

## Article

# Machine Learning Approaches for Evaluating the Progress of Surgical Training on a Virtual Reality Simulator

Konstantina Prevezanou <sup>1</sup>, Ioannis Seimenis <sup>1</sup>, Pantelis Karaikos <sup>1</sup>, Emmanouil Pikoulis <sup>2</sup>,  
Panagis M. Lykoudis <sup>2</sup> and Constantinos Loukas <sup>1,\*</sup>

<sup>1</sup> Laboratory of Medical Physics, Medical School, National and Kapodistrian University of Athens, 15772 Athens, Greece

<sup>2</sup> 3rd Department of Surgery, University General Hospital "Attikon", Medical School, National and Kapodistrian University of Athens, Rimini 1, 12461 Chaidari Attica, Greece

\* Correspondence: cloukas@med.uoa.gr

**Abstract:** Automated assessment of surgical skills is crucial for the successful training of junior surgeons. Twenty-three medical students followed a structured training curriculum on a laparoscopic virtual reality (VR) simulator. Three surgical tasks with significant educational merit were considered (Tasks 5, 6, and 7). We evaluated seven machine learning (ML) models for classifying the students' trials into two and three classes based on the progress of training (Beginning vs. End and Beginning vs. Middle vs. End). Additionally, we evaluated the same ML framework and a deep learning approach (LSTM) for predicting the remaining number of trials required to complete the training proficiently. A model-agnostic technique from the domain of explainable artificial intelligence (XAI) was also utilized to obtain interpretations of the employed black-box ML classifiers. For 2-class classification, the best model showed an accuracy of 97.1%, 96.9%, and 75.7% for Task 5, 6, and 7, respectively, whereas for 3-class classification, the corresponding accuracy was 96.3%, 95.9%, and 99.7%, respectively. The best regression algorithm was LSTM with a Mean Absolute Error of 4 (Task 5) and 3.6 trials (Tasks 6, 7). According to XAI, the kinematic parameters have a stronger impact on the classification decision than the goal-oriented metrics.

**Keywords:** minimally invasive surgery; machine learning; VR simulator; skills assessment



**Citation:** Prevezanou, K.; Seimenis, I.; Karaikos, P.; Pikoulis, E.; Lykoudis, P.M.; Loukas, C. Machine Learning Approaches for Evaluating the Progress of Surgical Training on a Virtual Reality Simulator. *Appl. Sci.* **2024**, *14*, 9677. <https://doi.org/10.3390/app14219677>

Academic Editor: Fausto Famà

Received: 4 September 2024

Revised: 12 October 2024

Accepted: 16 October 2024

Published: 23 October 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Minimally invasive surgery (MIS) has become popular for specific surgical procedures, particularly those targeting the abdomen, such as laparoscopic surgery. Performing such procedures requires greater dexterity due to the elongated instruments employed and the constrained working space. Moreover, surgeons face additional challenges such as a two-dimensional view of the operation field and minimized depth perception. Given these unique challenges, training programs and skills assessment methodologies for aspiring surgeons must accommodate the acquisition of these specialized skills while maintaining objectivity to ensure patient safety. In recent years, a variety of approaches have been employed in laparoscopy training. Initially, box trainers, which consisted of a box with apertures for trocar insertion to simulate the abdominal cavity, were utilized. Advancements led to the evolution of Physical Reality Surgical Simulators, where basic boxes were replaced with models of human torsos to provide a more realistic training environment [1,2].

To enhance the training experience further, Virtual Reality (VR) simulators were developed, integrating custom laparoscopic instruments, display monitors, and haptic feedback devices [3–5]. These devices marked a significant advancement in laparoscopy training by enabling the capture of specific metrics during training sessions using specialized hardware. Hence, an automated assessment can be formulated, allowing for personalized guidance for trainees throughout their training procedure. The most common approach for automated skills assessment involves the analysis of hand and tool motion data generated from the

simulator by employing machine learning (ML) algorithms [6,7]. An alternative approach for gathering kinematic data involves the use of external sensors mounted on the instruments or the surgeons' hands [8–10]. A new approach, emerging in recent years, involves deep learning models for analyzing kinematic data [11–14] or video recordings [15].

In the past, a few studies have applied ML for skills assessment using the performance metrics generated by the VR simulator as the input. For example, in neurosurgery training, Bissonnette et al. used the simulator's data to develop ML models for classifying skilled and novice trainees [16]. The support vector machine (SVM) algorithm exhibited the highest accuracy in this task. However, due to the experimental protocol employed, not all available instruments were utilized, resulting in an incomplete representation of psychomotor skills. Similarly, Siyar et al. constructed an ML model to classify experts vs. novices [17]. Again, SVM achieved the highest accuracy using 15 out of 150 performance metrics generated by the simulator. Winkler-Schwartz et al. created models to classify participants into four classes (experts, seniors, juniors, or students). Among the four ML algorithms employed, the k-Nearest Neighbor achieved the highest accuracy [18]. In another study, Topalli et al. utilized performance metrics generated by a computer-based endoscopic simulator to classify participants into novices and experts [19]. Among the used ML algorithms, the AdaBoostMI classifier showed superior performance for either all or selected metrics as the input.

In pursuit of optimal performance and a deeper understanding of surgical skill acquisition, recent efforts have implemented deep learning models. Mirchi et al. successfully trained an artificial neural network using input data from the Sim-Ortho VR simulator to classify participants into three distinct categories. Despite the promising results, the VR simulator allows only for one-handed maneuvers, thus disregarding important information about bimanual dexterity [20]. Moreover, Moglia et al. utilized output data from a robotic-assisted surgery VR simulator to develop an ensemble deep neural network (DNN) for predicting the number of attempts and training time required to attain proficiency. The ensemble DNN demonstrated superior performance compared to other ML algorithms, except in cases where Random Forest and Gradient-Boosted Regression Tree algorithms achieved comparable scores [21].

In the aforementioned studies, several limitations can be identified. When utilizing box trainers, external sensors must be attached to the instruments to collect kinematic data. However, these sensors are not always available and can restrict surgeons' hand movements. For VR simulators, access to the API kinematic acquisition module is essential to collect kinematic data, but this is not always provided by the owning company. Additionally, interpreting the recorded data by VR simulators poses challenges for both students and instructors, prompting the need to process performance metrics to extract meaningful features. Furthermore, VR simulators lack the ability to provide students with a clear understanding of their training progress, as they do not indicate the number of remaining trials needed to successfully complete a training course and advance to the next level. Moreover, another limitation relates to the ground-truth labeling method. The majority of skill assessment studies are based on self-proclaimed skills [22] and skill-levels defined by arbitrary thresholds on performance scores evaluated by experts [23]. Some other studies aim to classify junior vs. senior residents, but the exposure to laparoscopy training is not uniform to all the participants within each group [24].

In contrast to the previous studies where trainees are classified into different skill levels (e.g., novices, intermediates, and experts), in this work, we considered an alternative approach. The developed framework provides students with valuable information during their training on the VR simulator about their progress and the current skill level achieved. Similar studies have also highlighted the significance of intra-training feedback since it provides immediate information to the trainee about the skill level achieved as well as the improvements needed to successfully complete the training curriculum [25].

The developed framework involves three different experiments. The first experiment involved the use of a binary classification model to classify students at the Beginning vs.

End of laparoscopic training. The second experiment performed a three-class classification among students at the Beginning vs. Middle vs. End of training. Lastly, the third experiment involved the development of a regression model that aimed to predict the remaining number of trials required to complete training on a laparoscopic task at a proficient level. Using these models, our research aims to make a valuable contribution to the domain of surgical education and training. We aspire to offer a comprehensive framework for evaluating and guiding the progression of medical students while they acquire key laparoscopic skills. These models have the potential to improve training methodologies, provide personalized learning paths, and elevate the overall quality of surgical education.

## 2. Materials and Methods

### 2.1. VR Simulator and Data Acquisition

The input data originated from 23 medical students with no prior experience in laparoscopy or VR simulation. The students followed a structured VR simulation curriculum during their undergraduate training at the Surgical Simulation Center of Athens University Medical School (AKISA), University General Hospital “Attikon”, Greece. The employed VR simulator was the LAP Mentor<sup>TM</sup> laparoscopic simulator (Surgical Science, Göteborg, Sweden). The training tasks are integral components of the module ‘Laparoscopic Skills 1’, which comprises nine distinct tasks with each one focusing on a specific laparoscopic skill. The tasks are of increasing difficulty, starting from Task 1 where the student needs to locate ten balls using a camera, all through to Task 9 where the student needs to translocate objects using both hands. For this study, we identified three tasks indicative of the acquired surgical skills and with important educational value. Task 5 (*Clipping and Grasping*, Figure 1a) centers on the objective of securely holding and sealing leaking ducts within specific-colored segments, utilizing a pair of instruments. Task 6 (*Two-Handed Maneuvers*, Figure 1b) engages students in locating balls within a jelly-like substance and transferring them into a designated basket using a pair of instruments. Task 7 (*Cutting*, Figure 1c) requires safely cutting a circular shape without causing any damage, with the aid of two instruments (scissors and a grasper).

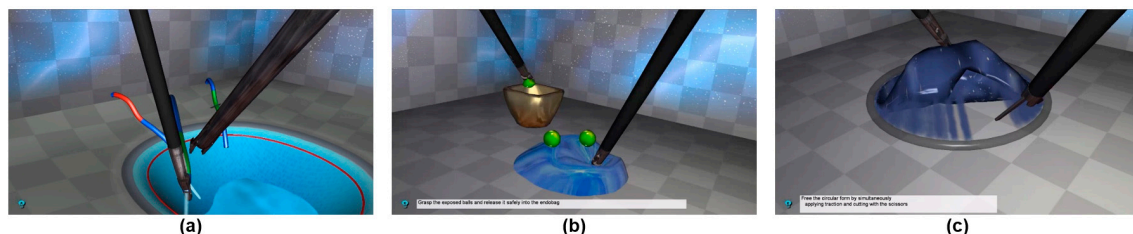


**Figure 1.** The three VR training tasks evaluated in this study: (a) Task 5 (Clipping and Grasping), (b) Task 6 (Two-Handed Maneuvers), and (c) Task 7 (Cutting).

Two important studies have examined the construct validity of the tasks included in the simulator’s psychomotor skills module (‘Laparoscopic Skills 1’) [26,27]. Task 5, followed by Tasks 6 and 7, has the most confirmed construct validity on more than one parameter. The postulated reason for this is the complexity of the tasks, which is translated into a greater number of maneuvers. Contrary to binary metrics (e.g., task completion) and scale parameters with a small range of values (e.g., number of dropped objects), other performance metrics such as the number of movements and number of cutting maneuvers can follow the progress of skill acquisition more reliably, as depicted in the learning curves published in [26,27]. Moreover, metrics with higher and more varying values comprise a favorable substrate for computational analysis, including AI algorithms.

The students’ progress through the module adheres to a structured curriculum designed by the surgical education board of the simulation center based on the training guidelines proposed in [26]. Specifically, advancement through the tasks is contingent upon the *successful completion* of the previous task (starting from Task 1 and continuing

sequentially up to Task 9). Failure to complete a task requires the student to restart the module from Task 1. This process is reiterated until the student successfully completes all module tasks by performing two consecutive successful trials in every task. Hence, each student concludes the module at their individual pace, with the number of attempts not being constrained.

It should be noted that the successful completion of a particular task is signified by several task-specific educational goals imposed by the educational curriculum. Specifically, the VR simulator provides an assessment report with a large number of task-specific performance metrics (Supplementary Materials, Table S1). In accordance with the educational objectives, the board has chosen a limited set of metrics for each task and established thresholds for each one. These thresholds, detailed in Table 1, delineate the criteria that a student must satisfy to consider a task as ‘*successfully completed*’.

**Table 1.** Thresholds for performance metrics that determine the *successful completion* of a training trial.

| Task | Metric                                     | Value     |
|------|--------------------------------------------|-----------|
| 5    | Total number of movements                  | $\leq 70$ |
|      | Number of clipped ducts                    | $\geq 9$  |
|      | Number of lost clips                       | $< 1$     |
| 6    | Total number of movements                  | $\leq 50$ |
|      | Number of lost balls which miss the basket | $< 1$     |
| 7    | Total Number of cutting maneuvers          | $\leq 20$ |

## 2.2. Experiments

Based on the performance metrics provided by the VR simulator, various datasets were derived to address the goals of the three experiments. Table 2 provides a statistical overview of the training trials performed by the participants. Firstly, we developed a binary classification model to differentiate the trials that are close to the Beginning (*B*) or the End (*E*) of the training progress. Hence, this dataset included 3 trials randomly selected from the Beginning and another 3 trials from the End of training of each participant ( $N_1 = 138$  trials in total). The borderline between the two classes was different for each subject and it was defined by the median (‘middle’) training attempt, splitting the student’s trials into two equal subsets.

**Table 2.** Ground-truth data statistics. Number (#) of trials performed by the subjects for each surgical task (SD: standard deviation).

| Metric (# Trials) | Task 5      | Task 6      | Task 7     |
|-------------------|-------------|-------------|------------|
| Minimum           | 9           | 14          | 6          |
| Maximum           | 52          | 57          | 34         |
| Median            | 23          | 23          | 16         |
| Mean $\pm$ SD     | $25 \pm 11$ | $26 \pm 10$ | $17 \pm 7$ |

The second experiment aimed to classify student trials into three classes: the Beginning (*B*), Middle (*M*), and End (*E*) of training. For this purpose, the training trials of each subject were first split into three equal parts. Then, each of the three classes included 3 random trials from each part, so the total number of trials for this dataset was  $N_2 = 207$ .

For the two classification experiments, the datasets were balanced across classes, and they are representative of the individualized training phases (classes) of a trainee on a particular surgical task. Moreover, in the context of surgical training if, for example, some student wants to start training on a particular task, the prior probability of being classified into one of the two/three training classes is essentially equal. In this context, our ML framework will be able to provide feedback to the trainee about his/her expertise level via the inferred training phase (class).

The third experiment focused on online regression to highlight the assessment process from a different perspective. In particular, the goal here was to provide a prediction for the remaining number of trials (RNT) required by a subject to *successfully* complete a training task. The prediction was generated throughout the entire training process, starting from the very first trial of each subject. For this experiment, we utilized metrics from all available trials separately for each task ( $N_{\text{Task5}} = 437$ ,  $N_{\text{Task6}} = 503$ ,  $N_{\text{Task7}} = 313$ ).

### 2.3. Dimensionality Reduction

The VR simulator generates numerous performance metrics (Task 5: 26, Task 6: 18, and Task 7: 16) with different scaling ranges. To improve the training process, the collected data were subjected to z-normalization, a statistical technique that centers the data around a mean of zero and scales it to have a standard deviation of one. Furthermore, several performance metrics present significant correlations. Consequently, we employed two different techniques for dimensionality reduction. The first method was Principal Component Analysis (PCA), which transforms the data into a new coordinate system linearly. The PCA method provided feature vectors with 6–7 dimensions (depending on the surgical task and training fold) that accounted for >95% of the total variance.

The second approach involved feature selection based on Fisher’s score. Using the Fisher criterion, the features are ranked in descending order according to how well they can differentiate between different classes. In this study, we defined those that were present within the top six features as ‘best features’ (according to Fisher’s score), for at least half of the training folds for the first experiment (2-class classification). This process provided about 6–7 ideal features (depending on the surgical task), as shown in Table 3.

**Table 3.** ‘Best features’ selected via the Fisher’s score. ‘✓’ denotes whether the metric is included in the corresponding task.

| Description                                                  | Abbreviation | Task 5 | Task 6 | Task 7 |
|--------------------------------------------------------------|--------------|--------|--------|--------|
| Accuracy rate—cuts without injury (%)                        | CutsNoInj    |        |        | ✓      |
| Average speed of right instrument movement (cm/s)            | Speed-R      | ✓      | ✓      |        |
| Average speed of left instrument movement (cm/s)             | Speed-L      |        |        | ✓      |
| Economy of movement—grasper (%)                              | EOM-G        | ✓      |        |        |
| Economy of movement—left instrument (%)                      | EOM-L        |        | ✓      |        |
| Ideal path length of clipper (cm)                            | IPL-C        | ✓      |        |        |
| Ideal path length of grasper (cm)                            | IPL-G        | ✓      |        |        |
| Ideal path length of left instrument (cm)                    | IPL-L        |        | ✓      |        |
| Number of exposed green balls that are collected             | #GreenB      |        | ✓      |        |
| Number of clipped ducts                                      | #ClipD       | ✓      |        |        |
| Number of movements of left instrument                       | #Move-L      | ✓      |        | ✓      |
| Number of movements of right instrument                      | #Move-R      |        | ✓      |        |
| Number of lost balls which miss the basket                   | #LostB       |        | ✓      |        |
| Number of cutting maneuvers performed without causing injury | #CuttM-NoInj |        |        | ✓      |
| Relevant path length—clipper (cm)                            | RPL-C        | ✓      |        |        |
| Safe retraction—overstretch (%)                              | SafeRetr     |        |        | ✓      |
| Trial number                                                 | TN           |        |        | ✓      |

### 2.4. Algorithms

We employed various ML algorithms that have been applied for similar applications in the past, such as skill classification based on kinematic data [28] and regression tasks in the clinical surgery domain [29,30]. For the two classification experiments we employed the k Nearest Neighbor (kNN), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (GNB), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) methods.

For the regression experiment, we employed the following regression algorithms: Support Vector Regressor (SVR) with three kernels (Linear, RBF, and Polynomial), Linear Regression (LinR), Second-degree Polynomial Regression (PolR), Random Forest (RF),

kNN regression (kNNR), and XGBoost (XGB). In addition, we applied a deep learning approach based on the LSTM network, which has shown promising performance in similar predictions tasks in surgical research [31]. The input (feature vectors) and output (RNT) were the same as that of the ML models and the RNT predictions were also generated in ‘real time’ (i.e., without using information from forthcoming trials). A significant advantage of LSTM is that it can handle long-term dependencies by removing or adding information to its internal cell state as the information flows from one trial to the next one. In this study, we employed an LSTM network with 256 units and a dense layer with one node and linear activation for RNT prediction. The dense layer’s weights are shared across the multiple trial steps, whereas at each trial, the LSTM output was forwarded to the dense layer for generating the current prediction.

For kNN, we used the Euclidian distance with  $k = 1$  (first nearest neighbor). For RF, the number of estimators was set to one hundred while leaving the decision tree’s maximum depth unspecified. For SVM, we employed two kernels for the classification task (linear and radial basis function (RBF)) and three kernels for the regression task (linear, RBF, and polynomial), leading essentially to two and three SVM variations, respectively. The kernel parameter gamma was set to  $1/(\# \text{ features})$  and the penalty  $C = 1$  in all runs. For XGB, the maximum depth was set at 6 and the learning rate at 0.3. The LSTM network was trained for 30 epochs with batch size 1, learning rate = 0.001, Adam optimization, and the Huber loss function. Due to the limited size of the dataset, we did not employ an extra validation set for parameter fine tuning. Instead, the parameters of the models were set according to those proposed in similar studies, and when not available, to the default ones in the employed scikit-learn library.

For the regression experiment, we also evaluated a naive approach based on the historical data of the trials performed by all subjects. For a particular surgical task, the remaining number of trials (RNT) required to successfully complete training by some subject ( $i$ ) is computed based on the difference between the current training trial ( $t_i$ ) and the median of the trials performed by all subjects in this task, as follows:

$$RNT_i = \max(0, medT - t_i), \quad (1)$$

where  $medT$  denotes the median of the trials performed by all subjects in a particular task and the  $\max(.,.)$  operator simply ensures that  $RNT_i$  is always nonnegative.

All models were implemented in Python 3.8 using the scikit-learn library (ML algorithms) and TensorFlow platform (LSTM model).

### 2.5. Evaluation Protocol

To assess the effectiveness of the models developed for each experiment, cross-validation techniques were implemented. In the case of the first two experiments (classification), twenty folds of stratified random sampling were performed. This involved data segregation into subgroups based on the ground truth labels, data shuffling and splitting into two sets, 75% for training and 25% for testing. For the third experiment (regression), random subsampling was utilized to split the data into two datasets twenty times (train and test), maintaining the same proportions as the classification experiments.

In addition to cross-validation, specific evaluation metrics were chosen for each experiment to quantify the models’ performance. For the classification experiments, a confusion matrix was computed for each model providing a representation of the algorithm’s predictions compared to the ground truth. In the 2-class classification, the positive class denotes the Beginning of training, and the negative class represents the End of training. In addition to the confusion matrix, the algorithms’ performance was evaluated using the following metrics: accuracy, specificity, and sensitivity.

For the third experiment, we used the Mean Absolute Error (MAE), which calculates, for each subject, the mean of the absolute differences between the model predictions and the corresponding target values (Equation (2)).

$$MAE = \frac{1}{n} \sum_i |y_i - \hat{y}_i|, \quad (2)$$

where  $y_i$  denotes the predicted RNT after a specific training session  $i$ ,  $\hat{y}_i$  denotes the ground truth RNT, and  $n$  is the subject's total number of attempts during training.

Figure 2 provides a visual overview of the overall experimental design followed in this study.

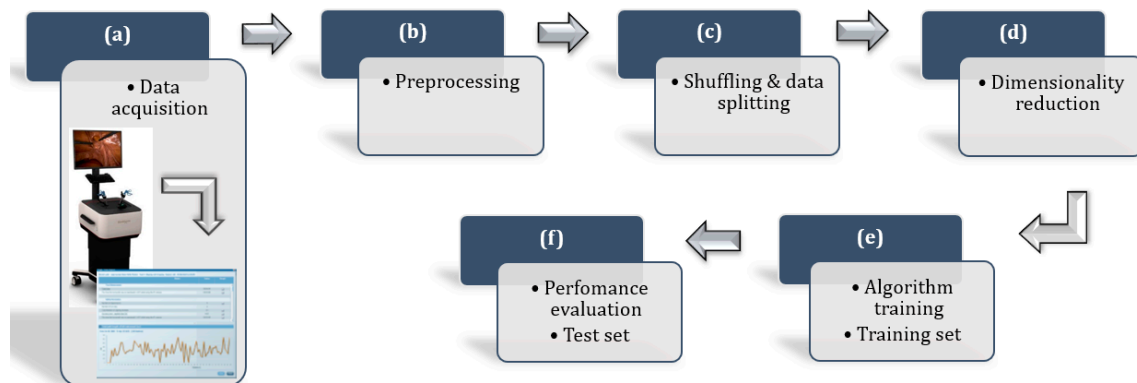


**Figure 2.** Model design process: (a) data acquisition from the VR simulator, (b) z-normalization data scaling, (c) division of the main dataset for algorithm training and testing, (d) application of dimensionality reduction techniques, (e) model training using the training set, and (f) performance evaluation based on metrics using the test set.

### 3. Results

#### 3.1. Classification Experiments

Table 4 shows the results for the 2-class classification based on PCA for the three surgical tasks. Overall, the accuracy for Tasks 5 and 6 is higher than for Task 7. The specificity, representing the true negative rate, is superior to the sensitivity across all tasks, denoting better classification for the End of training class. When analyzing each task separately, the performance differences across the algorithms are minor, but certain algorithms present better performance. Specifically, in Task 5, SVM (Linear) is the best-performing algorithm with an accuracy of 97.1%, while LDA shows the lowest accuracy (89.2%). Additionally, QDA achieves the highest sensitivity (98.1%), whereas LDA exhibits superior specificity (99.2%). In Task 6, SVM (Linear) maintains the highest accuracy (96.9%) while GNB scores the lowest accuracy (92.5%). SVM (Linear) demonstrates the highest specificity (98.9%), while RF and QDA excel in sensitivity (95.3%). In Task 7, lower performance scores are presented across all models. LDA scored the highest accuracy (75.7%) while LR performed the poorest (63.6%). In terms of specificity, LDA outperforms the other algorithms (87.8%), whereas QDA achieves the highest sensitivity (81.1%).

Table 5 presents the results for the 2-class classification experiment, utilizing Fisher's score, which seem aligned with those based on PCA, but with slightly lower values. Accuracy is higher for Tasks 5 and 6 compared to Task 7. Furthermore, sensitivity percentages are consistently higher than specificity across all tasks, indicating superior classification ability for the Beginning of the training class. For Task 5, the best performing algorithm is LR (93.6%) while QDA has the lowest performance (83.6%). Similarly, for Task 7, LR achieves the highest accuracy (72.1%), while QDA the lowest one (52.9%). Lastly, SVM (RBF) stands out as the best performing algorithm in Task 6, achieving an accuracy of 95.3%.

**Table 4.** Mean performance comparison for 2-class classification based on PCA (Tasks 5, 6, and 7). Best row-wise results are in bold.

| Task | Evaluation Metric | kNN  | LDA         | QDA         | LR   | GNB  | RF          | SVM (Linear) | SVM (RBF) |
|------|-------------------|------|-------------|-------------|------|------|-------------|--------------|-----------|
| 5    | Accuracy          | 95.1 | 89.2        | 94.3        | 94.4 | 93.2 | 93.1        | <b>97.1</b>  | 94.7      |
|      | Specificity       | 95.0 | <b>99.2</b> | 90.6        | 95.6 | 95.3 | 94.6        | 98.6         | 98.1      |
|      | Sensitivity       | 95.3 | 79.2        | <b>98.1</b> | 93.3 | 91.1 | 92.2        | 95.6         | 91.4      |
| 6    | Accuracy          | 92.6 | 94.6        | 95.1        | 95.4 | 92.5 | 95.6        | <b>96.9</b>  | 95.0      |
|      | Specificity       | 91.4 | 97.8        | 95.0        | 97.2 | 92.5 | 95.8        | <b>98.9</b>  | 96.4      |
|      | Sensitivity       | 93.9 | 91.4        | <b>95.3</b> | 93.6 | 92.5 | <b>95.3</b> | 95.0         | 93.6      |
| 7    | Accuracy          | 67.9 | <b>75.7</b> | 73.3        | 63.6 | 74.7 | 73.6        | 73.3         | 73.5      |
|      | Specificity       | 75.0 | <b>87.8</b> | 46.1        | 81.4 | 85.8 | 84.4        | 76.7         | 79.7      |
|      | Sensitivity       | 60.8 | 63.6        | <b>81.1</b> | 68.1 | 60.8 | 62.8        | 70.0         | 67.2      |

**Table 5.** Mean performance comparison for 2-class classification based on Fisher’s score (Tasks 5, 6, and 7). Best results row-wise are shown in bold.

| Task | Evaluation Metric | kNN  | LDA         | QDA         | LR          | GNB  | RF          | SVM (Linear) | SVM (RBF)   |
|------|-------------------|------|-------------|-------------|-------------|------|-------------|--------------|-------------|
| 5    | Accuracy          | 90.7 | 91.4        | 83.2        | <b>93.6</b> | 91.9 | 92.6        | 92.8         | 92.8        |
|      | Specificity       | 89.7 | <b>95.3</b> | 70.8        | 91.7        | 93.1 | 93.6        | 91.1         | <b>95.3</b> |
|      | Sensitivity       | 91.7 | 87.5        | <b>95.6</b> | <b>95.6</b> | 90.8 | 91.7        | 94.4         | 90.3        |
| 6    | Accuracy          | 93.9 | 92.8        | 82.2        | 95.1        | 91.8 | 94.4        | 94.6         | <b>95.3</b> |
|      | Specificity       | 92.5 | 93.1        | 84.2        | 94.2        | 90.0 | 92.8        | 92.8         | <b>95.0</b> |
|      | Sensitivity       | 95.3 | 92.5        | 80.0        | <b>96.1</b> | 93.6 | <b>96.1</b> | 96.4         | 95.6        |
| 7    | Accuracy          | 63.3 | 70.8        | 52.9        | <b>72.1</b> | 55.0 | 68.5        | 71.9         | 67.9        |
|      | Specificity       | 66.7 | <b>82.2</b> | 8.9         | 79.4        | 19.4 | 73.6        | 81.1         | 80.3        |
|      | Sensitivity       | 60.0 | 59.4        | <b>96.9</b> | 64.7        | 90.6 | 63.3        | 63.1         | 55.6        |

Figure 3 presents the normalized confusion matrices for the most effective algorithm, based on PCA, for the 2-class classification experiment. The matrices, which were generated by summing the raw confusion matrices from the twenty folds, reveal better classification performance for the End of the training class across all three tasks. The confusion matrices of all other algorithms can be found in the Supplementary Materials, Figure S1.

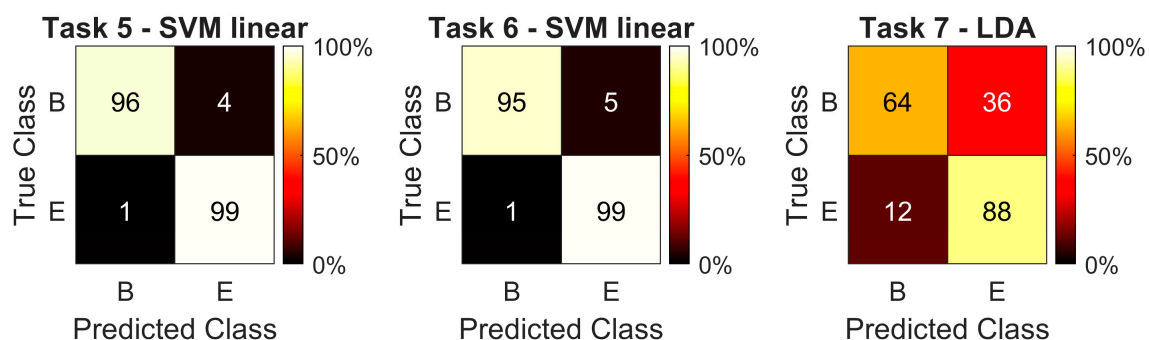


**Figure 3.** Color-coded confusion matrices (%) for the best-performing algorithm for the 2-class classification (Tasks 5, 6, and 7). B and E denote the ‘Beginning of training’ and ‘End of training’, respectively.

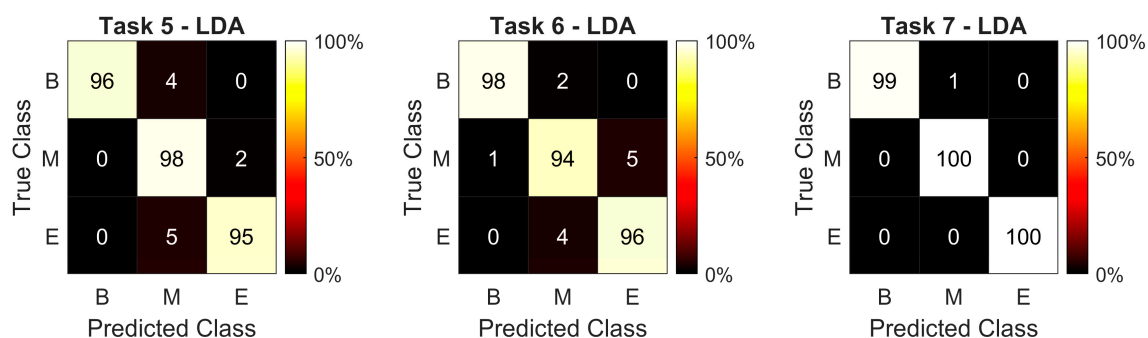
The results for the 3-class classification (Table 6) were produced following a specific experimental protocol. Initially, the feature dimensionality was reduced using Fisher’s score, resulting in 6–7 final features including the feature *Trial number*. Further reduction was performed based on PCA, resulting in 4–6 principal components.

**Table 6.** Mean performance comparison for 3-class classification (Tasks 5, 6, and 7). Best results row-wise are shown in bold.

| Task | Evaluation Metric | kNN  | LDA         | QDA  | LR   | GNB         | RF   | SVM (Linear) | SVM (RBF) |
|------|-------------------|------|-------------|------|------|-------------|------|--------------|-----------|
| 5    | Accuracy          | 89.9 | <b>96.3</b> | 94.8 | 94.8 | 94.0        | 85.7 | 95.7         | 90.8      |
|      | Specificity       | 95.0 | <b>98.2</b> | 97.3 | 97.3 | 97.0        | 93.0 | 97.8         | 95.5      |
|      | Sensitivity       | 90.0 | <b>96.3</b> | 94.7 | 94.7 | 94.0        | 85.9 | 95.7         | 91.0      |
| 6    | Accuracy          | 90.9 | <b>95.9</b> | 91.7 | 91.7 | 94.4        | 86.5 | 94.4         | 91.7      |
|      | Specificity       | 95.3 | 82.7        | 96.0 | 96.0 | <b>97.3</b> | 93.2 | 97.2         | 96.2      |
|      | Sensitivity       | 90.7 | <b>96.0</b> | 92.0 | 92.0 | 94.6        | 86.4 | 94.3         | 92.0      |
| 7    | Accuracy          | 81.0 | <b>99.7</b> | 78.6 | 78.6 | 98.7        | 82.5 | 98.8         | 93.6      |
|      | Specificity       | 90.4 | <b>99.8</b> | 89.3 | 89.3 | 99.3        | 91.1 | 99.3         | 96.8      |
|      | Sensitivity       | 80.7 | <b>99.7</b> | 78.6 | 78.6 | 98.7        | 82.1 | 98.7         | 93.7      |

Across all three tasks, accuracy consistently exceeds 80%, with the top-performing algorithms achieving at least 96%. Once again, specificity outweighs sensitivity. Specifically, for Tasks 5, 6, and 7, the best-performing algorithm is LDA, achieving accuracies of 96.3%, 95.9%, and 99.7%, respectively. LDA also shows the highest sensitivity in all three tasks (96.3%, 96.0%, and 99.7%, respectively), and for Tasks 5 and 7, LDA achieves the highest specificity (98.2% and 99.8%, respectively). In Task 6, the highest specificity is presented by GNB (97.3%).

Figure 4 presents the normalized confusion matrices for the best-performing algorithm for the 3-class classification experiment. Task 5 demonstrates superior classification performance for the Middle of training class, while both the Middle of training and End of training classes are perfectly classified in Task 7. In contrast, the Beginning of training class shows better classification performance in Task 6. The confusion matrices of all of the other algorithms can be found in Supplementary Materials, Figure S2.

**Figure 4.** Color-coded confusion matrices (%) for the best performing algorithm for the 3-class classification (Tasks 5, 6, and 7); B: ‘Beginning of training’, M: ‘Middle of training’ and E: ‘End of training’.

### 3.2. Regression Experiment

Table 7 shows the results for the regression experiments. For each technique, the MAE is presented as mean  $\pm$  SD (standard deviation) over the 20 test-set folds. In addition, the table includes the MAE for *fast* learners (*short* training sessions) and *slow* learners (*long* training sessions). The distinction was based on the total number of trials of the trainee ( $NT_i$ ) with respect to the median number of trials (mdNT): for *fast* learners  $NT_i < \text{mdNT}$  and for *slow* learners  $NT_i \geq \text{mdNT}$  (mdNT = 23, 23, and 16 for Task 5, 6, and 7, respectively).

**Table 7.** RNT prediction results (mean  $\pm$  SD MAE) for the three surgical tasks and for all, short, and long training sessions. The best MAE is shown in bold.

| Task | Algorithm      | MAE (All)                       | MAE (Short)                     | MAE (Long)                      |
|------|----------------|---------------------------------|---------------------------------|---------------------------------|
| 5    | SVR (Linear)   | 5.7 $\pm$ 1.1                   | 6.6 $\pm$ 2.7                   | 4.9 $\pm$ 1.2                   |
|      | SVR (RBF)      | 5.4 $\pm$ 0.9                   | 5.6 $\pm$ 2.1                   | 5.3 $\pm$ 1.2                   |
|      | SVR (Poly)     | 5.8 $\pm$ 0.6                   | 5.2 $\pm$ 1.1                   | 6.5 $\pm$ 1.3                   |
|      | kNNR           | 6.6 $\pm$ 1.2                   | 6.7 $\pm$ 2.1                   | 6.4 $\pm$ 0.9                   |
|      | PolR           | 9.0 $\pm$ 5.0                   | 9.2 $\pm$ 4.4                   | 8.6 $\pm$ 6.0                   |
|      | LinR           | 5.8 $\pm$ 1.3                   | 7.2 $\pm$ 3.0                   | <b>4.7 <math>\pm</math> 1.1</b> |
|      | RF             | 5.5 $\pm$ 1.3                   | 6.3 $\pm$ 2.7                   | 4.8 $\pm$ 0.9                   |
|      | XGB            | 5.7 $\pm$ 1.2                   | 6.7 $\pm$ 2.6                   | 4.8 $\pm$ 1.0                   |
|      | LSTM           | <b>4.0 <math>\pm</math> 1.1</b> | <b>2.6 <math>\pm</math> 0.9</b> | 4.8 $\pm$ 1.7                   |
|      | Naïve approach | 7.0 $\pm$ 1.1                   | 8.3 $\pm$ 2.8                   | 6.0 $\pm$ 2.6                   |
| 6    | SVR (Linear)   | 4.3 $\pm$ 0.8                   | 4.8 $\pm$ 1.6                   | <b>3.8 <math>\pm</math> 1.0</b> |
|      | SVR (RBF)      | 4.3 $\pm$ 0.7                   | 4.2 $\pm$ 1.3                   | 4.5 $\pm$ 1.1                   |
|      | SVR (Poly)     | 6.1 $\pm$ 0.9                   | 5.4 $\pm$ 0.9                   | 6.7 $\pm$ 1.3                   |
|      | kNNR           | 5.4 $\pm$ 0.6                   | 5.4 $\pm$ 1.1                   | 5.2 $\pm$ 0.7                   |
|      | PolR           | 4.9 $\pm$ 1.0                   | 5.5 $\pm$ 2.1                   | 4.3 $\pm$ 1.1                   |
|      | LinR           | 4.3 $\pm$ 0.8                   | 4.9 $\pm$ 1.6                   | <b>3.8 <math>\pm</math> 1.0</b> |
|      | RF             | 4.3 $\pm$ 0.6                   | 4.6 $\pm$ 1.2                   | 4.1 $\pm$ 0.9                   |
|      | XGB            | 4.7 $\pm$ 0.6                   | 4.9 $\pm$ 1.4                   | 4.1 $\pm$ 0.9                   |
|      | LSTM           | <b>3.6 <math>\pm</math> 0.8</b> | <b>2.4 <math>\pm</math> 1.0</b> | 4.6 $\pm$ 1.4                   |
|      | Naïve approach | 5.5 $\pm$ 0.9                   | 5.1 $\pm$ 1.6                   | 5.9 $\pm$ 1.8                   |
| 7    | SVR (Linear)   | 4.0 $\pm$ 0.8                   | 4.6 $\pm$ 1.4                   | 3.4 $\pm$ 0.6                   |
|      | SVR (RBF)      | <b>3.6 <math>\pm</math> 0.6</b> | 3.9 $\pm$ 1.1                   | <b>3.2 <math>\pm</math> 0.5</b> |
|      | SVR (Poly)     | 4.1 $\pm$ 0.8                   | 3.7 $\pm$ 0.8                   | 4.4 $\pm$ 1.5                   |
|      | kNNR           | 4.7 $\pm$ 0.6                   | 4.8 $\pm$ 0.9                   | 4.5 $\pm$ 0.5                   |
|      | PolR           | 4.5 $\pm$ 1.1                   | 5.0 $\pm$ 1.7                   | 4.0 $\pm$ 1.2                   |
|      | LinR           | 3.9 $\pm$ 0.8                   | 4.6 $\pm$ 1.5                   | 3.3 $\pm$ 0.6                   |
|      | RF             | 3.8 $\pm$ 0.7                   | 4.3 $\pm$ 1.3                   | 3.3 $\pm$ 0.5                   |
|      | XGB            | 4.1 $\pm$ 0.7                   | 4.8 $\pm$ 1.3                   | 3.5 $\pm$ 0.6                   |
|      | LSTM           | <b>3.6 <math>\pm</math> 0.7</b> | <b>2.8 <math>\pm</math> 1.6</b> | 4.6 $\pm$ 1.6                   |
|      | Naïve approach | 4.1 $\pm$ 1.0                   | 5.0 $\pm$ 1.8                   | <b>3.2 <math>\pm</math> 1.1</b> |

Regarding the results for *all* training sessions, most of the ML algorithms show a MAE value close to 5.4–5.8 for Task 5, 4.3–4.9 for Task 6, and 3.6–4.1 for Task 7. SVR shows the best performance among the ML techniques with MAE = 5.4 (Task 5), 4.3 (Task 6), and 3.6 (Task 7). The naïve approach, which is based only on historical data, seems to have the lowest performance for Task 5 (MAE = 7) and Task 6 (MAE = 5.5), whereas for Task 7, it is almost comparable to the ML algorithms (MAE = 4.1). The LSTM model presents the best performance against all other techniques with MAE = 4.0 (Task 5), 3.6 (Task 6), and 3.6 (Task 7). These values correspond to error percentages of 17.4% for Task 5, 15.7% for Task 6, and 22.5% for Task 7, which are calculated based on the median number of trials for each task as presented in Table 2.

For *short* training sessions, all ML algorithms exhibit an MAE similar to the MAE for *all* training sessions (MAE = 5.2–6.3 for Task 5, MAE = 4.2–4.9 for Task 6, and MAE = 3.7–4.8 for Task 7). Again, SVR is the best performing ML algorithm with MAE = 5.2 (Task 5), 4.2 (Task 6), and 3.7 (Task 7). In contrast to *all* training sessions, the naïve approach shows the lowest performance for Task 7 (MAE = 5.0), while for Task 5 (MAE = 8.3) and 6 (MAE = 5.1), the performance is comparable to some ML algorithms. The LSTM model outperforms all other algorithms with MAE = 2.6 (Task 5), 3.6 (Task 6), and 3.6 (Task 7) corresponding to error percentages of 11.3% for Task 5, 15.7% for Task 6, and 22.5% for Task 7.

For *long* training sessions, most ML algorithms achieve MAE values close to 4.7–5.3 (Task 5), 3.8–4.6 (Task 6), and 3.2–4.0 (Task 7). The best performing algorithms for Tasks 5, 6, and 7 are LinR (MAE = 4.7; 20.4% error), LinR/SVR (MAE = 3.8; 16.5% error), and SVR/Naïve (MAE = 3.2; 20.0% error), respectively. For Task 5, the LSTM model is slightly

worse than LinR, whereas for Tasks 6 and 7, it is significantly worse than the best ML models. However, as described previously, LSTM shows the best overall performance in all three surgical Tasks, irrespective of the number of training trials.

### 3.3. Model Explainability

Given the superior performance of the ML models in the classification tasks, we moved a step further toward explaining the overall models' behavior and prediction outcomes via explainable artificial intelligence (XAI). XAI refers to techniques that provide human users with the ability to comprehend and derive rational decisions about the overall performance of an AI model, the intrinsic mechanisms of which are very hard to explain (i.e., black box), as well as on individual predictions. The ultimate outcome of XAI techniques is commonly in the form of graphs due to their easily human-readable format and straightforward interpretation.

For the sake of simplicity and easy interpretation of the XAI outcome, we focused on explaining the best performing model for each surgical task (LR for Tasks 5, 7, and SVM for Task 6) regarding the binary classification scheme (i.e., Beginning vs. End of training) using only the 'best features' (based on Fisher's score).

In this study, the SHapley Additive exPlanations (SHAP) was employed [32], which is a model-agnostic XAI method based on the game theoretically optimal Shapley values [33]. The goal of SHAP is to explain the prediction of an instance by computing the contribution of each feature to the prediction. Features with large absolute Shapley values are important. Although the technique provides a Shapley value  $\varphi^{(i)}$  for any instance  $x^{(i)}$ , one can obtain the global importance of each feature by averaging the absolute Shapley values, as follows:

$$I_j = \frac{\sum_n |\varphi_j^{(i)}|}{n}, \quad (3)$$

where  $I_j$  essentially denotes the average impact on model output magnitude with respect to feature  $j$  (i.e., *feature importance*) and  $n$  is the total number of feature values (i.e., instances) on the dataset evaluated.

Figure 5a–c shows a bar plot for each task where the best 6–7 features are ranked in descending order according to their impact on model decisions (i.e., global feature importance ranking). The feature names presented in Figure 5 are abbreviated, while full feature names can be found in Table 3. For Task 5, the top 3 features are Economy of movement—grasper, Number of movements of left instrument, and Relevant path length—clipper. For Task 6, the top 2 features are Economy of movement—left instrument and Number of movements of right instrument. For Task 7, the top feature is Trial Number. It is obvious that as trainees gain more experience, the variation in the performance metric values decreases. As a result, fewer features can differentiate between classes and significantly influence the classification process (i.e., Task 7: only one feature has substantial impact).

Figure 5d–f presents a summary plot for each task, combining both the feature importance and the effects of each feature across the entire dataset. The  $x$ -axis represents the Shapley values of each feature ( $y$ -axis) along with a color bar where each dot is colored based on its value (high or low). A Shapley value less than 0, equal to 0, or greater than 0 corresponds to a negative contribution, no contribution, and positive contribution, respectively. For Task 5, a trial will most likely be classified as End of training when the trainee uses both the clipper and the grasper close to the ideal (shortest) path length (high values for EOM-G and low values for RPL-C) and when the left instrument is only moved when necessary (low values for #Move-L). In Task 6, the same kinematic patterns are followed and therefore a trial is classified as End of training when the trainee moves the left instrument along the shortest distance (high values for EOM-L) and the right instrument is moved only a handful of times (low values for #Move-R). For Task 7, high values for TN

correspond to the End of training while other features do not present a clear separation between the two training stages.

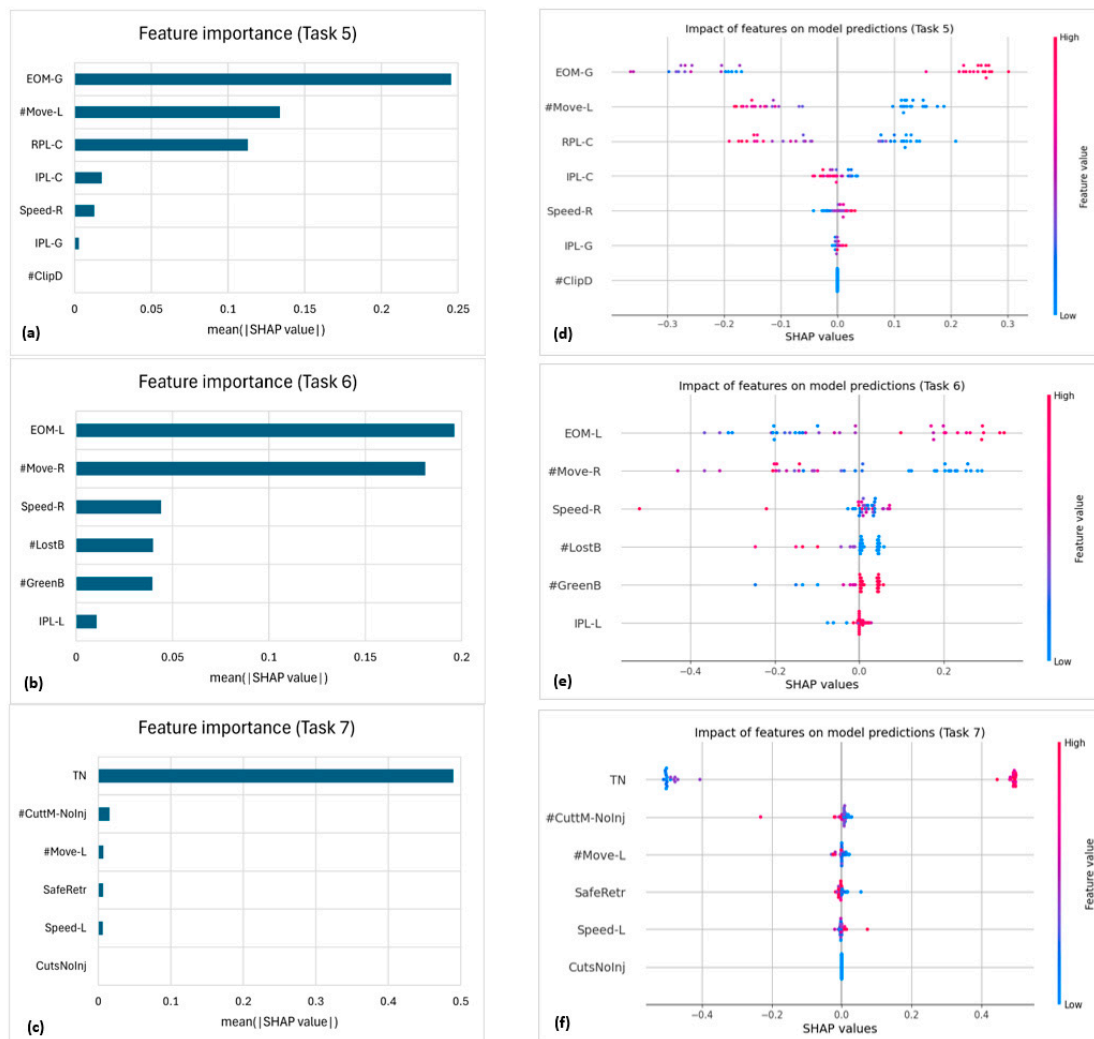


**Figure 5.** Bar plots of the average feature importance of the developed binary classifiers for (a) Task 5, (b) Task 6, and (c) Task 7. Feature impact scatter plots for (d) Task 5, (e) Task 6, and (f) Task 7.

#### 4. Discussion

In this study, we propose three distinct experiments to evaluate the progress of surgical skill acquisition by students who train on a laparoscopic VR simulator. For the first two experiments, we focus on the classification of the trainees' trials at different stages of surgical training: a 2-class classification (Beginning vs. End) and a 3-class classification (Beginning vs. Middle vs. End). Moreover, a regression model was evaluated to predict the remaining number of trials a student needs to *successfully* complete a training task.

Most of the ML algorithms used in the 2-class classification experiment presented accuracy percentages close to 80–98% (based on PCA) and 70–92% (based on Fisher's score), showing a high-quality classification performance. The best-performing algorithm is SVM with the highest achieved accuracy of 97.1% (Task 5, based on PCA). After examining the results, it is evident that Task 7 presents the lowest accuracy, mainly due to the students' familiarity and extensive practice with the VR simulator across the previous six tasks. By the time the trainees reach Task 7, they have successfully completed Tasks 1 through 6, hence acquiring more skills that are essential for the successful completion of Task 7 (such as eye-hand coordination, depth perception, bimanual skills, laparoscopic orientation, and instrument manipulation). Consequently, there is a small difference between the

performance metrics in the Beginning vs. the End of their training on this task. Moreover, when comparing results between the two-dimensionality reduction methods, PCA provides better results than Fisher's score. This was expected due to the way that these techniques operate: PCA combines features while Fisher's score ranks them before selection, resulting in a greater loss of variance. Moreover, the performance at the End of the training class is better probably due to the similar performance metrics achieved when trainees approach the successful completion of the task.

For 3-class classification, LDA emerges as the top-performing algorithm (99.7%, Task 7), excelling not only in accuracy but also in specificity and sensitivity across all three tasks. Contrary to the trends observed in the 2-class classification, Task 7 demonstrates the highest scores. This can be attributed to the inclusion of the 'Trial number' feature, which was also highlighted by the Fisher's score metric.

Regarding the regression experiment, the SVR variants outperformed the other ML models. However, the LSTM model outperformed all ML models in Tasks 5 (MAE = 4) and 6 (MAE = 3.6) while in Task 7, its performance was similar to SVR (MAE = 3.6). These findings denote a promising performance that is also similar to that in other prediction tasks, such as the prediction of remaining duration in surgical operations [31]. In contrast, the naïve approach performed poorly in Tasks 5 and 6 except for Task 7 where its performance was almost comparable to that of the ML algorithms (MAE = 4.1). When comparing the results for *all*, *short*, and *long* training sessions, the performance of the ML algorithms is similar in all cases, and the same applies to the naïve approach. In contrast, the LSTM model shows a clearly superior performance in shorter training performances (*fast* learners) compared to the longer ones (*slow* learners). This is due to its ability to better capture the variation in the trainee's performance metrics for a small number of trials whereas, for each ML model, the input is based solely on the current trial of the trainee.

Besides the evaluation of models' performance, we leveraged techniques from the XAI domain to obtain a better understanding of the model prediction output and further interpretation of the key features involved in surgical skill acquisition. In Tasks 5 and 6, kinematic metrics (i.e., path length, number of movements) have the greatest impact on the model's prediction while goal-oriented features do not influence the final decision. This outcome is logical because, even though trainees achieve the task goals in every trial, they do not always do so in the optimal way, which is the primary objective of the training process. Thus, for both tasks, a trainee is likely to be classified at the End of training if they move the instruments along the optimal (shortest) path, avoiding any unnecessary movement. For Task 7, neither kinematic nor goal-specific metrics have any impact on the classification. This suggests little difference between performance metric values for the two classes, which is expected because, as mentioned above, trainees have acquired many surgical skills that they can put into practice. Consequently, the models were able to classify the students' trials based solely on the 'Trial number' feature.

Despite the study's findings, some limitations must be addressed. First, the dataset includes training trials from 23 medical students, which may not fully capture the variability in surgical skill acquisition. Second, the study focuses on basic training tasks rather than procedural skills where surgical maneuvers may be more complex. To enhance VR training and ML-based assessment, it would be beneficial to include the Objective Structured Assessment of Technical Skill (OSATS) score in order to provide further insight on how well the ML models' outputs correlate with expert opinions [34]. However, this was not feasible in the current study due to the large number of trials (total number = 1253) and mostly, the inability to have access to the video recordings of the trainees' trials.

Moreover, testing the models in a non-virtual environment would be useful to further validate their effectiveness. However, previous studies using video-based ML assessment were unable to extract performance metrics comparable to those automatically provided by VR simulators [28]. Although some studies have successfully utilized kinematic features to provide valuable insights, these approaches often lack the detailed metrics available in VR environments, particularly those related to interaction events (e.g., number of clipping

attempts, cutting maneuvers, etc.) [6]. Reaching this level of detail through video analysis would require manual observation and recording of each event, an impractical solution due to the significant human annotation workload involved.

In future research, we aim to address these limitations by implementing several enhancements. First, we plan to include follow-up assessments involving postgraduate students, residents, and expert surgeons performing both procedural tasks and full surgical procedures. This will help validate the skills acquired and enable better generalization of our findings. Furthermore, we intend to incorporate both performance metrics and video data from training sessions. Additionally, we aim to extend the study by using vision-based feature extraction through deep learning techniques, ultimately developing a multi-input fusion deep learning model, a method that has shown promising results in similar research [13].

Overall, the findings presented in this work are promising for all three experiments. The created models could be used in real-life scenarios, such as training rooms, to provide an automated form of assessment for each trainee individually. With the help of XAI to guide professors, students will be provided with an insight into how their skills are evolving with each trial on the VR simulator, elevating the overall quality of surgical education.

**Supplementary Materials:** The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/app14219677/s1>, Table S1: Performance metrics for Tasks 5, 6, and 7; Figure S1: Color-coded confusion matrices (%) for all algorithms, except the best ones shown in the paper, for the 2-class classification experiment based on PCA (Tasks 5, 6, and 7); Figure S2: Color-coded confusion matrices (%) for all algorithms, except the best ones shown in the paper, for the 3-class classification experiment (Tasks 5, 6, and 7).

**Author Contributions:** Conceptualization, C.L.; methodology, C.L. and K.P.; software, K.P.; validation, C.L., K.P., I.S., P.K., E.P. and P.M.L.; formal analysis, C.L., K.P., I.S., P.K., E.P. and P.M.L.; investigation, C.L. and K.P.; resources, C.L., I.S., P.K., E.P. and P.M.L.; data curation, K.P.; writing—original draft preparation, C.L. and K.P.; writing—review and editing, C.L., K.P., I.S., P.K., E.P. and P.M.L.; visualization, C.L. and K.P.; supervision, C.L.; project administration, C.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The Bioethics Committee of the University General Hospital “Attikon” granted an exception for the research study described in this paper as it concerns the analysis of educational data from a virtual reality simulator following students’ training and informed consent was obtained from all participants.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author due to privacy restrictions.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Aggarwal, R.; Moorthy, K.; Darzi, A. Laparoscopic Skills Training and Assessment. *Br. J. Surg.* **2004**, *91*, 1549–1558. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Varras, M.; Nikiteas, N.; Varra, V.K.; Varra, F.N.; Georgiou, E.; Loukas, C. Role of Laparoscopic Simulators in the Development and Assessment of Laparoscopic Surgical Skills in Laparoscopic Surgery and Gynecology (Review). *World Acad. Sci. J.* **2020**, *2*, 65–76. [\[CrossRef\]](#)
3. Kantamaneni, K.; Jalla, K.; Renzu, M.; Jena, R.; Kannan, A.; Jain, R.; Muralidharan, S.; Yanamala, V.L.; Zubair, Z.; Dominic, J.L.; et al. Virtual Reality as an Affirmative Spin-Off to Laparoscopic Training: An Updated Review. *Cureus* **2021**, *13*, e17239. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Loukas, C.; Nikiteas, N.; Kanakis, M.; Georgiou, E. The Contribution of Simulation Training in Enhancing Key Components of Laparoscopic Competence. *Am. Surg.* **2011**, *77*, 708–715. [\[CrossRef\]](#)
5. Escobar-Castillejos, D.; Noguez, J.; Neri, L.; Magana, A.; Benes, B. A Review of Simulators with Haptic Devices for Medical Training. *J. Med. Syst.* **2016**, *40*, 104. [\[CrossRef\]](#)

6. Fard, M.J.; Ameri, S.; Darin Ellis, R.; Chinnam, R.B.; Pandya, A.K.; Klein, M.D. Automated Robot-Assisted Surgical Skill Evaluation: Predictive Analytics Approach. *Int. J. Med. Robot. Comput. Assist. Surg.* **2018**, *14*, e1850. [\[CrossRef\]](#)
7. Schuler, N.; Shepard, L.; Saxton, A.; Russo, J.; Johnston, D.; Saba, P.; Holler, T.; Smith, A.; Kulason, S.; Yee, A.; et al. Predicting Surgical Experience after Robotic Nerve-Sparing Radical Prostatectomy Simulation Using a Machine Learning-Based Multimodal Analysis of Objective Performance Metrics. *Urol. Pract.* **2023**, *10*, 447–455. [\[CrossRef\]](#)
8. Kowalewski, K.F.; Garrow, C.R.; Schmidt, M.W.; Benner, L.; Müller-Stich, B.P.; Nickel, F.; Friedrich, K.; Carly, K.; Mona, R.G.; Laura, W.S.; et al. Sensor-Based Machine Learning for Workflow Detection and as Key to Detect Expert Level in Laparoscopic Suturing and Knot-Tying. *Surg. Endosc.* **2019**, *33*, 3732–3740. [\[CrossRef\]](#)
9. Ebina, K.; Abe, T.; Hotta, K.; Higuchi, M.; Furumido, J.; Iwahara, N.; Kon, M.; Miyaji, K.; Shibuya, S.; Lingbo, Y.; et al. Objective Evaluation of Laparoscopic Surgical Skills in Wet Lab Training Based on Motion Analysis and Machine Learning. *Langenbeck's Arch. Surg.* **2022**, *407*, 2123–2132. [\[CrossRef\]](#)
10. Allen, B.; Nistor, V.; Dutson, E.; Carman, G.; Lewis, C.; Faloutsos, P. Support Vector Machines Improve the Accuracy of Evaluation for the Performance of Laparoscopic Training Tasks. *Surg. Endosc.* **2010**, *24*, 170–178. [\[CrossRef\]](#)
11. Benmansour, M.; Malti, A.; Jannin, P. Deep Neural Network Architecture for Automated Soft Surgical Skills Evaluation Using Objective Structured Assessment of Technical Skills Criteria. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 929–937. [\[CrossRef\]](#)
12. Benmansour, M.; Malti, A. Skills Evaluation of Specific Surgical Tasks Using Long Short Term Memory Networks. In *Proceedings of the Advances in Intelligent Systems and Computing*; Advances in Intelligent Systems and Computing; Springer Science and Business Media Deutschland GmbH: Berlin/Heidelberg, Germany, 2020; Volume 1103, pp. 331–339.
13. Kasa, K.; Burns, D.; Goldenberg, M.G.; Selim, O.; Whyne, C.; Hardisty, M. Multi-Modal Deep Learning for Assessing Surgeon Technical Skill. *Sensors* **2022**, *22*, 7328. [\[CrossRef\]](#)
14. Wang, Z.; Majewicz Fey, A. Deep Learning with Convolutional Neural Network for Objective Skill Evaluation in Robot-Assisted Surgery. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 1959–1970. [\[CrossRef\]](#)
15. Yanik, E.; Kruger, U.; Intes, X.; Rahul, R.; De, S. Video-Based Formative and Summative Assessment of Surgical Tasks Using Deep Learning. *Sci. Rep.* **2023**, *13*, 1038. [\[CrossRef\]](#)
16. Bissonnette, V.; Mirchi, N.; Ledwos, N.; Alsidieri, G.; Winkler-Schwartz, A.; Del Maestro, R.F. Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. *J. Bone Jt. Surg.—Am. Vol.* **2019**, *101*, 82–95. [\[CrossRef\]](#)
17. Siyar, S.; Azarnoush, H.; Rashidi, S.; Winkler-Schwartz, A.; Bissonnette, V.; Ponnudurai, N.; Del Maestro, R.F. Machine Learning Distinguishes Neurosurgical Skill Levels in a Virtual Reality Tumor Resection Task. *Med. Biol. Eng. Comput.* **2020**, *58*, 1357–1367. [\[CrossRef\]](#)
18. Winkler-Schwartz, A.; Yilmaz, R.; Mirchi, N.; Bissonnette, V.; Ledwos, N.; Siyar, S.; Azarnoush, H.; Karlik, B.; Del Maestro, R. Machine Learning Identification of Surgical and Operative Factors Associated with Surgical Expertise in Virtual Reality Simulation. *JAMA Netw. Open* **2019**, *2*, e198363. [\[CrossRef\]](#)
19. Topalli, D.; Cagiltay, N.E. Classification of Intermediate and Novice Surgeons' Skill Assessment Through Performance Metrics. *Surg. Innov.* **2019**, *26*, 621–629. [\[CrossRef\]](#)
20. Mirchi, N.; Bissonnette, V.; Ledwos, N.; Winkler-Schwartz, A.; Yilmaz, R.; Karlik, B.; Del Maestro, R.F. Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance. *Oper. Neurosurg.* **2020**, *19*, 65–75. [\[CrossRef\]](#)
21. Moglia, A.; Morelli, L.; D'Ischia, R.; Fatucchi, L.M.; Pucci, V.; Berchiolli, R.; Ferrari, M.; Cuschieri, A. Ensemble Deep Learning for the Prediction of Proficiency at a Virtual Simulator for Robot-Assisted Surgery. *Surg. Endosc.* **2022**, *36*, 6473–6479. [\[CrossRef\]](#)
22. Zia, A.; Essa, I. Automated Surgical Skill Assessment in RMIS Training. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 731–739. [\[CrossRef\]](#)
23. Hoffmann, H.; Funke, I.; Peters, P.; Venkatesh, D.K.; Egger, J.; Rivoir, D.; Röhrig, R.; Hölzle, F.; Bodenstedt, S.; Willemer, M.C.; et al. AIxSuture: Vision-Based Assessment of Open Suturing Skills. *Int. J. Comput. Assist. Radiol. Surg.* **2024**, *19*, 1045–1052. [\[CrossRef\]](#)
24. Alkadri, S.; Del Maestro, R.F.; Driscoll, M. Unveiling Surgical Expertise through Machine Learning in a Novel VR/AR Spinal Simulator: A Multilayered Approach Using Transfer Learning and Connection Weights Analysis. *Comput. Biol. Med.* **2024**, *179*, 108809. [\[CrossRef\]](#)
25. Mirchi, N.; Bissonnette, V.; Yilmaz, R.; Ledwos, N.; Winkler-Schwartz, A.; Del Maestro, R.F. The Virtual Operative Assistant: An Explainable Artificial Intelligence Tool for Simulation-Based Training in Surgery and Medicine. *PLoS ONE* **2020**, *15*, e0229596. [\[CrossRef\]](#)
26. Aggarwal, R.; Crochet, P.; Dias, A.; Misra, A.; Ziprin, P.; Darzi, A. Development of a Virtual Reality Training Curriculum for Laparoscopic Cholecystectomy. *Br. J. Surg.* **2009**, *96*, 1086–1093. [\[CrossRef\]](#)
27. Sinitsky, D.M.; Fernando, B.; Potts, H.; Lykoudis, P.; Hamilton, G.; Berlingieri, P. Development of a Structured Virtual Reality Curriculum for Laparoscopic Appendectomy. *Am. J. Surg.* **2020**, *219*, 613–621. [\[CrossRef\]](#)
28. Lam, K.; Chen, J.; Wang, Z.; Iqbal, F.M.; Darzi, A.; Lo, B.; Purkayastha, S.; Kinross, J.M. Machine Learning for Technical Skill Assessment in Surgery: A Systematic Review. *NPJ Digit. Med.* **2022**, *5*, 24. [\[CrossRef\]](#)
29. Martinez, O.; Martinez, C.; Parra, C.A.; Rugeles, S.; Suarez, D.R. Machine Learning for Surgical Time Prediction. *Comput. Methods Programs Biomed.* **2021**, *208*, 106220. [\[CrossRef\]](#)
30. Kilic, A.; Goyal, A.; Miller, J.K.; Gjekmarkaj, E.; Tam, W.L.; Gleason, T.G.; Sultan, I.; Dubrawski, A. Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery. *Ann. Thorac. Surg.* **2020**, *109*, 1811–1819. [\[CrossRef\]](#)

31. Twinanda, A.P.; Yengera, G.; Mutter, D.; Marescaux, J.; Padoy, N. RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations. *IEEE Trans. Med. Imaging* **2019**, *38*, 1069–1078. [[CrossRef](#)]
32. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
33. Thomson, W.; Roth, A.E. The Shapley Value: Essays in Honor of Lloyd S. Shapley. *Economica* **1991**, *58*, 123. [[CrossRef](#)]
34. Martin, J.A.; Regehr, G.; Reznick, R.; Macrae, H.; Murnaghan, J.; Hutchison, C.; Brown, M. Objective Structured Assessment of Technical Skill (OSATS) for Surgical Residents. *Br. J. Surg.* **1997**, *84*, 273–278. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.