www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(4), Sep-Dec 2025, 273-287. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2025.v16i4.940

A Survey of Machine Learning Based Systems for Evaluating Expertise Classification in Medical Simulators

Monserrat Ríos-Hernández[1], Juan Manuel Jacinto-Villegas[2], Adriana Herlinda Vilchis González[1], Otniel Portillo Rodríguez[1]
[1] Faculty of Engineering, Universidad Autónoma del Estado de México, Mexico
[2] SECIHTI, Mexico
mriosh003@profesor.uaemex.mx, jmjacintov@uaemex.mx, avilchisg@uaemex.mx, oportillor@uaemex.mx

Abstract. Medical simulators provide a safe environment for practising crucial procedures, particularly in virtual simulators where objective and quantitative data can be collected for developing machine learning algorithms for automatic expertise classification. This survey analyses 13 automatic evaluation systems used in medical simulators and identifies best practices for integrating ML algorithms. Among these systems, nine employed commercial simulators, particularly NeuroVR and the Da Vinci robotic systems, while four utilised custom simulators. The survey outlines the main steps in the integration of machine learning algorithms: data collection, metric generation and selection, training, and testing. Metric selection was identified as a crucial factor affecting both the accuracy of the algorithm and the comprehension of the evaluation. Typically, multiple machine learning algorithms were applied to the same dataset to compare results and identify the most effective model. Overall, this survey suggests that transparent algorithms are preferable, as they enhance physicians' understanding.
Keywords: medical simulators, machine learning, automatic evaluation

Article Info
Received February 25, 2025
Accepted July 6, 2025

1 Introduction

Medical simulators provide a safe and controlled environment for practicing and evaluating complex procedures, as well as for acquiring clinical skills, without putting real patients at risk. These simulators may range from physical box trainers with realistic components that mimic physical interaction to fully virtual environments that simulate entire procedures (Oquendo et al., 2018). Virtual simulators provide a structured environment for skill acquisition and evaluation by collecting objective, high-fidelity data on user performance.

However, the analysis of such data is time-consuming and susceptible to human bias. Conventionally, performance evaluation by expert observers relies on checklists for evaluating procedural steps, which are useful in identifying omissions among novices. Nevertheless, checklists may disadvantage experts' practitioners, as they are time-consuming and can potentially undervalue expertise. Therefore, it is crucial to consider factors such as speed, efficiency, and overall performance in evaluations (Gerard et al., 2013; Kelly et al., 2020). Moreover, evaluating medical procedural training that involves multiple skills using isolating metrics may not capture the complexity of skill acquisition. This limitation has driven the development of systems capable of processing quantitative data from multiple sources (Bissonnette et al., 2019). While theoretical knowledge is evaluated through objective testing, practical skills remain largely dependent on subjective evaluator judgements, highlighting the need for automation in assessment (Nguyen et al., 2019).

In response to these limitations, Artificial Intelligence (AI), and specifically Machine Learning (ML), has emerged as a powerful solution for enabling automated and objective performance evaluation in medical simulators. AI is defined as a set of algorithms capable of taking intelligent decisions, and ML is a part of AI that identifies and learns patterns from different groups (Mirchi, Bissonnette, Ledwos, et al., 2020). The ML algorithms must be trained on real data to recognize and categorize new data accurately, with the objective to improve some activities, such as the disease detection (Castillo et al., 2024), risk factor analysis

(Vidal & Gordillo, 2023), agriculture improvement (Zavala-Díaz et al., 2024), and performance optimization in sports (Ruiz-Vanoye et al., 2017).

In medical simulators, the integration of ML enables automated evaluations by accurately classifying expertise, distinguishing between non-experts (novices) and experts (Winkler-Schwartz, Bissonnette, et al., 2019). Expertise in medical procedures correlates with improved clinical outcomes and decision-making. These skills can be quantified through metrics such as instrument motion, force application, and procedural efficiency, to obtain an objective evaluation (Mintz & Brodie, 2019). The analysis of these parameters enhances both training programs and patient safety. Additionally, integrating ML for assessing user performance helps understand the components of expertise and reduces skills gaps caused by disparity in training (Fazlollahi et al., 2022).

Previous studies have demonstrated that exists a difference in hand motion patterns in individuals with different expertise in surgical skills, including open surgery (Genovese et al., 2016), laparoscopic surgery (Mason et al., 2013) and robot assisted minimally invasive surgery (Liang et al., 2018). Recent studies have proposed various approaches to skill and performance classification, differing in algorithm selection and methodological implementation. Despite these differences, most authors follow a common structure for developing automated evaluation systems. This general process involves four key steps: data collection, metrics generation and selection, training and testing (see Fig. 1).

The contribution of this survey is to identify and analyze how ML algorithms are integrated into virtual simulators for medical training. It identifies best practices in data collection, feature selection, model training, and evaluation. For this survey a bibliographic search on the electronic Google Scholar database using the date intervals 2018 to 2024 was performed with the next keywords *Machine learning, medical simulators, virtual reality simulators, classifiers, skills assessment*, where only medical simulators for training medical skills were selected. Narrative data were extracted from each paper using the steps mentioned in the Fig. 1.

The reminder of this paper is organized as follows; Section 2 presents the data collection process. Section 3 presents the metrics generation and selection. Section 4 describes the testing process. Section 5 provides a discussion and analysis of the advantages and disadvantages of machine learning-bases evaluation. Finally, Section 6 shows the conclusions.

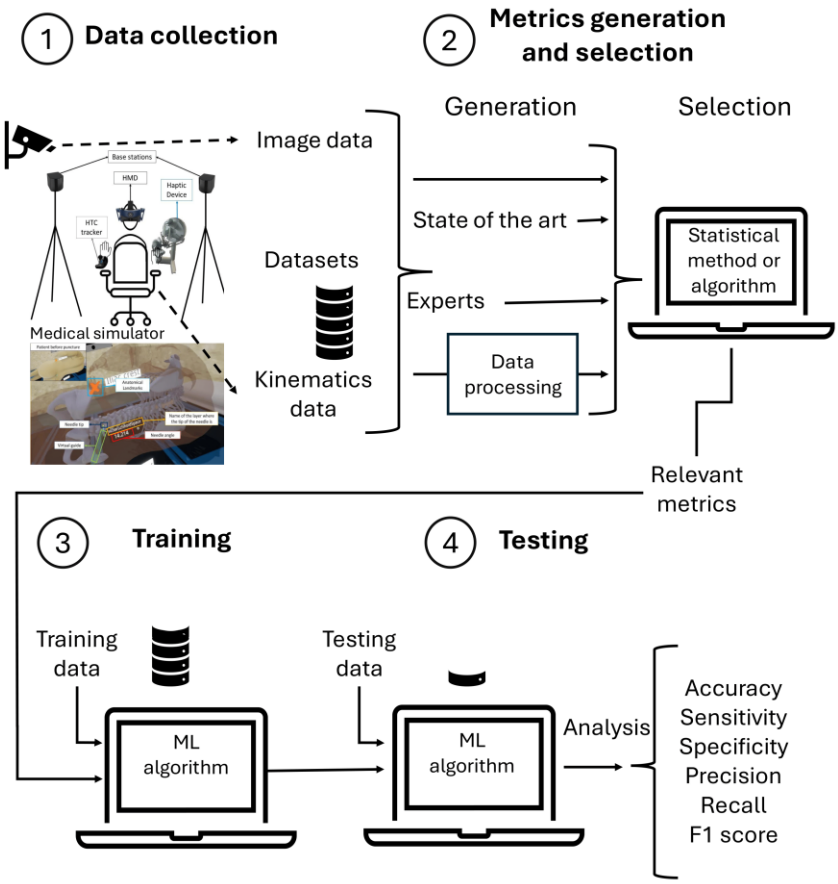

Fig. 1. Steps for integrating an ML algorithm in medical simulators.

2 Data collection

The base of ML algorithms in medical simulators relies on the quality and diversity of the collected data. Data acquisition can be achieved through different sources, including commercial simulators, publicly available datasets and custom-designed sensor-integrated systems.

Several studies (Kelly et al., 2020; Funke et al., 2019; Gorantla & Esfahani, 2019; Khalid et al., 2020) have utilized publicly available datasets, with John Hopkins University and Intuitive Surgical Inc gesture and Skill Assessment Working Set (JIGSAWS) being the most frequently used. This dataset includes data extracted from three elementary surgical tasks (suturing, knot-tying and needle passing) performed using the DaVinci® surgical system.

Another commonly used source is commercial simulators, such as the Neuro VR® simulator, which has been frequently applied in neurosurgical training research (Bissonnette et al., 2019; Winkler-Schwartz, Bissonnette, et al., 2019; Siyar et al., 2020). Alternatively, some researchers have opted to develop their own systems to collect data tailored to their specific evaluation needs. These systems are often integrated with sensors to collect specific task-metrics that may not be available in commercial platforms (Oquendo et al., 2018; Chen et al., 2021; Loukas et al., 2020; Uemura et al., 2018).

Data collection forms the base for defining variables required to develop ML metrics. Generally, three main types of data are obtained: a) image or video data: captured through camera-based systems to analyze motion, gestures, and instrument usage; b) kinematic data: includes motion parameters such as velocity, acceleration, and trajectory, related to the physical attributes of the task; c) event data: records discrete actions such as button presses or instrument activation.

Each of these data types plays a crucial role in shaping the performance metrics used to train ML algorithms. Properly structuring and processing these data sources ensures the reliability and accuracy of automated evaluation systems in medical simulators.

3 Metrics selection

Metrics selection is a fundamental step in the development of ML algorithms for medical simulators. The accuracy and computational efficiency of ML-based evaluation systems depend on the relevance and quality of the metrics selected. These metrics serve as the quantitative representation of user performance, derived from the raw data collected during simulation sessions (Bissonnette et al., 2019). They not only enable the assessment of technical skills but also help create evaluation tools that are easy to understand for both teachers (medical experts) and medical students (Winkler-Schwartz, Bissonnette, et al., 2019).

3.1 Metrics generation

Metrics used in these systema can be originate from various sources, depending on the specific application of the ML algorithm and the medical procedure. The generation of these metrics typically follows three main approaches (Mirchi, Bissonnette, Yilmaz, et al., 2020, 2020): first, expert consultation, where physicians identify critical performance indicators based on clinical experience and components involved in safe procedures; second, literature review, which involves prior research of the state of the art to identify metrics that have been used by previous studies to create automatic evaluation systems with medical simulators; and third, development of new metrics, where new metrics are designed to help to identify the degree of expertise of physicians, particularly for procedures that lack standardized assessment criteria. Several studies categorize metrics into distinct groups such as efficiency, precision, safety, movement and motion of tools (Gerard et al., 2013; Bissonnette et al., 2019; Mirchi, Bissonnette, Ledwos, et al., 2020; Gorantla & Esfahani, 2019; Yilmaz et al., 2022). Table 1 summarizes the most frequently used metrics across research studies, organized by categories.

3.2 Metrics selection

Once potential metrics are defined, a selection process is required to filter out redundant, irrelevant, or noisy data while preserving the most meaningful features. This step often involves data preprocessing, which includes eliminating erroneous entries, normalizing datasets, and applying statistical or algorithmics techniques to refine metric selection (Mirchi, Bissonnette, Yilmaz, et al., 2020). The most used methods for metric selection are in

Various ML algorithms have been employed in medical simulators to classify expertise levels based on performance metrics. The selection of the algorithm depends on the complexity of dataset, computational efficiency and sample size. The most used include the Support Vector Machine, K-nearest neighbors, Naïve Bayes, Artificial Neural Networks, Random Forest, Boosting, and Discriminant analysis (see Table 3). Once the model has been trained, it must be evaluated to verify its ability to classify unseen data correctly, this step is detailed in the following section.

4 Testing

Once the algorithm with the best performance has been selected, the next step is to evaluate its effectiveness through testing. This step is essential to observe the utility and reliability in the context of medical simulators.

Testing involves assessing the model's ability to accurately classify new data into the predefined expertise groups. In general, this evaluation is conducted using 20% of data reserved from the original dataset. As detailed in Table 4, some studies used the same algorithm for training and testing, while others applied methods such as the LOOCV (described in Section 3) or 5-fold-cross validation (Brown et al., 2020). This algorithm consists of splitting all the data into five folds or subsets; during each iteration, four folds are used for training while the remaining one is used for testing. This process is repeated five times, using each fold to test the algorithm. The model´s overall performance is then computed by aggregating the results across all folds.

Additional metrics such as sensitivity (recall), specificity and F1 score provide a more comprehensive evaluation (Loukas et al., 2020). The sensitivity (see Equation 2) also known as recall (Khalid et al., 2020) is defined as ratio of T_P divided by P, it indicates the proportion of positive instances that were correctly classified. Meanwhile, Specificity (see Equation 3) is the ratio of T_N to the total number of Negatives (N). It measures the proportion of negative instances correctly classified divided by the total of the negative instances. The precision (see Equation 4) is a measure that represents the T_P is divided by the sum of T_P and F_P (Khalid

et al., 2020). It represents the proportion of positive instances correctly identified out all the instances that are classified as positive. Finally, the F1 score is the representation of the balance between the sensitivity and precision scores (see Equation 5).

$$\text{Accuracy} = \frac{T_P + T_N}{P+N} \tag{1}$$

$$\text{Sensitivity} = \frac{T_P}{P} \tag{2}$$

$$\text{Specifity} = \frac{T_N}{N} \tag{3}$$

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{4}$$

$$\text{F1 score} = \frac{2}{\frac{1}{Sensitivity} + \frac{1}{Precision}} \tag{5}$$

Table *2*.

The effectiveness of an ML algorithm is closely tied to the quality and relevance of the selected metrics. Studies have shown that selecting the right subset of metrics not only improves classification accuracy but also enhances the interpretability of evaluation results, making them more intuitive for medical professionals (Bissonnette et al., 2019; Loukas et al., 2020).

Table 1. Metrics used in medical simulators and their categories (Gerard et al., 2013; Bissonnette et al., 2019; Mirchi, Bissonnette, Ledwos, et al., 2020; Gorantla & Esfahani, 2019; Yilmaz et al., 2022)

Metric	Category
Force change in a tool (instrument)	Safety
Mean force applied on a tissue	Safety
Maximum force applied on a tissue	Safety
Number of times a tissue was touched	Safety
Bleeding speed	Safety
Total blood loss	Safety
Tissue volume removed	Safety & Quality
Force used by a tool (N)	Safety & Efficiency
Time no force is applied by any tool	Efficiency
Sum of every change in position of a tool	Efficiency
The amount of time spent removing something	Efficiency
Time to completion	Efficiency
Total path traveled by an instrument	Efficiency & Overall motion
Instrument tip separation change	Bimanual cognitive
Instrument tip separation distance Bimanual	Cognitive
Instrument choice	Cognition
Tool acceleration	Movement
Tool velocity	Movement
Distance between 2 acceleration peaks	Motion of tools
Angular velocity	Motion of tools
Mean acceleration	Motion of tools & Dynamic features
Mean velocity	Motion of tools & Dynamic features
Variance of angles of an instrument	Motion of tools & Turning angle features
Coordination between two instruments	Coordination
Degree of smoothness	Dynamic features

Various ML algorithms have been employed in medical simulators to classify expertise levels based on performance metrics. The selection of the algorithm depends on the complexity of dataset, computational efficiency and sample size. The most used include the Support Vector Machine, K-nearest neighbors, Naïve Bayes, Artificial Neural Networks, Random Forest, Boosting, and Discriminant analysis (see Table 3). Once the model has been trained, it must be evaluated to verify its ability to classify unseen data correctly, this step is detailed in the following section.

5 Testing

Once the algorithm with the best performance has been selected, the next step is to evaluate its effectiveness through testing. This step is essential to observe the utility and reliability in the context of medical simulators.

Testing involves assessing the model's ability to accurately classify new data into the predefined expertise groups. In general, this evaluation is conducted using 20% of data reserved from the original dataset. As detailed in Table 4, some studies used the same algorithm for training and testing, while others applied methods such as the LOOCV (described in Section 3) or 5-fold-cross

validation (Brown et al., 2020). This algorithm consists of splitting all the data into five folds or subsets; during each iteration, four folds are used for training while the remaining one is used for testing. This process is repeated five times, using each fold to test the algorithm. The model´s overall performance is then computed by aggregating the results across all folds.

Additional metrics such as sensitivity (recall), specificity and F1 score provide a more comprehensive evaluation (Loukas et al., 2020). The sensitivity (see Equation 2) also known as recall (Khalid et al., 2020) is defined as ratio of T_P divided by P, it indicates the proportion of positive instances that were correctly classified. Meanwhile, Specificity (see Equation 3) is the ratio of T_N to the total number of Negatives (N). It measures the proportion of negative instances correctly classified divided by the total of the negative instances. The precision (see Equation 4) is a measure that represents the T_P is divided by the sum of T_P and F_P (Khalid et al., 2020). It represents the proportion of positive instances correctly identified out all the instances that are classified as positive. Finally, the F1 score is the representation of the balance between the sensitivity and precision scores (see Equation 5).

$$\text{Accuracy} = \frac{T_P + T_N}{P + N} \tag{1}$$

$$\text{Sensitivity} = \frac{T_P}{P} \tag{2}$$

$$\text{Specifity} = \frac{T_N}{N} \tag{3}$$

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{4}$$

$$\text{F1 score} = \frac{2}{\frac{1}{Sensitivity} + \frac{1}{Precision}} \tag{5}$$

Table 2. Methods for metric selection

Method	Description
Forward or backward feature selection (Mirchi, Bissonnette, Yilmaz, et al., 2020)	Iterative algorithms that evaluate model performance at each step. Forward selection begins with an empty set and adds one feature at a time. In contrast, backward selection starts with all the features and progressively removes them one by one.
Connection Weight Algorithm (Mirchi, Bissonnette, Ledwos, et al., 2020)	Assesses the contribution of each metric to the classification model, ranking them by importance.
Leave-One-Out-Cross-Validation (LOOCV) (Winkler-Schwartz, Bissonnette, et al., 2019)	Trains the algorithm on all available data except for one participant, whose data is used for testing. This process is repeated iteratively for all participants to determine the most robust metrics.
Sequential Forward feature Selection (Gorantla & Esfahani, 2019)	Features are added one by one to an initially empty set of candidates. This algorithm continues until the addition of new features no longer improves the classification rate.
Recursive Feature Elimination (Brown et al., 2020)	This recursive algorithm ranks the metrics by importance, progressively reducing the number of features until a desired subset is reached.
Statistical t-Test (Siyar et al., 2020)	This test helps to select features with statistical differentiation by comparing the means of different groups.
Wilcoxon Rank-Sum test (Cuzick, 1985)	A statistical method for comparing two groups (smokers - non smokers) when apparently there is not a clear model.
Semantic segmentation (Khalid et al., 2020)	Used in image processing. This method identifies key points essential for estimating surgical performance, such as the orientation, position and size of surgical tools. It allows for outlining and labelling specific regions in an image.

Table 3. ML algorithms employed in medical simulators

Machine learning algorithms	Description
Support Vector machine (SV) (Bissonnette et al., 2019)	Uses a hyperplane to separate data in 2 or more groups, maximizing the distance between the closest points of each group.
K-nearest neighbors (KNN) (Bissonnette et al., 2019)	Determines the class of a participant based on the closest neighbors using Euclidean distance. The parameter k represents the number of neighbors considered, and the classification is based on the relationship with the nearest participants in a multidimensional space.
Naive Bayes (Bissonnette et al., 2019)	Classifies participants with the probability that the chosen metrics belong to either experts or novice surgeons. It assumes that all the chosen metrics are independent from each other.

Artificial Neural Network (ANN) (Uemura et al., 2018)	Comprises multiple computing cells (neurons) that work in parallel to process ir and generate a result, often handling raw data.
Random forest (Chen et al., 2021)	This algorithm generates separate independent decision trees, with each tree contributing to the result.
Gradient boosting (Chen et al., 2021)	It is a boosting model that is used to train a series of weak classifiers. Each classifier learns from residual errors from the previous step. The final prediction is the sum of all previous predictions.
Discriminant analysis (Bissonnette et al., 2019)	This algorithm consists of projecting the data on a single dimension to maximize the distance between the means of the groups.
AdaBoost (Chen et al., 2021)	This boosting model updates the weights of data points by weighing each classifier based on the related errors in sequence. The final prediction is a weighted majority of all classifier's input.

Statistical measures such as the coefficient of determination (R^2) are also used to describe and evaluate how the model's output fits the actual data (Kowalewski et al., 2019). Additionally, the Root Mean Square Error (RMSE) (Zhang et al., 2020) quantifies the square root of the variance of the standard error between the desired and obtained value. The analysis of the variance (ANOVA) compares mean classification accuracy from different models and datasets (Chen et al., 2021).

For visual interpretation, confusion matrices are commonly used to illustrate the relation between real and predicted instances (Bissonnette et al., 2019; Mirchi, Bissonnette, Ledwos, et al., 2020; Lee et al., 2020), as shown in Fig. 2. Alternatively, some authors (Nguyen et al., 2019) utilize a performance matrix to display metric values across different expertise groups, as seen in Fig. 3. Other evaluation strategies include balanced accuracy score (average of recall between training and testing data) and Matthews's correlation coefficient that gives a value between -1 (poor prediction) and 1 (perfect prediction) (Brown et al., 2020).

6 Discussion

This section analyzes the automatic evaluation systems developed for classifying expertise medical simulators based on the review of relevant literature. It is important to mention that it is not possible to make a comparison between the automatic evaluation systems used in medical simulators since they use different metrics and different numbers of participants, which can be reflected in the results obtained; however, a discussion can be made regarding them to identify their most relevant characteristics. Table 4 provides a summary of the main characteristics of these systems, it highlights aspects involved in the steps mentioned before, such as metric selection methods, the number of metrics proposed and the way that the results are shown.

Among the 13 systems reviewed (see

Table **4**, Table 5 and

Table **6**), nine collected data from commercial simulators, with the NeuroVR shown in

Table **4** (Bissonnette et al., 2019; Siyar et al., 2020; Winkler-Schwartz, Yilmaz, et al., 2019) and DaVinci robotic systems inTable 5, where the data comes from the public dataset JIGSAWS (Khalid et al., 2020; Mirchi, Bissonnette, Yilmaz, et al., 2020; Brown et al., 2020; Lee et al., 2020; Ismail Fawaz et al., 2019). In contrast, four studies developed their own systems to collect specific data necessary for creating the metrics that they identified, all of them being box trainers for laparoscopic procedure (Oquendo et al., 2018; Loukas et al., 2020; Uemura et al., 2018; Kowalewski et al., 2019). As shown in

Table **6** each simulator presents tasks that allow for evaluating surgical skills, line needle passing (NO), knot tying (KT) and suturing.

The number of participants varied considerably across the studies (see Fig. 4). Three studies included fewer than 20 participants, six involved between 20 and 50 participants, two had between 51 and 100 individuals participated, and two studies included more than 100 participants. While the number of participants often depends on their availability, having a larger number of participants generally allows the ML algorithm to be trained with greater accuracy. This is because more data provides a better representation of different expertise levels, making the algorithm able to generalize. However, fields such as medicine, recruiting many participants can be particularly challenging due to time constraints, ethical considerations, and the specialized nature of the population.

Metric selection is a critical step in developing effective ML algorithms for medical simulators. The use of a larger number of metrics can increase computational costs, but the number of metrics also impacts the accuracy of the algorithm. Only one study selected metrics based on expert opinions (

Table **6**), identifying metrics that distinguish between experts and novices (Gorantla & Esfahani, 2019). The methods for metrics selection used in the reviewed studies are presented in Fig. 5, some studies combine two methods for this selection (Bissonnette et al., 2019; Siyar et al., 2020), including the backward or forward selection algorithm, while another study selected metrics according to the Global Evaluative Assessment of Robotic Surgery (GEARS) and the Object Structured Assessment (OSATS) of Technical Skills, scales used in the medical education. Additionally, two studies consider the expert opinion in order to detect metrics related to a specific medical procedure.

Although this survey presents the most used metrics of the 13 systems reviewed, the selected metrics will depend on the medical procedure the simulator is focused on. For instance, the metrics will not be the same for a diagnostic procedure where palpation is the basis, as for a procedure where one of the most relevant aspects is the use of instruments (such as robotic surgery). For this reason, during the metric selection process, it is important to consider input from medical experts. Their insights can provide valuable guidance on what aspects are essential in specific medical procedures. Additionally, involving medical experts helps ensure an understanding of how the metrics impact the evaluation of medical skills. However, two systems that employed neural networks did not perform any metric selection (Uemura et al., 2018; Ismail Fawaz et al., 2019), likely relying on the neural network´s ability to process raw data. On the other hand, studies that used video data obtain the metrics through video analysis (Chen et al., 2021; Lee et al., 2020), dynamic time wrapping data (Kowalewski et al., 2019), and semantic segmentation (Khalid et al., 2020) (see Table 5 and

Table **6**).

Regarding the metrics selection, the number of proposed metrics ranged from 9 to 369, depending on the data collected and the source. Studies proposing fewer than 20 metrics (Siyar et al., 2020; Loukas et al., 2020; Lee et al., 2020) generally did not perform metric selection, training and testing the algorithms with all the metrics. In contrast, studies proposing more than 20 metrics reduced the number of metrics through selection process, aiming to identify the most relevant and indispensable metrics for greater algorithm accuracy.

Fourteen different algorithms were used across the reviewed studies, as shown in Fig. 6. The most common were different types of neural network (Khalid et al., 2020; Mirchi, Bissonnette, Yilmaz, et al., 2020; Kowalewski et al., 2019) such as fully convolutional (Ismail Fawaz et al., 2019) and deep neural network (Uemura et al., 2018). Four studies used Support Vector Machine (Bissonnette et al., 2019; Kowalewski et al., 2019; Lee et al., 2020; Winkler-Schwartz, Yilmaz, et al., 2019), the K-nearest neighbors' algorithm was also used in four (Bissonnette et al., 2019; Siyar et al., 2020; Loukas et al., 2020; Yilmaz et al., 2022). It is important to highlight that in various studies (Oquendo et al., 2018; Bissonnette et al., 2019; Winkler-Schwartz, Bissonnette, et al., 2019; Mintz & Brodie, 2019; Fazlollahi et al., 2022; Chen et al., 2021; Kowalewski et al., 2019; Lee et al., 2020), the same dataset is used for training and testing across different ML algorithms, with the same or varying metrics.

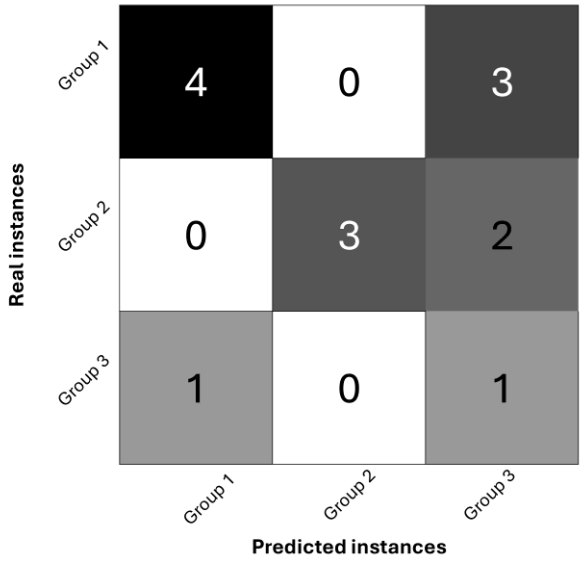

Fig. 2. Example of a Confusion Matrix.

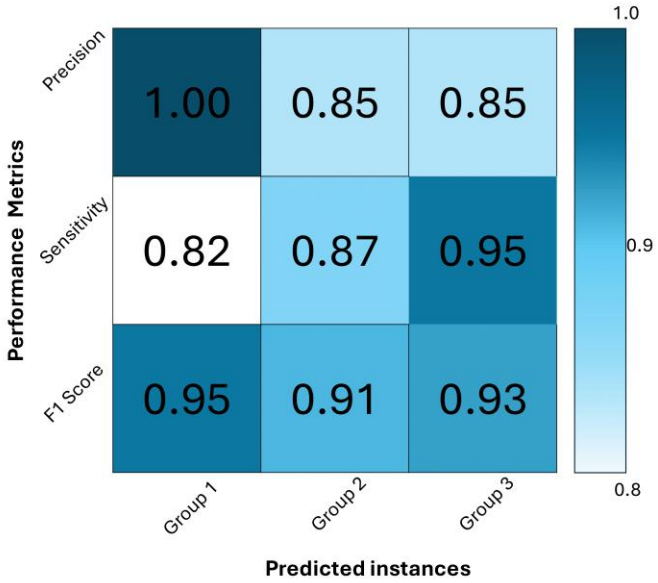

Fig. 3. Example of a Performance Matrix.

As shown in Fig. 7, the results obtained from the ML algorithm, accuracy is the most common metric reported by all authors. Additionally, three studies reported sensitivity (Bissonnette et al., 2019; Khalid et al., 2020; Loukas et al., 2020), two reported specificities (Bissonnette et al., 2019; Loukas et al., 2020), and one reported precision and F1 Score (Khalid et al., 2020). Despite the relatively limited variety of reported metrics, it is important to emphasize that many additional performance indicators—such as sensitivity, specificity, precision—can be derived if the confusion matrix is available, even if these metrics are not explicitly stated in the publications. The confusion matrix provides the fundamental counts of true positives, true negatives, false positives, and false negatives, serving as the basis for calculating a comprehensive set of performance measures. However, as depicted in Fig. 8, only a small subset of the studies included in this review reported the confusion matrix as part of their results.

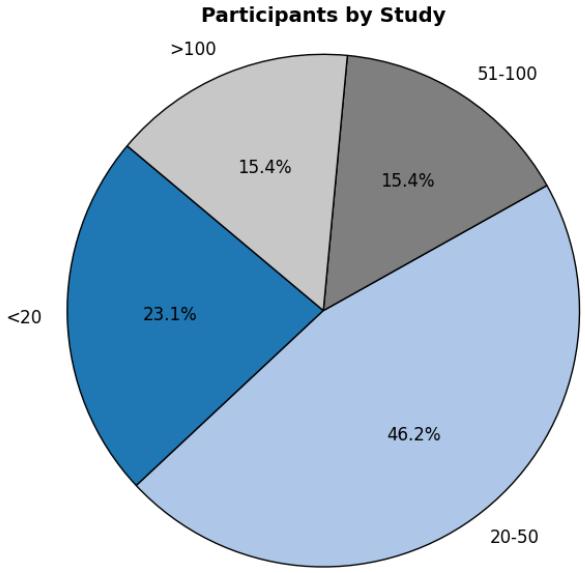

Fig. 4. Number of participants by study.

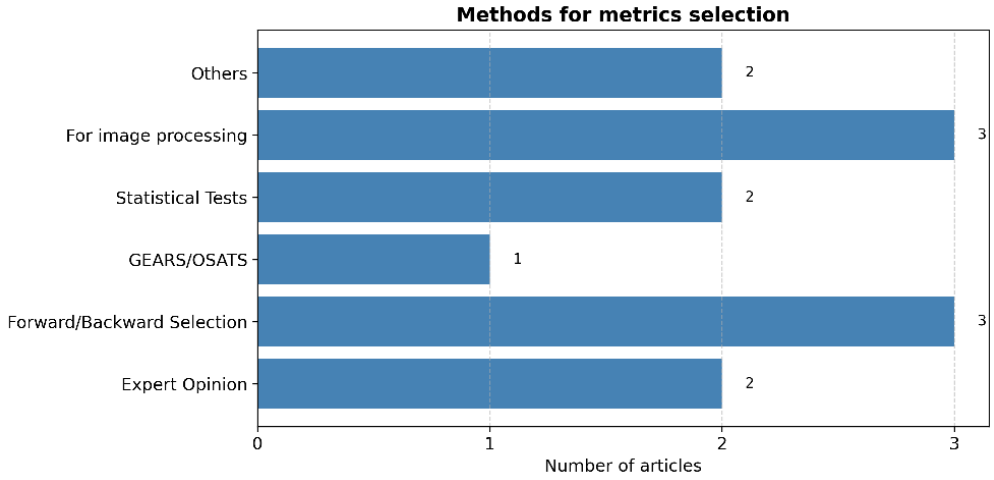

Fig. 5. Methods used for metrics selection.

In the context of medical education, ML-based systems have the potential to enhance objectivity, explainable and individualized performance feedback. However, to ensure these systems are effective and suitable for physicians and educational integration, it is crucial to align them with ethical principles such as interpretability, explainability, responsibility and trustworthiness. Explainability is essential for enabling physicians and experts to understand how the system evaluates performance and how to interpret its outputs. This transparency improves the acceptability of the system with the ML algorithm, promotes trust in the results, and can positively influence skill acquisition (Mirchi, Bissonnette, Yilmaz, et al., 2020). Furthermore, aligning AI-based systems with existing assessment frameworks, ethics and involving medical professionals in their design and implementation can ensure these tools are used responsibly and effectively (Rasheed et al., 2022).

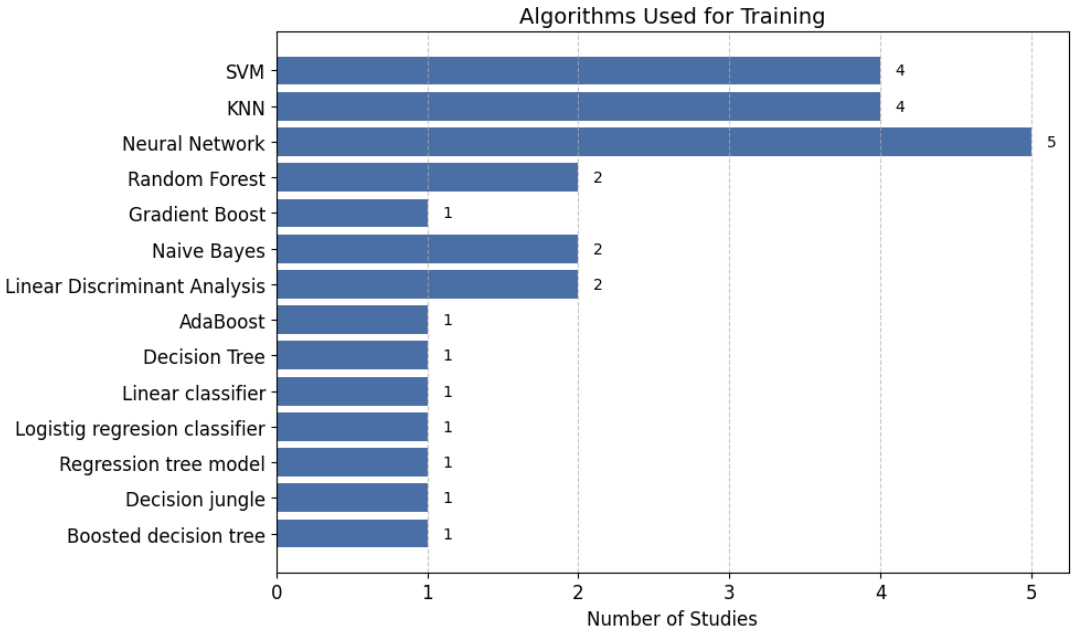

Fig. 6. Algorithms used for training with the metrics selected.

In addition, it is important to note that the reviewed systems focus primarily on accurately classifying expertise and identifying representative performance metrics. However, they do not report consistent integration into medical curricula. This reveals a critical gap: determining whether these systems are truly prepared to distinguish between users who possess true procedural skills and those who merely demonstrate technical ability to operate the simulator. Furthermore, most evaluations were conducted over short periods or involved only a single training session, limiting the evidence of long-term effectiveness. To eliminate this gap, ML-based teaching platforms must undergo rigorous validation processes involving experts and demonstrate that the skills acquired in simulators transfer effectively to real-world.

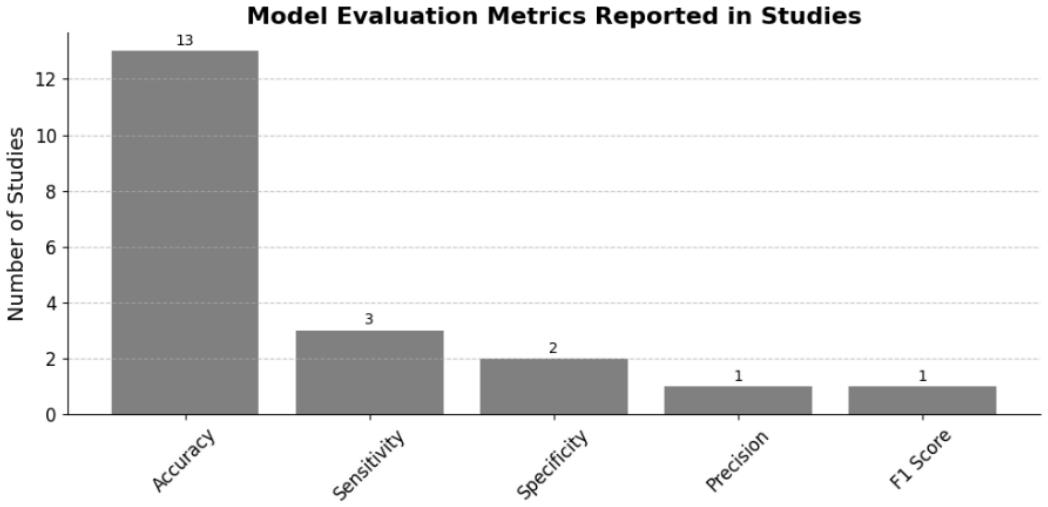

Fig. 7. Evaluation of metrics for model performance.

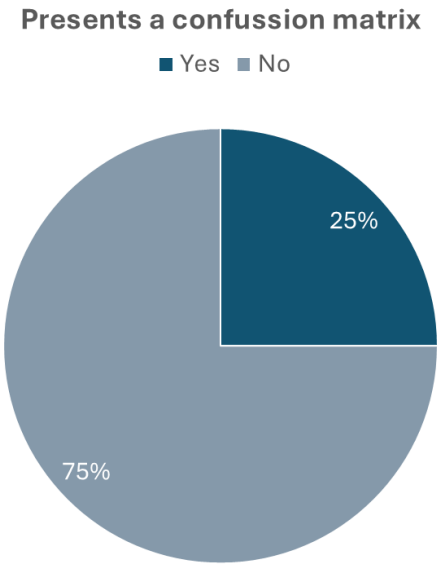

Fig. 8. Number of studies reporting a confusion matrix as part of their results.

7 Conclusions

Automated evaluation systems in medical simulators are essential for providing standardized and objective assessments of expertise in specialized procedures such as laparoscopic and robotic surgery. These systems not only enable precise and consistent evaluations but also help physicians identify the key metrics necessary for effective training. By identifying these metrics, medical educators and simulation developers of medical procedures can design targeted strategies to enhance training programs, ultimately improving the overall quality of medical education.

The studies presented in this article have shown different characteristics, varying from the type of simulated medical procedure to the algorithms employed for training and testing. However, a common aspect among these studies is the way results are presented: all report the accuracy of the algorithms used. While accuracy is a common measure in ML, it is less intuitive for physicians, who are more familiar with terms such as specificity and sensitivity (Winkler-Schwartz, Bissonnette, et al., 2019). Therefore, it is crucial to provide measures that are easily understandable for physicians and medical professionals, offering a comprehensive perspective on the evaluation tool. This includes not only correct classification (accuracy) but also misclassified data. In this regard, the confusion matrix serves as a visual representation of test outcomes, allowing physicians to analyze both classified and misclassified data in an easily understandable format.

Regarding this situation, algorithms with non-transparent processes, such as neural networks, may not be the best option if physicians need to understand how results are obtained. Instead, using transparent algorithms with metrics selection can facilitate better comprehension and trust among medical professionals. A widely adopted approach in the development of automated evaluation systems for medical simulators involves testing multiple ML on a common dataset to identify the most effective model. This approach enables a direct comparison of the algorithm's performance, facilitating the selection of the most suitable algorithm based on the desired outcomes.

Finally, following the appropriate steps for integrating a ML algorithm into a medical simulator enables significant improvements in the simulator's functionality. The integration of ML-based evaluation enables comprehensive post-training assessment, facilitating targeted feedback and skill refinement. Consequently, it opens the possibility of enhancing the feedback mechanism and correcting errors that occurred during the training. This iterative improvement will help make the simulator an effective tool for medical education and skill development.

In summary, while ML-based assessment systems present promising characteristics for improving medical training, their successful implementation depends not only on technical performance but also on ethical integrity, curricular integration, and validation of real-world skill transfer. Addressing these aspects will be essential to ensure that such systems truly support medical education and contribute meaningfully to clinical competence development.

Table 4. Automatic evaluation systems for skill assessment with the Neuro VR simulator

Studies	Participants - Classification groups	Metrics selection	Data source	M_P (M_S)	Testing	Algorithm training (M_S)	Results
(Yilmaz et al., 2022)	50 - Neurosurgeons, Senior residents, Junior residents, medical students	Backward and forward selection	Simulator	270 (122)	LOOCV	K-nearest neighbors (6)	A: 90%
						Naive Bayes (9)	A: 84%
						Discriminant analysis (8)	A: 78%
						SVM (8)	A: 76%
(Bissonnette et al., 2019)	41 - Senior and junior	By spine surgeons and backward selection	Simulator	41 (12)	LOOCV	K-nearest neighbors	A: 92.7%
						Naive Bayes	A: 70.7%
						Linear discriminant analysis	A: 87.8%
						SVM	A: 97.6%, S: 100%, SP: 94.7%
						Decision tree	A: 70.7%
(Siyar et al., 2020)	115 - Skilled and novices	Statistical and forward selection	Simulator	150 (68)	Cross validation	K-nearest neighbors (6)	A: ≈ 93%

M_P = Metrics proposed, M_S = Metrics selected, A = Accuracy, S = Sensitivity, SP = Specificity, SVP = Support Vector Machine, LOOCV = Leave-out-one-cross-validation

Table 5. Automatic evaluation systems for skill assessment with the DaVinci system

Studies	Participants-Classification groups	Metrics selection	Data source	M_P (M_S)	Testing	Algorithm training (M_S)	Results
(Khalid et al., 2020)	8 - Novice, intermediate and expert	Semantic segmentation (image processing)	JIGSAWS dataset	-------	Same algorithm used in training	Neural network	A: (ST, KT, NP) 36%; P: (ST) 100%, (KN, NP) 1%; S: (ST) 100%, (KT) 35%, (NP): 32%; F1: (ST) 100%, (KT) 13%, (NP) 3%
(Lee et al., 2020)	52 - Novice, skilled and expert	OSATS (O) and GEARS (G) score	Video data	9	LOOCV	Linear classifier	A: (O) 58%, (G) 67%
						SVM	A: (O) 75%, (G) 67%
						Random forest	A: (O) 83%, (G) 83%
(Ismail Fawaz et al., 2019)	8 - Novice, intermediate and expert	No apply	JIGSAWS dataset	76	Spearman's coefficient	Fully convolutional neural network	A: (ST) 100%, (NP) 100%, (KT) 92.1%
(Brown et al., 2020)	>100 - Trainee, expert surgeon and training specialist	Wilcoxon Rank and Recursive Feature elimination	Video and Kinematic data from simulator	43 (88-117)	Five-fold-cross-validation	Logistic regression classifier	A: 80.24-98.27%
(Chen et al., 2021)	17 - Super experts, ordinary experts, experts, novices	Video analysis	Previous dataset	-------	Same algorithms	AdaBoost	A: 69.87%
						Gradient boost	A: 67.24%
						Random forest	A: 72.75%

M_P = Metrics proposed, M_S = Metrics selected, A = Accuracy, S = Sensitivity, P = Precision, ST = Suturing, KT = Knot Tying, NP = Needle Passing, F1 = F1 score, SVP = Support Vector Machine, LOOCV = Leave-out-one-cross-validation

Table 6. Automatic evaluation systems for skill assessment with diverse simulators

Studies	Simulator	Participants and classification groups	Metrics selection	Data source	M_P (M_S)	Testing	Algori-thm trai-ning (M_S)	Results
(Loukas et al., 2020)	Laparoscopic box trainer	32 - Medical students and surgical resident	By an expert	Video data	10 (each task)	K-nearest neighbors	K-nearest neighbors	A: 71-86%; S: 80-100%; SP: 60 - 80%
(Uemura et al., 2018)	Box trainer	67 - Experts and non-expert	No apply	Box trainer	NA	Deep neural network	Deep neural network	A: 79%
(Mirchi, Bissonnette, Ledwos, et al., 2020)	Sim-Ortho platform	21 - Junior, senior, post resident	Stepwiseft function	Box trainer	369 (13)	Neural network	Neural network	A: 83.3%
(Oquendo et al., 2018)	Pediatric laparoscopic box trainer	32 - Medical students, residents, fellows	The LASSO elastic net technique	TD	280 (10)	LOOCV	Regression tree model	A: 52%
				TD and GD	280 (48)			A: 52%
				TD and TPD	280 (202)			A: 46%
				TD, MD, and TVD	280 (214)			A: 59%
				TD, MD and GD	280 (190)			A: 54%
				TD, MD, TVD and GD	280 (184)			A: 71%
(Kowalewski et al., 2019)	Laparoscopic box trainer	28 - Beginner, intermediate, experts	Dynamic time wrapping for video data	Video data	12	Cross validation	Decision jungle	A: 62%
							Neural network	A: 70%
							SVM	A: 60%
							Boosted decision tree	A: 66%

M_P = Metrics proposed, M_S = Metrics selected, A = Accuracy, S = Sensitivity, SVP = Support Vector Machine, TD = Time Data, GP = Grip Data, TD = Tip Data, MD = Tip Data Motion, TVD = Tool Visibility Data.

References

Oquendo, Y. A., Riddle, E. W., Hiller, D., Blinman, T. A., & Kuchenbecker, K. J. (2018). Automatically rating trainee skill at a pediatric laparoscopic suturing task. *Surgical Endoscopy*, 32(4), 1840–1857. https://doi.org/10.1007/s00464-017-5873-6

Gerard, J. M., Kessler, D. O., Braun, C., Mehta, R., Scalzo, A. J., & Auerbach, M. (2013). Validation of global rating scale and checklist instruments for the infant lumbar puncture procedure. *Simulation in Healthcare*, 8(3), 148–154. https://doi.org/10.1097/SIH.0b013e3182802d34

Kelly, J. D., Petersen, A., Lendvay, T. S., & Kowalewski, T. M. (2020). Bidirectional long short-term memory for surgical skill classification of temporally segmented tasks. *International journal of computer assisted radiology and surgery*, 15(12), 2079–2088. https://doi.org/10.1007/s11548-020-02269-x

Bissonnette, V., Mirchi, N., Ledwos, N., Alsidieri, G., Winkler-Schwartz, A., & Del Maestro, R. F. (2019). Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. *Journal of Bone and Joint Surgery*, 101(23), e127. https://doi.org/10.2106/JBJS.18.01197

Nguyen, X. A., Ljuhar, D., Pacilli, M., Nataraja, R. M., & Chauhan, S. (2019). Surgical skill levels: Classification and analysis using deep neural network model and motion signals. *Computer methods and programs in biomedicine*, 1771–8. https://doi.org/10.1016/j.cmpb.2019.05.008

Mirchi, N., Bissonnette, V., Ledwos, N., Winkler-Schwartz, A., Yilmaz, R., Karlik, B., & Del Maestro, R. F. (2020). Artificial neural networks to assess virtual reality anterior cervical discectomy performance. *Operative Neurosurgery*, 19(1), 65–75. https://doi.org/10.1093/ons/opz359

Castillo, A. E. P., Raggi, S. E. A., Robles, L. A., Flores, A. B., & Robledo, J. F. P. (2024). Comparative Study of Lung Image Representations for Automated Pneumonia Recognition. *International Journal of Combinatorial Optimization Problems and Informatics*, 15(5), 193–193. https://doi.org/10.61467/2007.1558.2024.v15i5.578

Vidal, M. T. & Gordillo, J. M. C. (2023). Detection of Risk Factors for Diabetes Mellitus with Machine Learning. *International Journal of Combinatorial Optimization Problems & Informatics*, 14(1), 66–75.

Zavala-Díaz, N. A., Olivares-Rojas, J. C., Zavala-Díaz, J., Reyes-Archundia, E., Téllez-Anguiano, A., Chávez-Campos, G. M., & Méndez-Patiño, A. (2024). Study of Machine Learning Techniques for the Estimation of Soil Moisture in Agriculture. *International Journal of Combinatorial Optimization Problems & Informatics*, 15(4), https://doi.org/10.61467/2007.1558.2024.v15i4.502

Ruiz-Vanoye, J. A., Díaz-Parra, O., Fuentes-Penna, A., Vélez-Díaz, D., Munguía, M. G., Ruiz-Díaz, J., & Ruiz-Díaz, F. (2017). Motivation Index to Improve the Soccer Performance. *International Journal of Combinatorial Optimization Problems and Informatics*, 8(3), 45–57.

Winkler-Schwartz, A., Bissonnette, V., Mirchi, N., Ponnudurai, N., Yilmaz, R., Ledwos, N., Siyar, S., Azarnoush, H., Karlik, B., & Del Maestro, R. F. (2019). Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. *Journal of Surgical Education*, 76(6), 1681–1690. https://doi.org/10.1016/j.jsurg.2019.05.015

Mintz, Y. & Brodie, R. (2019). Introduction to artificial intelligence in medicine. *Minimally Invasive Therapy & Allied Technologies*, 28(2), 73–81. https://doi.org/10.1080/13645706.2019

Fazlollahi, A. M., Bakhaidar, M., Alsayegh, A., Yilmaz, R., Winkler-Schwartz, A., Mirchi, N., Langleben, I., Ledwos, N., Sabbagh, A. J., Bajunaid, K., Harley, J. M., & Del Maestro, R. F. (2022). Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial. *JAMA Network Open*, 5(2), e2149008. https://doi.org/10.1001/jamanetworkopen.2021.49008

Genovese, B., Yin, S., Sareh, S., Devirgilio, M., Mukdad, L., Davis, J., Santos, V. J., & Benharash, P. (2016). Surgical hand tracking in open surgery using a versatile motion sensing system: Are we there yet? *The American Surgeon*, 82(10), 872–875. https://doi.org/10.1177/000313481608201002

Mason, J. D., Ansell, J., Warren, N., & Torkington, J. (2013). Is motion analysis a valid tool for assessing laparoscopic skill? *Surgical endoscopy*, 27, 1468–1477. https://doi.org/10.1007/s00464-012-2631-7

Liang, K., Xing, Y., Li, J., Wang, S., Li, A., & Li, J. (2018). Motion control skill assessment based on kinematic analysis of robotic end-effector movements. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 14(1), e1845. https://doi.org/10.1002/rcs.1845

Funke, I., Mees, S. T., Weitz, J., & Speidel, S. (2019). Video-based surgical skill assessment using 3D convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14, 1217–1225. https://doi.org/10.1007/s11548-019-01995-1

Gorantla, K. R. & Esfahani, E. T. (2019). Surgical Skill Assessment using Motor Control Features and Hidden Markov Model. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 5842–5845. https://doi.org/10.1109/EMBC.2019.8857629

Khalid, S., Goldenberg, M., Grantcharov, T., Taati, B., & Rudzicz, F. (2020). Evaluation of Deep Learning Models for Identifying Surgical Actions and Measuring Performance. *JAMA Network Open*, 3(3), e201664. https://doi.org/10.1001/jamanetworkopen.2020.1664

Siyar, S., Azarnoush, H., Rashidi, S., Winkler-Schwartz, A., Bissonnette, V., Ponnudurai, N., & Del Maestro, R. F. (2020). Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task. *Medical & Biological Engineering & Computing*, 58(6), 1357–1367. https://doi.org/10.1007/s11517-020-02155-3

Chen, A. B., Liang, S., Nguyen, J. H., Liu, Y., & Hung, A. J. (2021). Machine learning analyses of automated performance metrics during granular sub-stitch phases predict surgeon experience. *Surgery*, 169(5), 1245–1249. https://doi.org/10.1016/j.surg.2020.09.020

Loukas, C., Gazis, A., & Kanakis, M. A. (2020). Surgical Performance Analysis and Classification Based on Video Annotation of Laparoscopic Tasks. *JSLS : Journal of the Society of Laparoscopic & Robotic Surgeons*, 24(4), e2020.00057. https://doi.org/10.4293/JSLS.2020.00057

Uemura, M., Tomikawa, M., Miao, T., Souzaki, R., Ieiri, S., Akahoshi, T., Lefor, A. K., & Hashizume, M. (2018). Feasibility of an AI-Based Measure of the Hand Motions of Expert and Novice Surgeons. *Computational and Mathematical Methods in Medicine*, 2018, 1–6. https://doi.org/10.1155/2018/9873273

Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., & 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. https://doi.org/10.1371/journal.pone.0229596

Yilmaz, R., Winkler-Schwartz, A., Mirchi, N., Reich, A., Christie, S., Tran, D. H., Ledwos, N., Fazlollahi, A. M., Santaguida, C., Sabbagh, A. J., & others (2022). Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. *NPJ Digital Medicine*, 5(1), 54–54. https://doi.org/10.1038/s41746-022-00596-8

Brown, K. C., Bhattacharyya, K. D., Kulason, S., Zia, A., & Jarc, A. (2020). How to Bring Surgery to the Next Level: Interpretable Skills Assessment in Robotic-Assisted Surgery. *Visceral Medicine*, 36(6), 463–470. https://doi.org/10.1159/000512437

Kowalewski, K.-F., Garrow, C. R., Schmidt, M. W., Benner, L., Müller-Stich, B. P., & 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. https://doi.org/10.1007/s00464-019-06667-4

Zhang, D., Wu, Z., Chen, J., Gao, A., Chen, X., Li, P., Wang, Z., Yang, G., Lo, B., & Yang, G.-Z. (2020). Automatic Microsurgical Skill Assessment Based on Cross-Domain Transfer Learning. *IEEE Robotics and Automation Letters*, 5(3), 4148–4155. https://doi.org/10.1109/LRA.2020.2989075

Lee, D., Yu, H. W., Kwon, H., Kong, H.-J., Lee, K. E., & Kim, H. C. (2020). Evaluation of Surgical Skills during Robotic Surgery by Deep Learning-Based Multiple Surgical Instrument Tracking in Training and Actual Operations. *Journal of Clinical Medicine*, 9(6), 1964. https://doi.org/10.3390/jcm9061964

Winkler-Schwartz, A., Yilmaz, R., Mirchi, N., Bissonnette, V., Ledwos, N., Siyar, S., Azarnoush, H., Karlik, B., & 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. https://doi.org/10.1001/jamanetworkopen.2019.8363

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14, 1611–1617. https://doi.org/10.1007/s11548-019-02039-4

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*, 149. https://doi.org/10.1016/j.compbiomed.2022.106043