Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation

Alexander Winkler-Schwartz, MD, * *Vincent Bissonnette, MD,* *,† *Nykan Mirchi, BSc,* * *Nirros Ponnudurai, BEng,* * *Recai Yilmaz, MD,* * *Nicole Ledwos, BA,* * *Samaneh Siyar, MSc,* *,‡ *Hamed Azarnoush, PhD,* *,‡ *Bekir Karlik, PhD,* * *and Rolando F. Del Maestro, MD, PhD* *

*Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, Quebec, Canada; †Division of Orthopedic Surgery, Montreal General Hospital, McGill University, Montreal, Quebec, Canada; and ‡Department of Biomedical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

OBJECTIVE: Virtual reality simulators track all movements and forces of simulated instruments, generating enormous datasets which can be further analyzed with machine learning algorithms. These advancements may increase the understanding, assessment and training of psychomotor performance. Consequently, the application of machine learning techniques to evaluate performance on virtual reality simulators has led to an increase in the volume and complexity of publications which bridge the fields of computer science, medicine, and education. Although all disciplines stand to gain from research in this field, important differences in reporting exist, limiting interdisciplinary communication and knowledge transfer. Thus, our objective was to develop a checklist to provide a general framework when reporting or analyzing studies involving virtual reality surgical simulation and machine learning algorithms. By including a total score as well as clear subsections of the checklist, authors and reviewers can both easily assess the overall quality and specific deficiencies of a manuscript.

DESIGN: The Machine Learning to Assess Surgical Expertise (MLASE) checklist was developed to help computer science, medicine, and education researchers ensure quality when producing and reviewing virtual reality manuscripts involving machine learning to assess surgical expertise.

SETTING: This study was carried out at the McGill Neurosurgical Simulation and Artificial Intelligence Learning Centre.

PARTICIPANTS: The authors applied the checklist to 12 articles using machine learning to assess surgical expertise in virtual reality simulation, obtained through a systematic literature review.

RESULTS: Important differences in reporting were found between medical and computer science journals. The medical journals proved stronger in discussion quality and weaker in areas related to study design. The opposite trends were observed in computer science journals.

CONCLUSIONS: This checklist will aid in narrowing the knowledge divide between computer science, medicine, and education: helping facilitate the burgeoning field of machine learning assisted surgical education. (J Surg Ed 76:1681–1690. © 2019 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: simulation, surgery, education, artificial intelligence, assessment, machine learning

Funding/Support: This work was supported by the Di Giovanni Foundation, the Montreal English School Board, the Montreal Neurological Institute and Hospital, and the McGill Department of Orthopedics.

Other Disclosures: Samaneh Siyar is a Visiting Scholar in the Neurosurgical Simulation and Artificial Intelligence Learning Centre. Dr. Azarnoush previously held the Postdoctoral Neuro-Oncology Fellowship from the Montreal Neurological Institute and Hospital and is a Visiting Professor in the Neurosurgical Simulation and Artificial Intelligence Learning Centre. Dr. Winkler-Schwartz holds a Robert Maudsley Fellowship from the Royal College of Physicians and Surgeons of Canada, Nicole Ledwos holds the Christian Gaeda Brain Tumor Scholarship from the Montreal Neurological Institute and Nirros Ponnudurai is supported by a Heffez Family Bursary. Dr. Del Maestro is the William Feindel Emeritus Professor in Neuro-Oncology at McGill University.

Ethical Approval: Not applicable.

Disclaimer: None.

Previous Presentations: None.

Co-first authors.

Correspondence: Inquiries to Alexander Winkler-Schwartz, MD, Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, 3801 University Street, Room E2.89, Montreal, Quebec, Canada; e-mail: manuscriptinquiry@gmail.com

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2019.05.015

COMPETENCIES: Medical Knowledge, Interpersonal and Communication Skills, Patient Care

INTRODUCTION

The assessment and training of the complex psychomotor skills necessary to perform surgical procedures is critical to safe patient outcomes. As such, virtual reality simulators are being utilized to understand, evaluate, and train these skills.[1] Simulation platforms allow for the quantification of multiple aspects of surgical performance in safe environments. The combination of virtual reality simulators and machine learning has the potential to significantly augment current methods of surgical training.

In computer science, machine learning is a subset of artificial intelligence utilizing algorithms (such as classifiers) which give computers the capacity to "learn" patterns when provided with data. Broadly speaking, classifiers can be either supervised or unsupervised. Supervised classifiers use data which has been identified by the researchers' a priori to generate predictive models to identify novel unlabeled data. In its simplest application in an educational context this implies identifying "expert" and "nonexpert" participant data, thus generating models capable of categorizing individuals into these groups and, ostensibly, assessing expertise. Supervised classifiers lend themselves well to circumstances where groups can be clearly defined. Unsupervised algorithms require no a priori data labeling. Please refer to Table 1 for the definitions of relevant terms.

Increasingly, the application of artificial intelligence techniques to evaluate performance on virtual reality simulators has led to an increase in the volume and complexity of publications which bridge the fields of computer science, medicine, and education. Although all disciplines stand to gain from research in this field, important differences in reporting exist, limiting interdisciplinary communication and knowledge transfer. A standardized approach in the reporting of these publications will allow researchers from these fields to form a better shared understanding of the burgeoning field

TABLE 1. Definitions in the Context of Artificial Intelligence and Machine Learning

Keyword	Definition
Artificial intelligence	Intelligence demonstrated by a machine able to make decisions in a manner similar to human intelligence.
Machine learning	A sub-branch of artificial intelligence where machines process data and learn on their own, without constant human supervision.
Metric	A measurement to quantitate performance.
Feature	Input data that is fed to the artificial intelligence algorithm.
Label	A determinant of the class to which a data point belongs to in the classification process. Usually applied to a dataset in the context of supervised learning. In the context of surgical simulation, an individual's data could be labelled as "expert" or "novice".
Classifier	A machine learning algorithm which sorts data into predefined categories.
Supervised machine learning	A type of machine learning algorithm where the machine learns patterns to make prediction after being trained with labelled data.
Unsupervised machine learning	A type of machine learning algorithm where the machine learns patterns from unlabelled data.
Algorithm	A set of rules provided to artificial intelligence that allows machines to perform certain tasks such as classification.
Model	A previously trained machine learning algorithm.
Overfitting	A condition which occurs when a model is too closely fitted to a particular set of data and cannot be reliably applied to a new dataset.
Accuracy $\frac{\Sigma\text{True Positive}+\Sigma\text{True Negative}}{\Sigma\text{Total population}}$	A measure of ability of machine learning to correctly classify new data.
Sensitivity $\frac{\Sigma\text{True Positive}}{\Sigma\text{True Positive}+\Sigma\text{False Negative}}$	A measure of how many positive condition predictions are actually true positives.
Specificity $\frac{\Sigma\text{True Negative}}{\Sigma\text{True Negative}+\Sigma\text{False Positive}}$	A measure of how many negative condition predictions are actually true negatives.

The definitions and formulas were obtained from *An introduction to machine learning.*[6]

of machine learning assisted surgical education. As such, our goal is to diminish this gap by producing a framework known as the Machine Learning to Assess Surgical Expertise (MLASE) checklist which researchers can utilize when producing and reviewing virtual reality manuscripts involving machine learning to assess surgical expertise. By including a total score as well as clear subsections of the checklist, authors and reviewers can both easily assess the overall quality and specific deficiencies of a manuscript. The framework complements existing guidelines for best practices in reporting experimental design in medical education.[2] To our knowledge, this is the first attempt to create a conceptual structure to ensure quality of virtual reality studies utilizing machine learning to assess surgical skills.

In the manuscript we outline the MLASE checklist, and apply it to publications obtained through a systematic literature review on the use of machine learning to assess surgical expertise in virtual reality simulation.

METHODS

MLASE Checklist

Upon consultation with interdisciplinary groups of physicians, computer scientists, engineers, and specialists in artificial intelligence, we developed the "Machine Learning to Assess Surgical Expertise" (MLASE) checklist comprised of 20 essential key elements when reporting studies using machine learning algorithms to assess technical skills in virtual reality surgical simulators. The key elements were divided into 4 sections: Study Design, Data Structure, Supervised Machine Learning and Discussion Quality (Table 2).

Study Design

This section contains 5 elements: Literature Review, Sample Size, Expertise Definition, Simulator Description and Simulated Tasks Description.

Literature Review. A relevant literature review on the previous use of similar machine learning algorithms to

TABLE 2. Machine Learning to Assess Surgical Expertise (MLASE) Checklist

Section	Element	Yes?
Study design (5 points)	1. Is relevant literature on the use of artificial intelligence in simulation provided?	
	2. Is the sample size clearly stated (including number of groups and number of participants in each group)?	
	3. Is a definition of each group of expertise provided?	
	4. Is the simulator described?	
	5. Are the surgical tasks to be performed outlined?	
Data structure (6 points)	6. Is raw data acquisition described?	
	7. Is feature extraction mentioned?	
	8. Is an effort made to normalize the data?	
	9. Is feature selection mentioned?	
	10. Is the count of features used by the algorithm clearly stated?	
	11. Are the final selected features clearly described?	
Supervised machine learning (5 points)	12. Is the type of the classifier used mentioned and justified (either by comparing multiple classifiers or citing relevant literature)?	
	13. Is the mechanism of the classifier explained or is a relevant source provided?	
	14. Is an effort made to clearly describe the methods used to train and test the algorithm?	
	15. Is the accuracy of the classifier mentioned?	
	16. Is the sensitivity and specificity mentioned?	
Discussion quality (4 points)	17. Are efforts made to explain the educational rationale of the features used by the algorithm?	
	18. Is the educational application of classifiers in the context of surgical simulation stated, specifically its use as a summative or formative assessment tool?	
	19. Are methodological limitations discussed, including those pertaining to any above-points?	
	20. Are the future directions discussed?	
Total Score = _____/20		

The checklist contains 20 elements, separated into 4 sections. A point is awarded for every element completed in the article. The total score is calculated by adding the total number of elements checked.

evaluate skill level should be presented. An effort should be made to situate the current manuscript in the context of previous publications.

Sample Size. The number of groups and participant numbers per group should be clearly stated. In virtual reality surgical education trials, it is often easier to recruit nonexpert (medical student and junior resident) rather than expert (physician consultant) members. As such, algorithms using a dataset obtained from such groups may be biased to incorrectly categorize a new expert participant. Furthermore, as with statistical tests, certain algorithms function poorly with little input data. Thus, the sample size must be appropriate for the algorithm used. *Potential pitfall:* Having unbalanced groups will skew the algorithms' predictive ability towards the largest group, limiting its future predictive ability. Having a small sample size may be inappropriate for the algorithm used.

Expertise Definition. When utilizing supervised algorithms, judgments concerning what constitutes "expert" performance create algorithms which recapitulate the human assumptions that underlie them. A clear definition of each group is critically important, specifically what constitutes an "expert" vs a "nonexpert". For example, the algorithm accuracy may differ substantially if first year medical students are considered novices, compared to third year residents. *Potential pitfall:* The outcome of a supervised algorithm classifying process will vary according to the researchers' definition of expertise.

Simulator Description. A description of the simulator hardware and software tools used, type of data recorded, and the experimental environment setting should be elaborated. If available, previous publications outlining the aforementioned items can be cited instead. *Potential pitfall:* Study reproducibility can only be achieved with a clear description of the simulator platform utilized.

Simulated Tasks Description. Due to the variety of simulated scenarios on a given virtual reality system, an adequate description of surgical task should be provided. *Potential pitfall:* A lack of clear description of the simulated task may impact study reproducibility, applicability, and pedagogical insights.

A broad overview of the following 3 sections can be found in Figure 1.

Data Structure

The Data Structure section contains 6 elements: Raw Data Acquisition, Feature Extraction, Data Normalization, Feature Selection, Count, and Description of Final Features selected.

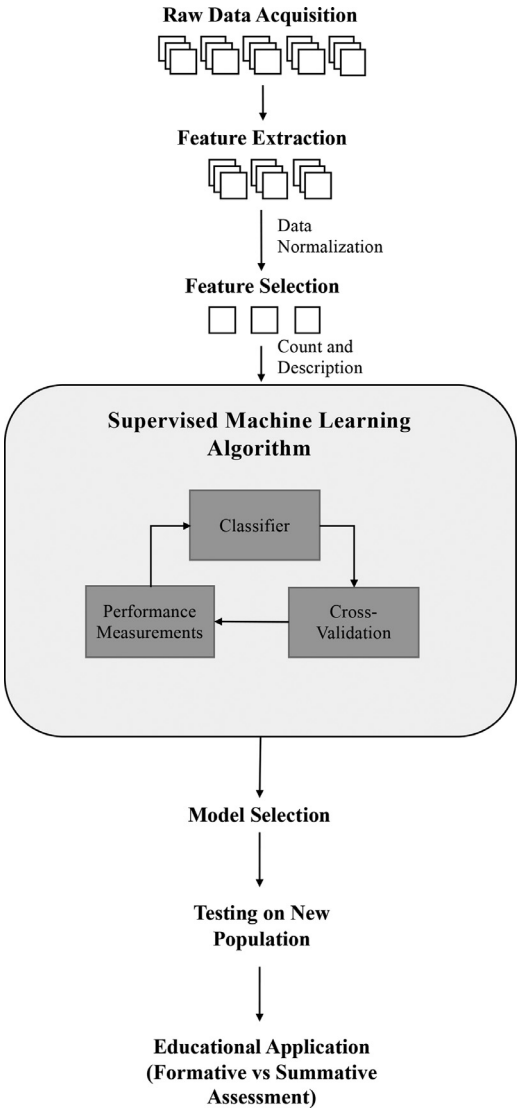

FIGURE 1. A broad overview of the application of machine learning technology in virtual reality surgical simulation according to the Machine Learning to Assess Surgical Expertise (MLASE).

Raw data acquired from the simulator is transformed into a format which can be inputted into the machine learning algorithm via feature extraction and selection. Following this, an iterative process involving cross-validation is utilized in which the machine learning classifier is optimized. Once a final model is selected it is retrained on the entire study dataset. After this, educational applications of the model can be tested in novel populations.

Raw Data Acquisition. The process of raw data acquisition should be briefly outlined. The most important information to provide is the fundamental structure of the data yielded by the simulator during a simulated task. Notable examples include positional data every second, and applied force vectors in 3 dimensions. *Potential pitfall:* As in any statistical test, general description of the nature of the data acquired is essential to best understand the functioning of the algorithm and the potential educational benefits.

Feature Extraction. Raw data from virtual reality simulators is often very complex, repetitive, and with varying degrees of 'signal to noise' ratio. Feature extraction is a method that reduces the dimensionality of a dataset by manipulating raw data, however this can be accomplished by many ways.[3] One can automatically reduce the dimensionality of data by using statistical procedures such as principal component analysis. Alternatively, data can be combined by experienced individuals to generate features in which there may be an a priori hypothesis in distinguishing between experts and novices, such as force applied close to a structure felt to be critical in an operation.[4] *Potential pitfall:* Failure to provide relevant input will force the algorithm to find patterns in features which may be irrelevant to surgical competency. This may, in addition to limiting the educational use of the model, negatively impact the accuracy and efficiency of the machine learning algorithm.

Data Normalization. Various features generated from feature extraction may be scaled differently, as such, feature normalization should be carried out before providing them as inputs into the algorithm. *Potential pitfall*: Failure to normalize data will result in diminished accuracy of the classification process.

Feature Selection. Feature selection is a method that highlights the most relevant features and eliminates those that are causing noise. Statistical methods can select only those features showing significant differences between groups (2 sample *t* test, for example), and are thus most likely to aid the algorithm classifying process. Numerous other feature selection techniques exist, however these are beyond the scope of this article.[5] *Potential pitfall:* Improper feature selection will negatively impact an algorithm's classification ability.

Count of Final Features Selected. It is of critical importance to include the final count of features used by the algorithm. Including an abundance of features may reduce the algorithms' predictive accuracy by adding noise (i.e., irrelevant information not helpful in the classification process) or by overfitting (a process in which an algorithm is able to detect small differences between groups on a study dataset at the expense of not capturing larger trends which are useful in classifying a novel dataset). *Potential pitfall:* If the number of final features is too large for a given sample size, the algorithm may appear to be extremely accurate using the study dataset, however its ability to make accurate predictions in a novel dataset may be compromised.

Description of Final Features Selected. We recognize that it may be impractical for authors to describe every final feature in detail if many were included in the final algorithm. However, efforts should be made to apply broad categories, such as features relating to force, movement, tissue removed, to name a few. *Potential pitfall:* Not including an adequate description of final features may miss interesting insights concerning surgical performance which may serve as the basis for trainee feedback.

Supervised Machine Learning

The Supervised Machine Learning section comprises 5 elements: Type of Classifier and Justification, Mechanism of the Classifier, Training and Testing Set, Accuracy, and Sensitivity/Specificity.

Type of Classifier and Justification. Various supervised machine learning classifiers, such as hidden Markov models, support vector machines, and artificial neural networks have been used to assess surgical expertise level in simulation studies.[4] Authors should not only state the type of classifier employed, but should also provide the rationale for their choice. Such justification can be provided by citing a relevant study using a classifier in a similar context. *Potential pitfall:* It is important to consider the variability of classifier performance depending on the surgical task. For instance, a classifier can accurately predict the expertise level in a laparoscopic surgery task but perform poorly in a brain tumor resection task. Therefore, an alternative would be to compare the performance of multiple classifiers and select the most accurate for a given task.

Mechanism of the Classifier. The manuscript should include a simple explanation with regards to how the machine learning algorithm works or refer the reader to a source that does so. *Potential pitfall:* Since artificial intelligence is a novel field in medicine, additional clarification may be necessary, thereby allowing the medical community to gain knowledge on this highly technical topic.

Training and Testing Set. In cases of supervised machine learning, training datasets consist of participant data where groups of expertise have been defined by the researchers. The algorithms' performance in a testing dataset will determine its ability to judge whether novel data will be classified as expert or nonexpert (or various gradations in between). Since this represents a crucial aspect of algorithm development, efforts should be made to clearly describe the process of training and testing. Two common methods are described.[6] If the sample size from each group of expertise is large enough, the sample can be divided in 2 subsamples where one is used for training and the other for testing. However, when the sample size is smaller, many different subsamples of training and testing sets can be used and averaged to obtain the accuracy. This process is known as cross-validation. Many cross-validation methods exist

and are beyond the scope of this publication.[6] *Potential pitfall:* Failure to provide a clear explanation of the training and testing sets does not allow the reader to understand and evaluate the methodology of the study. Ultimately, cross-validation is a technique used to estimate the accuracy of many models and select the one that is most likely to perform well on a new dataset. However, cross-validation is not an exact measurement of a model's accuracy in real-life application. Therefore, assumptions should not be made about the generalizability of a model that performs well in cross-validation.

Accuracy. Accuracy can be defined as the number of correct predictions made by the machine learning algorithm on all the predictions made (see Table 1). Accuracy is a key element because it evaluates the overall ability of the classifier to predict expertise level with a given set of features.

Sensitivity and Specificity. The engineering and medical literature differs based on their reporting of test success. Whereas the engineering community reports in terms of accuracy and equal error rates, these may be less intuitive to medical readers themselves familiar with sensitivity and specificity. For this reason, it is important to discuss sensitivity and specificity when reporting studies in medical journals. *Potential pitfall:* Authors should mention the percentage of experts and novices misclassified as it may assist readers in understanding whether the authors' conclusions for the use of the algorithm are justified. For example, a highly sensitive but poorly specific algorithm, namely one which misclassifies many nonexperts as expert, would be incompatible with its application as a summative assessment tool. If study design allows, another option is to present a full confusion matrix, which is similar in structure to a 2-by-2 table commonly used in medicine.[7]

Discussion Quality

The Discussion Quality section contains 4 elements: Educational Application of Machine Learning, Educational Rationale of the Selected Features, Methodological Limitations and Future Directions.

Educational Application of Machine Learning. Authors should clearly state the educational aim of their use of machine learning. Classifiers are designed to categorize data, thus lending themselves well as a summative assessment tool. As such, machine learning can be used as a summative assessment tool to evaluate a surgeon's performance. Although more challenging to execute in practice, machine learning can also be involved in formative assessment by facilitating feedback to help surgeons improve their skills. Both types of assessment have different requirements.[8] *Potential pitfall:* Summative assessment can have a drastic impact on surgeons' success, hence they require extremely high accuracy and reproducibility. On the other hand, formative assessment requires an understanding of the specific features used by the algorithm to help surgeons improve their technical skills.

Educational Rationale of the Selected Features. Authors should clearly describe why the chosen features are important in an educational context. *Potential pitfall:* Overly abstract features (such as eye movement) may serve well as summative assessments, however if the intended use is for a formative assessment then the chosen features must be easily teachable.

Methodological Limitations. Authors should always address the limitations of their study. Specifically, the shortcomings of the use of machine learning in surgical skill assessment should be outlined.

Future Directions. Future directions should be mentioned. This benefits the medical education community as it provides the reader with a clear understanding of how the field may continue to evolve.

Literature Review

In order to evaluate the current status of articles on the subject using our checklist, we performed a systematic review involving artificial intelligence or machine learning to distinguish experts and novices using virtual surgical simulators in the Medline, Embase, and Web of Science databases. Investigations were included if: (1) its purpose was to assess surgical skill, (2) employing a supervised machine learning algorithm, and (3) on tasks performed on a virtual reality simulator.

Two authors (V.B., N.M.) individually reviewed and scored each article using the MLASE checklist. The article was awarded 1 point for each element of the checklist. If differing article scores were present, an attempt was made by the 2 reviewers to come to a consensus. If none was obtained, then consensus was achieved with the remaining authors. Scores were compiled in a table and analyzed using descriptive statistics. Inter-rater reliability between the reviewers was calculated with Cohen's Kappa.

RESULTS

A total of 2642 articles were identified. Following review of abstracts and titles, 84 articles involving simulation and artificial intelligence or machine learning were assessed. A total of 54 articles were excluded as they did not involve

TABLE 3. Articles Assessed on the Use of Artificial Intelligence to Classify Expertise in Virtual Reality Surgical Simulation

Journal Category	Year Published	Classifier	Authors
Medical	2018	Naïve Bayes and support vector machines	Ershad et al.[9]
	2012	Decision tree	Kerwin et al.[10]
	2011	Hidden Markov models	Rhienmora et al.[11]
	2010	Linear discriminant analysis and artificial neural network	Richstone et al.[12]
	2010	Naïve Bayes, hidden Markov models and logistic regression	Sewell et al.[13]
	2005	Fuzzy	Huang et al.[14]
Engineering	2011	Support vector machines and hidden Markov models	Loukas et al.[15]
	2011	Hidden Markov models	Liang et al.[16]
	2011	Support vector machines and decision tree	Jog et al.[17]
	2007	Fuzzy	Hajshirmohammadi et al.[18]
	2006	Hidden Markov models	Megali et al.[19]
	2003	Hidden Markov models	Murphy et al.[20]

virtual reality surgical simulation. Of the remaining 30 articles, 21 were removed as they did not meet all the elements of the inclusion criteria. Three further articles were identified through manual searches of Google Scholar and Cochrane databases for a total of 12 articles.

These 12 articles[9-20] utilizing machine learning to assess surgical expertise in simulation were reviewed using the MLASE checklist (Table 3). Inter-rater reliability between the 2 reviewers was calculated with an observed agreement of 80% (Cohen's Kappa = 0.56). Six of the articles were published in medical and 6 in computer science or engineering journals. The results are summarized in Table 4 and Figure 2. The global average score for all articles was 73%. This can be further divided into sections where Study Design, Data Structure, Supervised Machine Learning, and Discussion Quality scored 78, 71, 75, and 67%, respectively.

The 3 lowest scoring elements were: explaining the educational rationale of the selected features (element 17, 5/12, 42% articles) explaining the methodological limitations (element 19, 5/12, 42% articles), normalization of the data (element 9, 6/12, 50% articles), and mentioning the specificity and sensitivity of the algorithm (element 16, 6/12, 50% articles).

We also analyzed articles based on their journal category. Articles from medical and engineering journals both scored 73% overall. Medical articles scored lowest in Data Structure (69%) and Supervised Machine Learning (70%) and highest in Discussion Quality (79%) whereas engineering articles scored lowest in Discussion Quality (54%) and highest in Supervised Machine Learning (80%) and Study Design (80%).

DISCUSSION

Our results indicate that a conceptual framework has the potential of improving the quality of future manuscripts. Though our checklist was tested on articles using machine learning assessing surgical expertise employing virtual reality simulation, we believe the MLASE checklist is also applicable to benchtop simulators, augmented reality, or any other studies which digitize physical surgical performance and use machine learning methods to assess surgical expertise.

TABLE 4. Results of Assessment of Articles Using the Machine Learning to Assess Surgical Expertise Checklist

	All	Journal Type Medical	Engineering
MLASE section score	Score/percentage mean (max – min)	Score/percentage mean (max – min)	Score/percentage mean (max – min)
Study design	3.92(5 – 2)/78(100 – 40)	3.83(5 – 2)/77(100 – 40)	4.00(5 – 2)/80(100 – 40)
Data structure	4.25(6 – 2)/71(100 – 33)	4.17(6 – 2)/69(100 – 33)	4.33(6 – 2)/72(100 – 33)
Supervised machine learning	3.75(5 – 2)/75(100 – 40)	3.5(5 – 2)/70(100 – 40)	4.00(5 – 3)/80(100 – 60)
Discussion	2.67(4 – 1)/67(100 – 25)	3.17(4 – 2)/79(100 – 50)	2.17(4 – 1)/54(100 – 25)
Overall	14.58(18 – 11)/73(90 – 55)	14.67(18 – 11)/73(90 – 55)	14.50(17 – 12)/73(85 – 60)

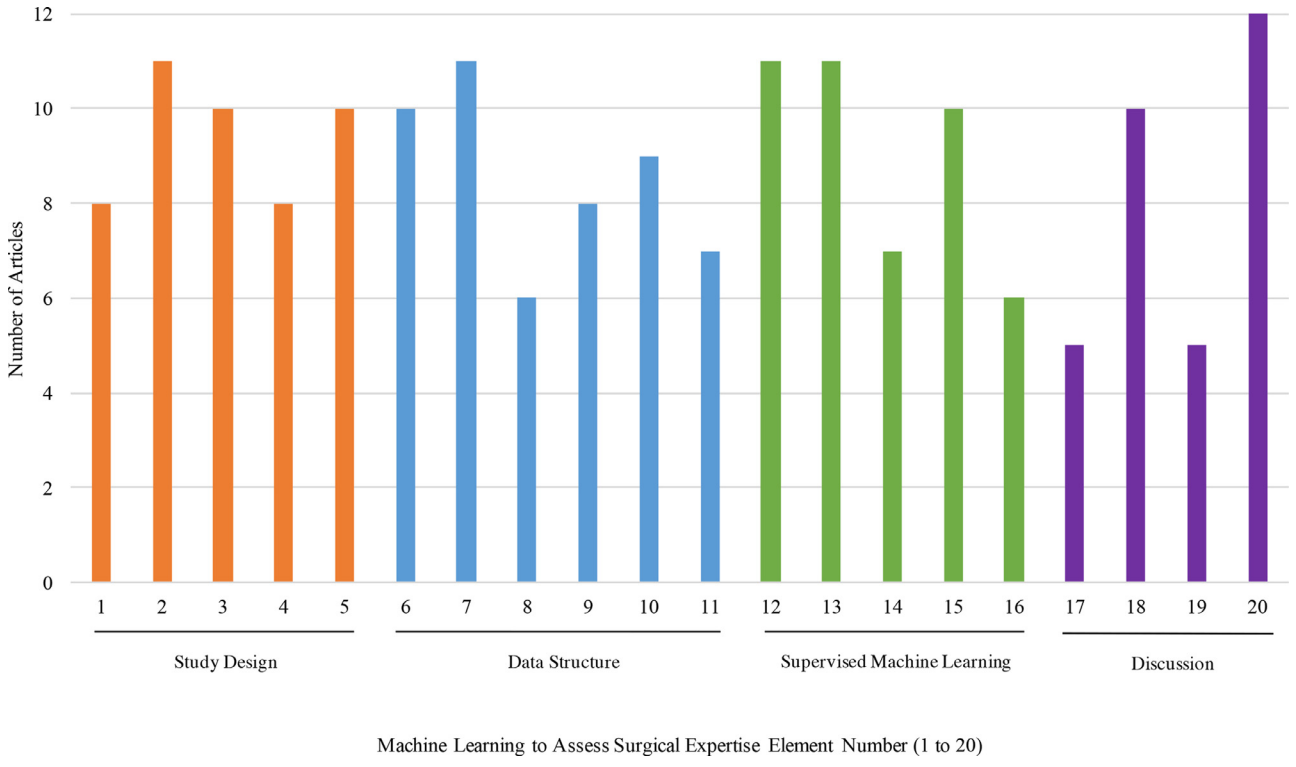

FIGURE 2. The authors applied the Machine Learning to Assess Surgical Expertise (MLASE) checklist to 12 articles on obtained from a systematic review involving artificial intelligence or machine learning to distinguish experts and novices using virtual reality surgical simulators.

Although manuscripts published in medical and computer science journals received, on average, the same overall MLASE total score, important differences in the subsections were noted. We identified the Discussion Quality section of the MLASE checklist as one which will require the most attention from computer scientists wishing to publish in the field of medicine. In medical journals, more detail is required in the Data Structure and Supervised Machine Learning section. The MLASE checklist makes it possible for researchers from these differing communities to ensure their publications reach the widest possible audience. Furthermore, this manuscript may serve as a guide for journal editors and reviewers to ensure that best practices in applying machine learning methodologies in a surgical-simulation context are adhered to. As such, improvements in reporting practices will ultimately facilitate interdisciplinary communication and knowledge transfer, helping to advance this field of research.

Further Suggestions for Future Authors and Reviewers to Enhance the Quality of Manuscripts

Following our article review, we identified new elements which may further enhance the quality of future manuscripts. Firstly, some studies[18,19] attempt to increase their sample size by allowing the same surgeon to perform a procedure several times. When such methods are used, it is crucial to explain how each trial is used in the analysis. Often, explanations are vague and it is unclear if different trials from the same surgeon were part of both, the training and testing sample. This would lead to overfitting of the algorithm as performance measures extracted from different trials of the same surgeon are highly correlated. This may hinder an algorithm's ability to accurately classify a new participant. Secondly, if sample size permits, having a third dataset excluded from the initial testing and training to run the chosen model may yield information regarding its generalizability. Thirdly, as an increasingly holistic understanding of expertise continues to be developed (i.e., one which is not based solely on the number of years of practices or on the number of procedures completed), supervised algorithms' predictive abilities will continue to improve. Finally, there are potential educational benefits in describing the individuals that were misclassified by the algorithm, particularly if the same participant is misclassified by different algorithms.

Limitations

The objective of the MLASE checklist is to provide a general framework when reporting or analyzing these studies

in the future. However, we acknowledge that the checklist is not extensive and further elements can be added to enhance the quality of a study. The checklist only presents the 20 elements deemed essential to bridge the knowledge gaps in different communities studying the use of artificial intelligence in surgical education. The MLASE checklist was designed and evaluated using only supervised machine learning articles. The MLASE checklist can be applied to studies utilizing unsupervised learning algorithms, however these algorithms do not necessarily always require feature extraction and feature selection.

Future Directions

Artificial intelligence systems may be developed to not only classify participants according to surgical expertise but to coach trainees to a defined surgical standard. These systems will allow for the conduct of studies to further elaborate the proper approach in using this technology in the teaching of psychomotor skills. Regardless of what the future holds, a clear understanding of surgery, artificial intelligence methodologies, and educational best practices will be crucial to the ultimate success of these systems.

CONCLUSIONS

The MLASE checklist was developed to help computer science, medical, and education researchers ensure quality when producing and reviewing virtual reality manuscripts involving the use of machine learning to assess surgical expertise in virtual reality simulation. We believe our checklist will narrow the knowledge divide between computer science, medicine, and education helping facilitate the burgeoning field of machine learning assisted surgical education.

ACKNOWLEDGMENT

The authors wish to thank Drs. Greg Berry and Jean Ouellet from the Division of Orthopaedic Surgery, Montreal General Hospital, McGill University, Canada for their input and Alex Amar, librarian at the Montreal Neurological Insitute and Hospital for their thorough search strategy and availability.

REFERENCES

1. Gallagher AG, Ritter EM, Champion H, et al. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Ann Surg*. 2005;241:364–372. https://doi.org/10.1097/01.sla.0000151982.85062.80.

2. Cook DA, Beckman TJ, Bordage G. Quality of reporting of experimental studies in medical education: a systematic review. *Med Educ*. 2007;41:737–745. https://doi.org/10.1111/j.1365-2923.2007.02777.x.

3. Ding S, Zhu H, Jia W, Su C. A survey on feature extraction for pattern recognition. *Artif Intell Rev*. 2012;37:169–180. https://doi.org/10.1007/s10462-011-9225-y.

4. Vedula SS, Ishii M, Hager GD. Objective assessment of surgical technical skill and competency in the operating room. *Annu Rev Biomed Eng*. 2017;19:301–325. https://doi.org/10.1146/annurev-bioeng-071516-044435.

5. Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electr Eng*. 2014;40:16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

6. Kubat M. An Introduction to Machine Learning. Switzerland: Springer International Publishing; 2015. https://doi.org/10.1007/978-3-319-63913-0.

7. Watson RA. Use of a machine learning algorithm to classify expertise : analysis of hand motion patterns during a simulated surgical task. *Acad Med*. 2014;89:1163–1167. https://doi.org/10.1097/ACM.0000000000000316.

8. Chauvin SW. Applying educational theory to simulation-based training and assessment in surgery. *Surg Clin North Am*. 2015;95:695–715. https://doi.org/10.1016/j.suc.2015.04.006.

9. Ershad M, Rege R, Fey AM. Meaningful assessment of robotic surgical style using the wisdom of crowds. *Int J Comput Assist Radiol Surg*. 2018;13:1037–1048. https://doi.org/10.1007/s11548-018-1738-2.

10. Kerwin T, Wiet G, Stredney D, Shen HW. Automatic scoring of virtual mastoidectomies using expert examples. *Int J Comput Assist Radiol Surg*. 2012;7:1–11. https://doi.org/10.1007/s11548-011-0566-4.

11. Rhienmora P, Haddawy P, Suebnukarn S, Dailey MN. Intelligent dental training simulator with objective skill assessment and feedback. *Artif Intell Med*. 2011;52:115–121. https://doi.org/10.1016/j.artmed.2011.04.003.

12. Richstone L, Schwartz MJ, Seideman C, Cadeddu J, Marshall S, Kavoussi LR. Eye metrics as an objective assessment of surgical skill. *Ann Surg*. 2010;252:177–182. https://doi.org/10.1097/SLA.0b013e3181e464fb.

13. Sewell C, Morris D, Blevins NH, et al. Providing metrics and performance feedback in a surgical

simulator. *Comput Aided Surg*. 2008;13:63–81. https://doi.org/10.1080/10929080801957712.

14. Huang J, Payandeh S, Doris P, Hajshirmohammadi I. Fuzzy classification: towards evaluating performance on a surgical simulator. *Stud Health Technol Inform*. 2005;111:194–200 http://ebooks.iospress.nl/publication/10064. Accessed August 23, 2018.

15. Loukas C, Georgiou E. Multivariate autoregressive modeling of hand kinematics for laparoscopic skills assessment of surgical trainees. *IEEE Trans Biomed Eng*. 2011;58:3289–3297. https://doi.org/10.1109/TBME.2011.2167324.

16. Liang H, Shi MY. Surgical skill evaluation model for virtual surgical training. *Appl Mech Mater*. 2011;40-41:812–819. https://doi.org/10.4028/www.scientific.net/AMM.40-41.812.

17. Jog A, Itkowitz B, Liu M, et al. Towards integrating task information in skills assessment for dexterous tasks in surgery and simulation. In: Proceedings—IEEE International Conference on Robotics and Automation; 2011. p. 5273–5278. https://doi.org/10.1109/ICRA.2011.5979967.

18. Hajshirmohammadi I, Payandeh S. Fuzzy set theory for performance evaluation in a surgical simulator. *Presence-Teleoperators Virtual Environ*. 2007;16:603–622 https://www.mitpressjournals.org/doi/pdf/10.1162/pres.16.6.603. Accessed August 23, 2018.

19. Megali G, Sinigaglia S, Tonet O, Dario P. Modelling and evaluation of surgical performance using hidden Markov models. *IEEE Trans Biomed Eng*. 2006;53:1911–1919. https://doi.org/10.1109/TBME.2006.881784.

20. Murphy TE, Vignes CM, Yuh DD, Okamura AM. Automatic motion recognition and skill evaluation for dynamic tasks. In: Paper presented at: Eurohaptics 2003 Conference; July 6-9; 2003. (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.8109&rep=rep1&type=pdf).

SUPPLEMENTARY INFORMATION

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.jsurg.2019.05.015.