other tool to grasp the exposed ball and place it in the Endobag near the jelly. The aim of Task 7 was to apply traction and cut safely and accurately a circular tissue-like form using a grasper and scissors. The grasper is used to retract the form and expose a safe cutting area while the scissors are used for cutting the form. In terms of educational objectives, the tasks aim to provide training on key technical skills such as hand-eye coordination, bimanual dexterity, tissue handling and laparoscopic orientation.

Table 1: Features employed per surgical task. ‘I’ denotes instrument and ‘+’ or ‘-’ denote whether the metric is available in the corresponding task or not, respectively.

Description	Code name	Task 5	Task 6	Task 7
Trial number	TN	+	+	+
Average speed of left I (cm/sec)	Speed-L	+	+	+
Average speed of right I (cm/sec)	Speed-R	+	+	+
Number of movements-left I	#Move-L	+	+	+
Number of movements-right I	#Move-R	+	+	+
Total pathlength of left I (cm)	PL-L	+	+	+
Total pathlength of right I (cm)	PL-R	+	+	+
Trial completion time	Time	+	+	+
Economy of movement-left I (%)	EOM-L	+	+	-
Economy of movement-right I (%)	EOM-R	+	+	-
# clipped ducts	#ClipD	+	-	-
Total # clipping attempts	#ClipAtt	+	-	-
# exposed green balls collected	#GreenB	-	+	-
# lost balls which miss the basket	#LostB	-	+	-
# cutting maneuvers	#CuttM	-	-	+
# cutting maneuvers without tissue injury	#CuttM-NoInj	-	-	+
# retract. operations	#React	-	-	+

The number of trials required to successfully complete each task varied from 9-57 (median=23, 23 and 16 for Task 5, 6 and 7, respectively). Table 1 shows the performance metrics (*i.e.* features) that were considered for further analysis. Overall, 12 features were utilized. Ten of these were common to all or two of the tasks (TN, Speed-L, Speed-R, #Move-L, #Move-R, PL-L, PL-R, Time, EOM-L, EOM-R), while the other two were task specific

2.2 ML Framework

We employed an ML methodology to classify the trainees' trials into different phases of training progression on the VR simulator. In particular, the first experiment aimed to classify trials as being close to the Beginning (BT) or the End (ET) of training on a particular task. For this purpose, we included three random trials before and after the median training attempt (trial) of each subject. The second experiment aimed to classify the students' trials into three classes: Beginning (BT), Middle (MT), and End (ET) of training. For this purpose, the training trials of each subject were first divided into three equal parts based on their order in the training sequence, and then three random trials from each part were selected. Given that our study included 23 subjects, the total number of samples (*i.e.* trials) for the first and second experiment were $n_1=138$ and $n_2=207$, respectively.

Five ML algorithms that were previously applied to similar classification tasks, such as self-proclaimed skill classification (Mirchi *et al.*, 2020), (Siyar *et al.*, 2020), (Winkler-Schwartz *et al.*, 2019), were employed in this study: Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Random Forest (RF), Linear Regression (LR) and Gaussian Naïve Bayes (GNB). For SVM we employed two variants, one with linear kernel (SVM-Lin) and another one with a radial basis function (SVM-RBF). For SVM, the regularization parameter was set to $C=1$, the penalty term was set to L2 and the kernel coefficient $\gamma=1/n_f$, where n_f is the number of features. For LR, the regularization parameter was set to $C=1$, the penalty term was squared L2, and the optimization solver was Limited-memory BFGS. For LDA we employed the Singular Value Decomposition (SVD) solver with no prior class probabilities. For RF, the number of trees was set to 100, the number of min samples required to split an internal node was 2, the number of minimum samples required to be at a leaf node was 1, and the number of features to consider when looking for the best split was set to $\sqrt{n_f}$. The GNB was based on a Gaussian kernel without prior class probabilities

2.3 XAI Techniques

Recent advancements in XAI employ *model-agnostic* interpretation methods to achieve explanations for complex ML models. Unlike methods that are model-specific, *model-agnostic* interpretations offer more flexibility by decoupling the model from its explanations (Ribeiro, Singh and Guestrin, 2016). Hence, one may apply the same XAI technique to the predictions of different ML models trained/tested on the same dataset, allowing for comparison of interpretation results. Model-agnostic interpretation methods can be categorized into local and global methods (Molnar, 2022). Local methods aim to explain individual predictions whereas global methods describe how features affect the prediction on average. In this study we employed two commonly used global explanation methods that provide summary plots of feature importance: *Permutation Feature Importance (PFI)* and *SHapley Additive exPlanations (SHAP)*.

PFI is based on the simple idea of measuring the decrease in the prediction accuracy of the model when the values of a feature are permuted, thereby breaking the relationship between the feature and the true outcome (Fisher, Rudin and Dominici, 2019). In this work each feature was permuted 15 times and the *feature importance score* for each feature was computed as the average accuracy based on the predictions of the permuted data in the test-set.

SHAP is based on the game theoretically optimal Shapley values and aims to explain the prediction of an instance by computing the contribution of each feature to the prediction (Lundberg, Allen and Lee, 2017). Features with large absolute Shapley values are important. Although the technique provides a Shapley value $\phi^{(i)}$ for any instance $x^{(i)}$, the global importance per feature (*feature importance score*) can be obtained by averaging the absolute Shapley values over all feature values (*i.e.* instances) in the evaluated dataset, which in our case was the test-set.

3 EXPERIMENTAL RESULTS

For each experiment evaluated in this study (2-class or 3-class classification), the dataset was randomly split into a training set (70%) and a test set (30%), ensuring that the class frequencies were preserved in both sets. The ML models' performance was measured in terms of Accuracy (Acc), Precision (Pre) and Recall (Rec).

For 2-class classification, the positive and negative class was BT and ET, respectively. For 3-

class classification Acc was calculated as the sum of the diagonal elements of the confusion matrix (CM) over the total number of test samples. For Pre and Rec we adopted a 'macro' average approach by calculating each metric individually for every label, which was considered as the positive class, and subsequently determining their unweighted mean.

3.1 Classification Results

Tables 2 and 3 show the results for the first (2-class classification) and the second (3-class classification) experiment, respectively. The results are shown separately for each surgical task. Note also that for these experiments we excluded the TN (trial number) feature, so in total 11 out of 12 features were considered (see Table 1). The reason for excluding TN was to examine the potential for broader applicability of the proposed methodology in evaluating the students' skills before entering the training curriculum (see also Discussion). Overall, the results in both experiments show better performance for Task 5 and 6 compared to Task 7. This may be due to the fact that by the time the trainees start training on Task 7, they have already mastered the skills required to achieve proficiency in

Table 2: Performance comparison for 2-class classification without using the trial number feature (w/o TN). Best results column-wise are shown in bold.

Method	Acc (%)	Pre (%)	Rec (%)
Task 5 (w/o TN)			
SVM-Lin	97.2	97.4	97.2
SVM-RBF	97.2	97.4	97.2
LDA	94.4	95.0	94.4
RF	97.2	97.4	97.2
LR	97.2	97.4	97.2
GNB	94.4	95.0	94.4
Task 6 (w/o TN)			
SVM- Lin	94.7	94.7	94.7
SVM-RBF	92.1	92.2	92.1
LDA	86.8	86.9	86.8
RF	94.7	94.7	94.7
LR	84.2	84.6	84.2
GNB	86.8	87.7	86.8
Task 7 (w/o TN)			
SVM- Lin	84.2	85.8	84.2
SVM-RBF	84.2	84.6	84.2
LDA	78.9	79.3	78.9
RF	81.6	81.7	81.6
LR	78.9	80.3	78.9
GNB	76.3	78.3	76.3

this task. Moreover, the results for 2-class classification are much better than the results for 3-class classification, as expected. Overall, the best model is SVM. SVM-Lin shows the best performance compared to all other algorithms as the Acc, Pre and Rec were the highest in five out of the six experimental runs (2 experiments for 3 surgical tasks). In particular, with regard to the first experiment (2 classes), Acc for Task 5, 6 and 7 was close to 97%, 95% and ~84%, respectively. For the second experiment (3-classes) Acc was lower, about 76%, 82% and 67%, for Task 5, 6 and 7, respectively. No significant difference was found between the Pre and Rec values in both experiments.

Table 3: Performance comparison for 3-class classification without using the trial number feature (w/o TN). Best results column-wise are shown in bold.

Method	Acc (%)	Pre (%)	Rec (%)
Task 5 (w/o TN)			
SVM-Lin	75.9	77.8	75.9
SVM-RBF	74.1	74.2	74.1
LDA	64.8	67.8	64.8
RF	72.2	72.9	72.2
LR	66.7	67.0	66.7
GNB	70.4	70.3	70.4
Task 6 (w/o TN)			
SVM- Lin	80.7	80.8	80.7
SVM-RBF	82.5	83.1	82.5
LDA	73.7	74.9	73.7
RF	75.4	75.5	75.4
LR	78.9	79.6	78.9
GNB	77.2	77.8	77.2
Task 7 (w/o TN)			
SVM- Lin	67.3	70.0	67.6
SVM-RBF	63.5	63.6	63.7
LDA	59.6	63.2	59.7
RF	61.5	62.1	61.8
LR	63.5	63.2	63.7
GNB	59.6	59.0	59.8

Figure 2 shows the CMs for SVM-Lin when using 11 features (*i.e.* without TN). For the 2-class classification experiment, the performance is similar for the two classes. For 3-class classification the best performance is for the BT class and the second-best for the ET class. Most of the confusion occurs between MT and ET, especially when the ground-truth is the MT class. This may be because trainees in the middle of training have acquired much more skills compared to the beginning, and thus are close to achieving the required proficiency to complete the task successfully.

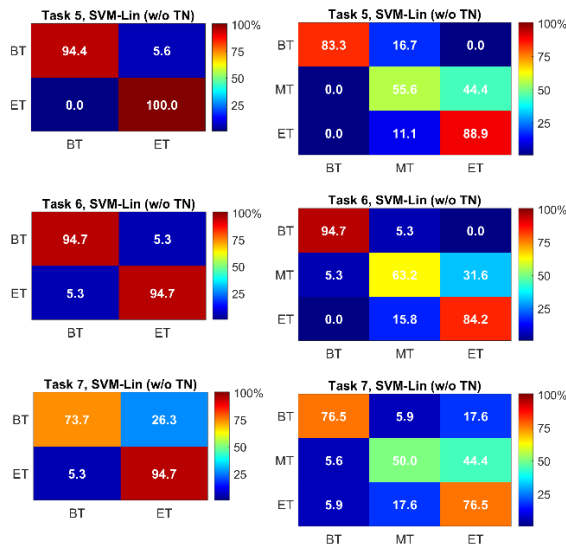


Figure 2: Color-coded confusion matrices for SVM-Lin when using 11 features (*i.e.* without TN). The X and Y-axis represent predicted and ground truth labels, respectively (left/right CMs: 2-/3-class classification).

Table 4: Performance comparison for 3-class classification when using all 12 features (*i.e.* including TN: w/ TN). Best results column-wise are shown in bold.

Method	Acc (%)	Pre (%)	Rec (%)
Task 5 (w/ TN)			
SVM-Lin	88.9	90.2	88.9
SVM-RBF	83.3	86.3	83.3
LDA	81.5	84.4	81.5
RF	87.0	88.0	87.0
LR	85.2	86.7	85.2
GNB	79.6	80.2	79.6
Task 6 (w/ TN)			
SVM-Lin	93.0	93.3	93.0
SVM-RBF	94.7	95.5	94.7
LDA	87.7	89.3	87.7
RF	91.2	91.9	91.2
LR	93.0	93.3	93.0
GNB	86.0	86.5	86.0
Task 7 (w/ TN)			
SVM-Lin	82.7	82.4	82.9
SVM-RBF	86.5	87.6	86.5
LDA	80.8	81.1	81.0
RF	82.7	83.3	83.0
LR	80.8	81.5	80.9
GNB	76.9	80.3	77.3

Table 4 presents the performance for 3-class classification when using all available features (*i.e.* 12 features, including TN). Compared to the results shown in Table 3, it may be seen that the performance of all algorithms has been improved in all surgical tasks by about 13% for Task 5 and 6 and close to 20%

for Task 7. SVM-RBF yields the best performance for Tasks 6 and 7 with Acc close to 95% and 87%, respectively. For Task 5 the best performance is shown by SVM-Lin with Acc close to 90%, whereas for Task 6 and 7 the best method is SVM-RBF.

Figure 3 shows the CMs for SVM-Lin when using all available features (*i.e.* with TN). Compared to the corresponding CMs shown in Figure 2, the best performance is presented when the ground-truth classes are BT and MT for Task 5 and 6 and ET for Task 7, but only slightly compared to MT. Similarly to the corresponding results in Figure 2, the greatest confusion occurs between the MT and ET classes, but now the misclassification is greater when the ground-truth class is ET, probably because the TN values of this class are more similar to those of MT compared to the other features.

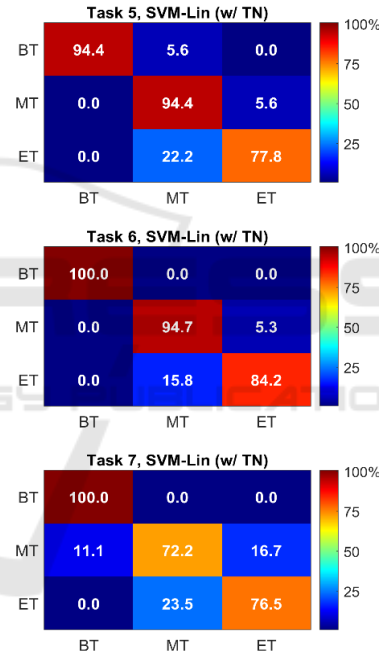


Figure 3: Color-coded confusion matrices for SVM-Lin when using all 12 features (*i.e.* with TN). The X and Y-axis represent predicted and ground truth labels, respectively (3-class classification CMs).

3.2 XAI Results

Figure 4 demonstrates the feature significance of SVM-Lin for the two XAI methods (PFI and SHAP), separately for each surgical task and classification experiment (2- and 3-class classification with or without the TN feature). To allow for better comparison between PFI and SHAP, the feature importance score was normalized by the score sum of all features used in each corresponding experiment.

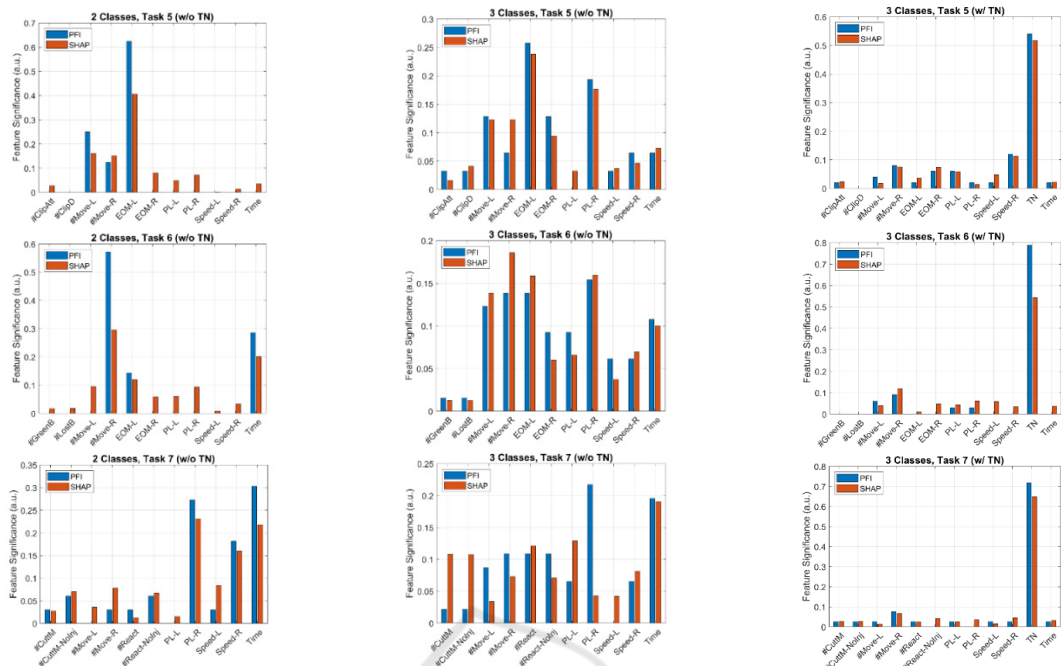


Figure 4: Normalized feature importance score (feature significance) of the two XAI methods for each classification experiment and surgical task.

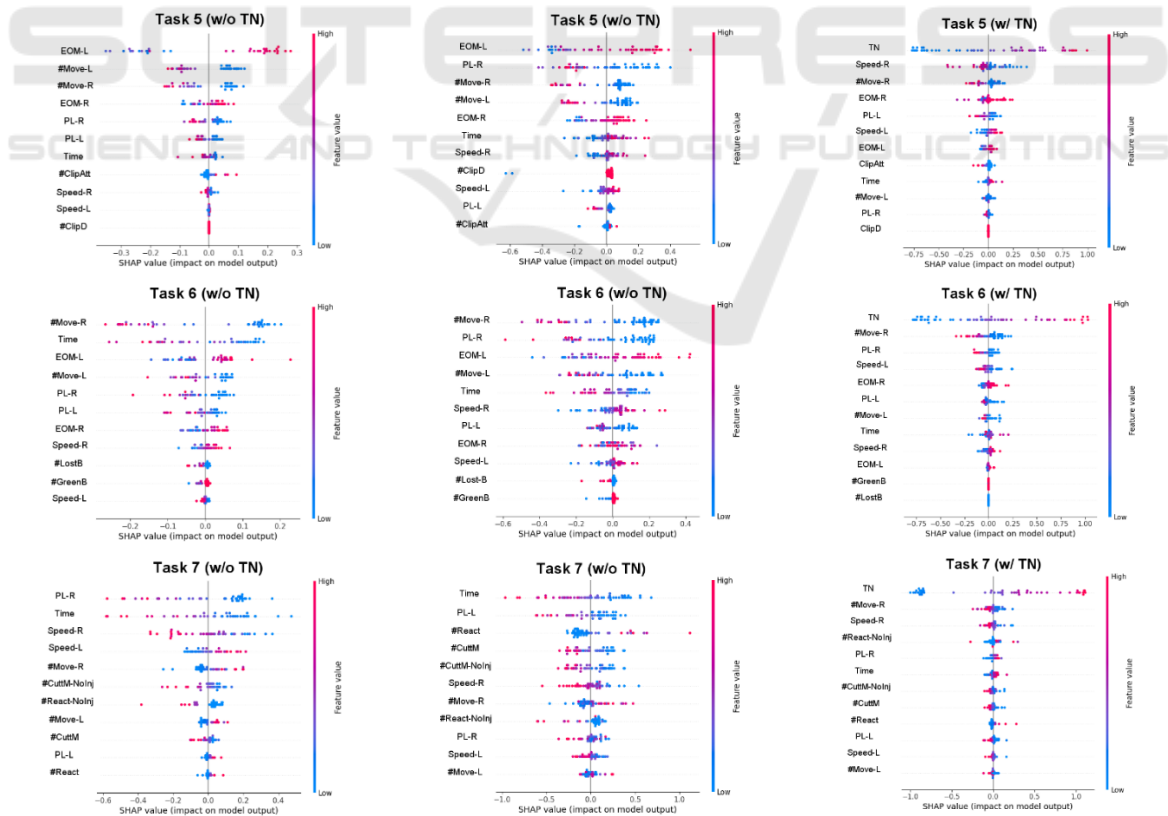


Figure 5: SHAP summary plots for SVM-Linear. Each plot corresponds to a different surgical task and classification experiment.

Overall, the two XAI methods seem to agree on the relative impact of each feature. For example, for 2-class classification the two features with the greatest impact are: EOM-L, #Move-L (Task 5), #Move-R, Time (Task 6) and Time, PL-R (Task 7). For 3-class classification without TN (w/o TN) the two features with the greatest impact are: EOM-L, PL-R (Task 5),

#Move-R, PL-R (Task 6) and Time, PL-R/PL-L (Task 7). Thus, features related to technical skill and bimanual dexterity (e.g. economy of movement, # of movements and instrument pathlength) seem to play the most important role, followed by the time parameter that relates to how fast the trainee completes the task. For 3-class classification with TN (w/ TN), the TN feature has by far the greatest impact among all features, having a relative score >50% for Task 5, and >70% for Task 6 and Task 7.

Figure 5 shows SHAP summary plots, based on the SVM-Lin model, for each surgical task and classification experiment (2- or 3-class classification and with or without using the TN feature). This type of plot can be used to visualize the relative impact of all features over the entire dataset (in this case the test-set). Features are sorted by the sum of their SHAP value magnitudes across all samples. SHAP values less than 0, equal to 0 and greater than 0 signify negative contribution, no contribution and positive contribution, respectively. For each instance, the given explanation is represented by a single dot on each feature row and the x position of the dot is determined by its SHAP value. The vertical colorbar to the right of the axes indicates the mapping of feature values (from low to high). In essence, more proficient technical skills, which are acquired towards the end of training phase, are indicated by higher values of EOM (economy of movement) and lower values of pathlength (PL) or/and number of movements (#Move), as expected.

4 CONCLUSIONS

In this paper we evaluate an ML framework for classifying trainees' performance with respect to the phase of training progression on a VR laparoscopic simulator using the output performance metrics (features). SVM showed the best performance with >84% accuracy in the 2-classification experiments when using 11 features (*i.e.* w/o TN) and >86% in the 3-class classification experiment when using 12 features (*i.e.* w/ TN), in all three surgical tasks evaluated. The reason for not using TN in the initial experiments was to allow for broader application

potential of the proposed methodology. For example, in addition to informing a trainee about the phase of his/her training progression on the simulator, the proposed framework could also be used to inform trainees outside the training curriculum about their skill level. In this case, the trainee could obtain potentially useful information about the training effort required to successfully complete a particular surgical task.

With respect to the XAI experiments, the two XAI methods (PFI and SHAP) seem to agree on the relative impact of each performance metric. Features related to technical skills and bimanual dexterity seem to play the most important role both in the 2- and 3-class classification experiments (e.g. EOM, PL and #Move). Goal-oriented features with respect to each task seem less important in explaining the classification results. When used, the TN feature seems to outperform all other features in the 3-class classification experiments, probably due to the similar number of trials performed in each phase of training progression by all students.

Despite the study's findings, some limitations must be addressed. First, our dataset includes training trials from 23 medical students and may not capture the overall variability in surgical skill acquisition. In future research we plan to increase the sample size by including more trainees and expert surgeons. Second, although the construct validity of VR surgical simulators has been addressed by several studies in the past (Aggarwal *et al.*, 2007), (Aggarwal *et al.*, 2009), in the future we aim to examine the educational value of our training curriculum by evaluating it on real-world surgical tasks for groups with and without VR training. Third, in this study we examined the classification of training progress in predefined classes. The classes were defined by hard thresholds, which separated each subject's trials into equal parts based on their order in the training sequence. As future work we intend to develop a framework that predicts the Objective Structured Assessment of Technical Skill (OSATS) score and thus obtain a grade of the training progress (Martin *et al.*, 1997). Furthermore, subsequent application of XAI techniques could provide students with valuable insight into the progression of their skills, thereby enhancing the overall quality of surgical training.

ACKNOWLEDGEMENTS

The author thanks Special Account for Research Grants and National and Kapodistrian University of Athens for funding to attend the meeting.

REFERENCES

- Aggarwal, R., Ward, J., Balasundaram, I., Sains, P., Athanasiou, T., and Darzi, A. (2007). Proving the effectiveness of virtual reality simulation for training in laparoscopic surgery. *Annals of Surgery*, 246(5), 771–779.
- Aggarwal, R., Crochet, P., Dias, A., Misra, A., Ziprin, P., and Darzi, A. (2009). Development of a virtual reality training curriculum for laparoscopic cholecystectomy. *British Journal of Surgery*, 96(9), 1086–1093.
- Benmansour, M., Malti, A., and Jannin, P. (2023). Deep neural network architecture for automated soft surgical skills evaluation using objective structured assessment of technical skills criteria. *International Journal of Computer Assisted Radiology and Surgery*, 18(5), 929–937.
- Dosis, A., Aggarwal, R., Bello, F., Moorthy, K., Munz, Y., Gillies, D., and Darzi, A. (2005). Synchronized video and motion analysis for the assessment of procedures in the operating theater. *Archives of Surgery*, 140(3), 293–299.
- Fard, M. J., Ameri, S., Darin Ellis, R., Chinnam, R. B., Pandya, A. K., and Klein, M. D. (2018). Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *International Journal of Medical Robotics and Computer Assisted Surgery*, 14(1), 1–10.
- Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20, 177.
- Fogelson, N., Loukas, C., Brown, J., and Brown, P. (2004). A common N400 EEG component reflecting contextual integration irrespective of symbolic form. *Clinical Neurophysiology*, 115(6), 1349–1358.
- Guedes, H. G., Câmara Costa Ferreira, Z. M., Ribeiro de Sousa Leão, L., Souza Montero, E. F., Otoch, J. P., and Luiz de Almeida Artifon, E. (2019). Virtual reality simulator versus box-trainer to teach minimally invasive procedures: A meta-analysis. *International Journal of Surgery*, 61, 60–68.
- Ikonen, T. S., Antikainen, T., Silvennoinen, M., Isojärvi, J., Mäkinen, E., and Scheinin, T. M. (2012). Virtual reality simulator training of laparoscopic cholecystectomies-A systematic review. *Scandinavian Journal of Surgery*, 101(1), 5–12.
- Kowalewski, K. F., Garrow, C. R., Schmidt, M. W., Benner, L., Müller-Stich, B. P., and Nickel, F. (2019). Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying. *Surgical Endoscopy*, 33(11), 3732–3740.
- Larsen, C. R., Grantcharov, T., Aggarwal, R., Tully, A., Sørensen, J. L., Dalsgaard, T., and Ottesen, B. (2006). Objective assessment of gynecologic laparoscopic skills using the LapSimGyn virtual reality simulator. *Surgical Endoscopy and Other Interventional Techniques*, 20(9), 1460–1466.
- Loukas, C., Nikiteas, N., Kanakis, M., and Georgiou, E. (2011). The contribution of simulation training in enhancing key components of laparoscopic competence. *The American Surgeon*, 77(6), 708–715.
- Lundberg, S. M., Allen, P. G., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.
- Martin, J.A., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., Brown, M. (1997). Objective Structured Assessment of Technical Skill (OSATS) for surgical residents. *British Journal of Surgery*, 84(2), 273–278.
- Matzke, J., Ziegler, C., Martin, K., Crawford, S., and Sutton, E. (2017). Usefulness of virtual reality in assessment of medical student laparoscopic skill. *Journal of Surgical Research*, 211, 191–195.
- Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., and Del Maestro, R. F. (2020). The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS ONE*, 15(2), e0229596.
- Moglia, A., Morelli, L., D'Ischia, R., Fatucchi, L. M., Pucci, V., Berchiolli, R., Ferrari, M., and Cuschieri, A. (2022). Ensemble deep learning for the prediction of proficiency at a virtual simulator for robot-assisted surgery. *Surgical Endoscopy*, 36(9), 6473–6479.
- Molnar, C. (2022). *Interpretable machine learning: a guide for making black box models explainable*. Ebook. <https://christophm.github.io/interpretable-ml-book/>.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint. arXiv:1606.05386v1*.
- Rosen, J., Brown, J. D., Chang, L., Sinanan, M. N., and Hannaford, B. (2006). Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model. *IEEE Transactions on Bio-Medical Engineering*, 53(3), 399–413.
- Shafiei, S. B., Durrani, M., Jing, Z., Mostowy, M., Doherty, P., Hussein, A. A., Elsayed, A. S., Iqbal, U., and Guru, K. (2021). Surgical hand gesture recognition utilizing electroencephalogram as input to the machine learning and network neuroscience algorithms. *Sensors*, 21(5), 1733.
- Siyar, S., Azarnoush, H., Rashidi, S., Winkler-Schwartz, A., Bissonnette, V., Ponnudurai, N., and Del Maestro, R. F. (2020). Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task. *Medical and Biological Engineering and Computing*, 58(6), 1357–1367.
- Soto Rodriguez, N. A., Arroyo Kuribreña, C., Porras Hernández, J. D., Gutiérrez-Gnecchi, J. A., Pérez-Escamirosa, F., Rigoberto, M. M., Minor-martinez, A., and Lorias-Espinoza, D. (2023). Objective evaluation of laparoscopic experience based on muscle electromyography and accelerometry performing circular pattern cutting Tasks: a pilot study. *Surgical Innovation*, 30(4), 493–500.

- Stefanidis, D., Korndorffer, J. R., Sierra, R., Touchard, C., Dunne, J. B., and Scott, D. J. (2005). Skill retention following proficiency-based laparoscopic simulator training. *Surgery*, 138(2), 165–170.
- Varras, M., Nikiteas, N., Varra, V. K., Varra, F. N., Georgiou, E., and Loukas, C. (2020). Role of laparoscopic simulators in the development and assessment of laparoscopic surgical skills in laparoscopic surgery and gynecology (Review). *World Academy of Sciences Journal*, 2(2), 65–76.
- Vedula, S. S., Ishii, M., and Hager, G. D. (2017). Objective assessment of surgical technical skill and competency in the operating room. *Annual Review of Biomedical Engineering*, 19(1), 301–325.
- Wang, Z., and Majewicz Fey, A. (2018). Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 13(12), 1959–1970.
- Winkler-Schwartz, A., Yilmaz, R., Mirchi, N., Bissonnette, V., Ledwos, N., Siyar, S., Azarnoush, H., Karlik, B., and Del Maestro, R. (2019). Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. *JAMA Network Open*, 2(8), e198363.
- Zia, A., Sharma, Y., Bettadapura, V., Sarin, E. L., and Essa, I. (2018). Video and accelerometer-based motion analysis for automated surgical skills assessment. *International Journal of Computer Assisted Radiology and Surgery*, 13(3), 443–455.

