ARTIFICIAL INTELLIGENCE DISTINGUISHES SURGICAL TRAINING LEVELS IN A VIRTUAL REALITY SPINAL TASK

Vincent Bissonnette, MD

Experimental Surgery, Surgical Education Concentration

McGill University, Montreal

July, 2019

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

Table of Contents

ABSTRACT

Background

With the emergence of competency-based training, the current evaluation scheme of surgical skills is evolving to include newer methods of assessment and training. The large amount of data collected from an individual's performance during virtual reality simulated tasks can be distilled into intuitive metrics. Since surgical procedures involve multiple psychomotor skills, effective assessment of surgical expertise is more appropriately realized through systems capable of revealing the complex relationships between multiple metrics. Artificial intelligence through machine learning algorithms can utilize extensive datasets to analyze operator performance. This study aims to address three questions, (1) Can artificial intelligence uncover novel metrics of surgical performance? (2) Can support vector machine algorithms be trained to differentiate "Senior" and "Junior" participants executing a virtual reality hemilaminectomy? (3) Can other algorithms achieve a good classification performance?

Methods

Participants from four Canadian universities were divided in two groups according to their training level (senior and junior) and were asked to perform a virtual reality hemilaminectomy. The position, angle and force application of the simulated burr and suction along with tissue volumes removed were recorded at twenty millisecond intervals. Raw data was manipulated to create metrics to train machine learning algorithms. Five algorithms, including a support vector machine, were trained to predict whether the task was performed by a senior or junior participant. The accuracy of each algorithm was assessed through leave-one-out cross-validation.

Results

Forty-one individuals were enrolled, 22 senior and 19 junior participants. Twelve metrics related to safety of the procedure, efficiency, motion of the tools and coordination were selected. Following cross-validation, the support vector machine achieved a 97.6% accuracy. The other algorithms achieved 92.7, 87.8, 70.7 and 65.9% accuracy, respectively.

Conclusion

Artificial intelligence defined novel metrics of surgical performance and outlined training levels in a virtual reality spinal simulation procedure.

Clinical Relevance

The significance of these results lies in the potential of artificial intelligence to compliment current educational paradigms and better prepare residents for patient procedures.

RÉSUMÉ

Introduction

Avec l'émergence de l'approche par compétences pour les résidences chirurgicales, les méthodes d'évaluation d'habiletés chirurgicales actuelles sont sujettes à évoluer pour inclure de nouvelles méthodes d'enseignement et de formation. L'importante quantité de données collectées lorsqu'un individu performe une opération simulée en réalité virtuelle peut être traduite en mesures de performance. Étant donné que les procédures chirurgicales incorporent plusieurs habiletés psychomotrices, l'évaluation de l'expertise chirurgicale est plus facilement réalisable en utilisant des systèmes capable de révéler les liens complexes entre plusieurs mesures de performance. L'intelligence artificielle à l'aide d'algorithme d'apprentissage machine peut utiliser de grandes quantité de données pour analyser la performance d'un individu. Cette étude a pour but de répondre à trois questions : (1) Est-ce que l'intelligence artificielle peut aider à découvrir de nouvelles mesures de performances chirurgicales? (2) Est-ce que les algorithmes de machines à vecteurs de support peuvent être entraîner à différencier des participants « senior » et « junior » qui exécute une hémi-laminectomie en réalité virtuelle? (3) Est-ce que d'autres algorithmes peuvent classifier ces participants adéquatement?

Méthodologie

Des participants de 4 universités canadiennes ont été divisé en 2 groupes en tenant compte de leur niveau de formation (senior et junior) et ont exécuté une hémi-laminectomie en réalité virtuelle. Les positions, les angles et la force appliquée par le « burr » et la succion ainsi que les volumes de tous les tissus retirés ont été enregistrés à toutes les 20 millisecondes. Les données brutes ont été manipulées pour créer des mesures de performance pour entraîner des algorithmes d'apprentissage

machine. Cinq algorithmes, incluant une machine à vecteurs de support, ont été entraînés pour prédire si la simulation avait été performée par un participant senior ou junior. La précision de chaque algorithme a été évalué par validation croisée.

Résultats

Quarante et un participants ont été recrutés, 22 senior et 19 junior. Douze mesures de performance concernant la sécurité de la procédure, l'efficacité, les mouvements des instruments et la coordination ont été sélectionnées. L'algorithme de machine à vecteurs de support a identifié correctement le niveau d'expertise (Junior ou Senior) à 97,6% lors de la validation croisée. Les autres ont identifié correctement à 92,7%, 87,8%, 70,7% et 65,9%, respectivement.

Conclusion

L'intelligence artificielle a aidé à définir de nouvelles mesures de performance chirurgicales et a adéquatement identifié le niveau de formation de participants dans une procédure de chirurgie spinale simulée en réalité virtuelle.

Impact clinique

Ces résultats démontrent le potentiel de l'utilisation de l'intelligence artificielle pour complimenter le curriculum chirurgical actuel afin de mieux préparer les résidents à effectuer des opérations sur de vrais patients.

ACKNOWLEDGMENTS

This work could not have been done without the precious support of numerous individuals.

First and foremost, I would like to thank my supervisor, Dr. Rolando Del Maestro, and my co-supervisor, Dr. Greg K. Berry, for giving me the opportunity to work with an incredible team at the Neurosurgical Simulation and Artificial Intelligence Learning Centre.

I would like to thank Dr. Del Maestro for his continuous guidance and expertise with respect to surgical education, surgical simulation and artificial intelligence. His daily advices regarding my master's project and life, as well as his passion for the history of medicine and art, have made this journey a life-learning experience. Through his knowledge and teaching skills, Dr. Del Maestro has given me a passion for surgical education –a passion that, I hope, will make me a better surgeon in the future.

I would also like to thank Dr. Berry, my co-supervisor, for his outstanding support. His great knowledge of surgical education and orthopaedic surgery research were extremely valuable. Dr. Berry was always available to answer my questions and gave me every tool to succeed. The enrollment of orthopaedic surgery residents and staff in the study could not have been done without his precious help. Dr. Berry also provided the funds for my master's project through the McGill Division of Orthopaedic Surgery.

I would like thank everyone that has been involved with the Neurosurgical Simulation and Artificial Intelligence Learning Centre, Nicole Ledwos for her support and help in the data

collection and writing process, Recai Yilmaz for his help with machine learning and code functions, Alexander Winkler-Schwartz for his great expertise in surgical education and metrics of surgical performance, Ghusn Alsidieri for his help in the data collection and development of the spinal scenario used in this study, Dr. Émile Lemoine, Samaneh Siyar, Dr. Bekir Karlik and Dr. Hamed Azarnoush for their expertise and useful inputs with respect to the machine learning methodology used in this study, Robin Sawaya, Dr. Fahad Alotaibi, Dr. Abdulgadir Bugdadi and Dr. Khalid Bajunaid for their help in the initial development of the spinal scenario used in this study, Nirros Ponnudurai for his help with the literature review, Dr. Jean Ouellet, Dr. Mohammad Maleki and Dr. Michael Weber for their precious help in identifying metrics that quantify performance in spine surgery.

A special thanks to Nykan Mirchi, a fellow master's student and co-first author of my manuscript, for his expertise in machine learning, his coding skills, and his help in the conceptualisation and writing of the manuscript. I wish him the best of luck with medical school at University of Toronto.

I would also like to thank the McGill's Division of Orthopaedic Surgery and the Franco Di Giovanni Foundation for the funding of my master's project, the team of the National Research Council Canada for the development of the NeuroTouch virtual reality platform used in this study, and Alex Amar for conducting a systematic search of manuscripts involving the use of machine learning to assess surgical expertise in simulation.

Finally, I would like to thank each and every one of the medical students, residents and consultant surgeons that participated in the study and made this thesis possible.

CONTRIBUTION OF AUTHORS

Nykan Mirchi was a co-first author on the manuscript utilized in this thesis. He helped with the conceptualisation of the project, data collection, data analysis, and writing of the manuscript. Nykan also helped with the review of literature and conceptualisation of the Machine Learning to Assess Surgical Expertise (MLASE) checklist. Nicole Ledwos helped with the collection of the data and writing of the manuscript. Ghusn Alsidieri helped with the collection of the data and development of the simulated task. Alexander Winkler-Schwartz helped with the data analysis, writing of the manuscript, and conceptualisation of the MLASE checklist. Recai Yilmaz helped with the data analysis, writing of the manuscript and the MLASE checklist. Samaneh Siyar helped with the conceptualization of the data analysis and the MLASE checklist. Hamed Azarnoush helped with the conceptualization of the data analysis and the MLASE checklist. Bekir Karlik helped with the conceptualization of the data analysis and the MLASE checklist. Robin Sawaya helped with the collection of the data and development of the simulated task. Fahad E Alotaibi helped with the collection of the data and development of the simulated task. Abdulgadir Bugdadi helped with the collection of the data and development of the simulated task. Khalid Bajunaid helped with the collection of the data and development of the simulated task. Nirros Ponnudurai helped with the literature review. Jean Ouellet helped with collection of the data, and writing of the manuscript. Greg K. Berry helped with collection of the data, and writing of the manuscript. Rolando F. Del Maestro helped with collection of the data, data analysis and writing of the manuscript.

I was involved with the MLASE checklist conceptualization, the systematic review of the literature, the conceptualization of this study, the data collection, the data analysis, writing the manuscript, correspondence with the journal and writing this thesis.

INTRODUCTION OF THESIS

The Surgical Education Paradigm Shift

For more than a century, surgical training has been inspired by Halsted and Osler's apprenticeship model, whereby residents learn through direct patient exposure supervised by expert surgeons with increased responsibilities according to the year of their residency training.[1] While this model has been widely used, it may not be optimal in terms of patient safety, residents' learning experience and costs. From a safety perspective, although residents are closely monitored by their supervisors in the operating room, allowing them to practice their skills on real patients raises ethical questions. Every patient deserves to be operated on by experienced surgeons less prone to make mistakes. From an economic standpoint, the apprenticeship model is expensive. Allen et al. reported that residents involvement in cases increased operating time by roughly 4.8 minutes which they calculated leads to an approximate annual cost of $492,889 at their institution.[2] From an educational point of view, since residents have seen their work-hour restricted in an effort to improve their wellness, they have less time to "learn by doing" as the traditional paradigm suggests.[3] Moreover, the number and complexity of case exposure for residents may vary considerably from one institution to another, leading to discrepancies in surgical training. Finally, given the ageing of the population, procedures with a higher degree of complexity will be increasingly performed. In spine surgery, Deyo et al. outlined a 15-fold increase in the rate of complex lumbar fusions in a span of 5 years.[4] This could result in less exposure for residents as their supervisor may not be keen to let them operate on these complex cases.

Given these limitations and mounting pressures from governments, insurance companies and the public seeking for improved patient outcomes at a reduced cost, medical entities have deemed it important to reassess the current scheme of surgical education. As such, a surgical education paradigm shift has been underway, seeking new methods of training and assessment. In Canada, this shift led to the emergence of competency by design curricula. [5]

In the previous paradigm, residents were assumed competent once they completed their five or six-year specialty training and successfully passed a series of examinations. With the shift towards competency-based curricula (referred to as competency by design in Canada), residents will need to prove their competence in a number of tasks (called entrustable professional activities) before graduating.[5] The requirement to effectively demonstrate competency at every stage of the training will lead to an important rise in the number of assessments, which may increase the workload of academic physicians and the cost of training. Technical surgical skill assessments such as direct observations, global rating scales and checklists are often performed by practising surgeon evaluators. Surgeon evaluators are costly and often lack the time and educational skills required to carry out periodic assessments of complex technical skills. This may lead to inconsistent, inadequate and delayed feedback of an individual's performance. In addition, human evaluators are prone to subjectivity. The need for more objective, automated, reliable, and cost-effective assessments has led to an increased interest in technologies capable of quantifying skill in surgical specialties.

Simulation-Based Training

Simulation-based training is increasingly utilized and studied in surgery.[6] Simulation is viewed as a useful adjunct to surgical training for many reasons. First, simulation gives trainees the opportunity to practice repeatedly in a safe environment. This allows them to enhance their skills prior to operating on patients which has the potential to lead to safer and more efficient surgical procedures with better patient outcomes. Second, simulation could improve residents learning by allowing for the decomposition of complex surgical procedures into many steps which gives residents the opportunity to perform deliberate practice.[7] Third, simulation provides the opportunity for standardization of training across programs. This is particularly interesting in the context of spine surgery in which surgeons come from two different surgical specialties (Orthopaedic Surgery and Neurosurgery) with very different training paradigms and surgical exposure.

Simulation-based training can be provided in many forms. Animal models, cadavers and benchtop models have all been utilized to teach surgical residents.[7–9] In spine surgery, these types of models have been employed to simulate anterior cervical discectomy and fusion, posterior foraminectomy and laminectomy, dural tear repair and pedical screw placement. [10–13]

Virtual Reality Simulators

With the advancement in computing power and graphics, virtual reality simulators have recently emerged as potential training platforms for surgical procedures. In spine surgery, many virtual reality platforms have been developed to simulate percutaneous vertebroplasty, pedicle screw placement/insertion and laminectomy.[14] While procedures on these platforms occur entirely in a virtual field, haptic feedback devices can be used to simulate the different tactile sensation an operator perceives when interacting with diverse tissues.[15] A unique attribute of virtual reality platforms is their ability to collect an enormous amount of data that quantifies multiple components of psychomotor performance during surgical procedures. Standards of reference to quantitate performance and progress (known as metrics) can thus be generated. In neurosurgery, metrics of performance have been extensively studied utilizing NeuroVR (CAE, Montreal, Canada) formerly known as NeuroTouch (National Research Council Canada, Bourcherville, Canada).[16–18] Since these metrics quantitate performance, they could potentially be employed to create proficiency benchmarks to assess surgical skill.[19] In fact, our group has demonstrated that metrics extracted from virtual reality simulators can better capture some components of the surgical performance than visual rating scales.[20] While individual metrics provide useful information on specific components of surgical skills, combining multiple metrics has the potential to provide a more holistic understanding of performance. The need to generate and analyze multiple metrics from a large amount of data has led to an increased interest in computer science techniques.

Artificial Intelligence and Machine Learning

Introduced in the 1950's, artificial intelligence is a concept that aims to give computers human problem solving skills.[21] Machine learning is a subtype of artificial intelligence whereby algorithms search for patterns in the data that allow computers to make decisions or predictions with no necessity of explicit instructions.[22–24] Machine learning can be broadly subdivided into supervised, unsupervised and reinforcement learning.[25] Supervised learning is utilized when the ground truth is known. In supervised learning, many examples belonging to different categories are provided to the algorithm along with data that corresponds to these examples. The idea is that the algorithm "learns" to identify patterns in the examples associated with each category.[26–28] This process is called the training of the algorithm. Once trained, the algorithm is tested.[28] Data from new examples is provided to the algorithm and it needs to decide to which category the new examples belong. For instance, in the surgical education context, the algorithm could be trained with data from many known expert and novice surgeons. Once trained, the algorithm could decide to which group (expert or novice) data from new individuals belongs. In unsupervised learning, the ground truth is unknown–the category to which an example belongs is a priori unknown. Given this, the algorithm has the task to analyze hidden patterns in the data and regroup examples that have similar patterns into categories.[25,27,29] In a surgical education context, we would provide the data of many surgeons, without a priori labelling them as expert or novices. The algorithm would then regroup individuals with similar patterns in their data. In reinforcement learning, the algorithm is put in an environment in which it takes actions and "learns" from trial and error based on a reward system.[25] In a surgical education context, the algorithm would be asked to decide whether one individual is a novice or an expert surgeon

without prior examples. Then, the algorithm would be "rewarded" if its answer is correct. This process results in decision accuracy improvements and can be repeated indefinitely. Although both unsupervised and reinforcement learning could be promising techniques for educational purposes, supervised learning techniques have been more widely utilized to assess physicians technical skills.[30,31]

Hypothesis and objectives

In this investigation, we aimed to test the combination of artificial intelligence methodology and virtual reality simulation as a potential objective assessment tool for surgical technical skills. We hypothesized that traditional machine learning algorithms could distinguish between two groups of different training levels in a virtual reality hemilaminectomy simulation. Three specific objectives were outlined to test this hypothesis:

1) to develop new complex metrics to quantitate psychomotor performance in a virtual reality hemilaminectomy simulated scenario.

2) with a selection of these metrics of performance, to train a support vector machine algorithm to classify participants performing a virtual reality hemilaminectomy as Senior (PGY-4 and above) or Junior (PGY-3 and below) and test its accuracy using cross-validation.

3) to train other traditional machine learning algorithms to perform the same task with the same metrics of performance and evaluate their accuracies using cross-validation.

The rationale of the simulated task

With population aging, the prevalence of back pain and the number of spinal procedures being performed have been rising.[32–34] In parallel, new surgical techniques and instrumentation have paved the way for the performance of more complex spine procedures. While spine procedures are performed as part of the training of both orthopaedic surgeons and neurosurgeons, there remain differences in terms of exposure and educational paradigms in each residency program. In fact, it has been reported that orthopaedic surgery residents have much less exposure to spine surgery compared their neurosurgery counterparts.[35,36] Given increased numbers and complexity of cases along with the lack of standardization in training, spine surgery could particularly benefit from the use of virtual reality surgical simulators to help residents improve their skill level.

The use of a burr and suction are involved in most spinal procedures and are basic skills residents need to acquire. One of the numerous surgical procedures simulated on the NeuroVR is a left L3 hemilaminectomy. The scenario involves the use of a simulated burr and simulated suction to remove part of the L3 lamina without damaging surrounding tissues. Alsidieri et al. provided an initial evaluation of the face, content and construct validity of this scenario.[37] The thirteen novel metrics of performance identified by the authors did not adequately differentiate between junior residents, senior residents and spine surgeons and thus may not be optimal to teach residents how to improve their skills. Four limitations could explain the lack of difference in scores amongst the three different levels of expertise. First, their sample size was relatively small. Second, the metrics they identified may not have been complex enough to identify specific differences amongst the groups. Third, the authors looked at the difference amongst three groups (junior residents, senior residents and surgeons). However, this simulated task may be too simple to differentiate between

senior residents and surgeons. Fourth, while surgery requires the interaction of multiple skills, metrics of performance were not combined. This may have prevented a holistic evaluation of skill level that may have differentiated between the groups more accurately. In this study, we increased our sample size, built more complex metrics of performance, assessed the differences between only two groups of different training levels and used machine learning methodology to combine multiple metrics of performance.

Machine learning methodology overview

As previously mentioned, supervised learning techniques require pre-identified categories or classes. If the supervised algorithm is trained to recognize one specific category, then it is said to be a one-class algorithm. An example of one-class algorithm would be to train a software to recognize images of dogs from a multitude of photographs by finding the attributes (features) that makes this animal a dog. If the supervised algorithm is trained to distinguish between two or more categories, then it is said to be a two-class or multi-class algorithm. Here the algorithm finds the inherent differences between the two or more categories. An example of multi-class algorithm would be to train a software to recognize and categorize series of photographs as cats, dogs or horses.

Another important concept to understand is the difference between traditional machine learning algorithms and the more novel deep learning algorithms. Whereas traditional machine learning algorithms required a lot of steps to select the inherent attributes of a category or the features that help distinguish between two or more categories, deep learning algorithms do not require this type pre-processing. However, deep learning algorithms typically work better with extremely large datasets and their decision-making process is much more complex to understand.[27] Our small sample size and desire to easily interpret the algorithms decision-making process led us to use traditional learning algorithms in this experiment. As such, a series of steps needed to be performed to transform the raw data into a handful of important features (metrics) that could be used to train traditional machine algorithms to distinguish between the two categories

identified a priori. These steps can include the extraction, normalization and selection of features or metrics. [38]

Feature extraction

When a large and multivariate dataset is acquired, it can be complex, repetitive and made up of noisy data.[26] Many techniques can be used to extract relevant and minimally repetitive features (metrics) from a large dataset.[38] Statistical methods may be employed to extract features that seem to differentiate between the two categories and eliminate noisy and repetitive data. Feature extraction can also be performed by hand, through combination of the raw data. For instance, by recording time and multiple positions of a tool, the average velocity of this tool can be extracted.

Feature normalization

In addition to the potential of recording repetitive and noisy information during the data acquisition process, the multiple features (metrics) extracted from the data are frequently on different scales of measurement. To effectively analyze, combine and compare these features, one must put them on the same scale.[38] This process of feature normalization can be done by utilizing z-scores for each feature (i.e. the difference between one's score and the mean of all scores divided by the standard deviation of all scores).

Feature selection

The feature selection step ensures to find the optimal combinations features upon which the algorithms can be trained to perform a specific task. When the task required from the algorithm is to distinguish between two different categories, researchers can hand-pick features that have

the potential to discriminate between the two categories based on their expertise or insights from previous studies. Feature selection algorithms may also be used during this step. Two commonly utilized category of feature selection algorithms are the forward and backward feature selection algorithms. The forward selection algorithms begin with one feature, test the accuracy of a machine learning algorithm at discriminating two or more categories based on that one feature, then adds features subsequently, one-by-one, until optimal accuracy is achieved. The backward selection algorithms begin with all features that were extracted from a dataset, test the accuracy of a machine learning algorithm at discriminating two or more categories based on the combination of all features, then removes features subsequently one-by-one, until optimal accuracy is achieved. These feature selection algorithms help find combinations of features (or metrics) that can best differentiate between the two or more pre-defined categories.

Machine learning algorithms

A multitude of traditional machine learning algorithms exist. Some of the most commonly reported in the medical education literature are the support vector machines, the k-nearest neighbors, linear discriminant analysis, naive bayes and decision trees. [31]

Support vector machines

Support vector machines find a linear (or non-linear) decision surface named hyperplane that can separate data points associated with individuals from two different categories and maximize the margins between this decision surface and the closest data points to this decision surface (support vectors).[23,26] In the context of this study, participants from a senior group and a junior group can be divided by an hyperplane based on a specific combination of metrics of performance.

K-nearest neighbors

The k-nearest neighbors algorithm predicts the category to which a data point belong by using distance functions to determine the closest neighbors to this data point in a multidimensional space (based on multiple variables).[25,26] The category of a participant can thus be determined based on the relationship with the nearest participants' category. A parameter (k) corresponds to the number of neighbors considered. In the context of this study, a participant can be classified as belonging to the junior or senior group based on the category to which belong the closest data points in a multivariable space (according to several metrics of performance). For instance, if, out of the 7 closest neighbors (k=7), 5 of these are labelled to the senior group, then the data point may be classified as belonging to the senior group.

Linear discriminant analysis

Linear discriminant analysis algorithms project multidimensional data (multiple metrics) on a single dimension and maximizes the distance between the means of the data points belonging to the different groups while minimizing the variance within each group.[23] In the context of this study, the data points associated with the junior group would be regrouped to minimize the variance within each other, while also maximizing the distance from the mean of the junior group data points and the mean of the senior group data points.

Naive bayes

The naive bayes algorithms utilize probabilities to predict the group to which an individual data point should belong.[23,26] In the context of this study, the algorithm is trained with multiple examples of scores on a series of metrics associated with a label (junior or senior). This process

allows the algorithm to learn the probability of a score on a certain metric being associated with a certain label (junior or senior). The algorithm then evaluates the probability associated with the scores on every metric to predict whether a participant belongs to the junior or senior group.

Decision trees

Decision trees classify individuals by building a series of nodes whereby participants are divided according to the value of a certain metric.[26] The algorithm is trained with multiple examples of values on a series of metrics associated with a label (junior or senior). This process helps the algorithm find the optimal values to divide participants in classes.

Algorithms' performance

Once algorithms are trained, their accuracies may be tested with new data points from an independent dataset.[28] Ideally, the algorithm would be trained with several junior and senior participants, then tested on new individuals. However, given the proof of concept nature of this manuscript and the relatively small sample size, cross-validation was utilized to train and test the algorithms. The algorithms' performance may be measured in multiple ways. Given the fact that the medical community is used to terminology such as accuracy, sensitivity and specificity, these results were displayed in a confusion matrix. The confusion matrix shows the number of senior classified as senior, junior classified as junior, senior misclassified as junior, junior misclassified as senior, and the number of correct predictions overall. The methods utilized in this experiment are explained more thoroughly in the methods section of the manuscript.

Comprehensive Review of the Literature: The Use of Machine Learning to Provide Objective Assessment of Surgical Skills

A systematic review was performed by our group to search for current articles that utilized supervised machine learning methodology to objectively assess surgical skill level in virtual reality simulation.[38] A total of 2642 articles were identified with a search through the Medline, Embase and Web of Science databases. After screening the titles and abstracts, 84 articles were identified involving the application of machine learning to assess surgical skill level using simulation technologies. Of these, 9 articles employed supervised machine learning to assess surgical expertise utilizing virtual reality simulators. Three other articles were found by manually searching on Google Scholar and Cochrane databases.

In 2003, Murphy et al. trained hidden markov models to recognize skill level by analyzing operator motions in a basic laparoscopic virtual task of moving a ball towards a target.[39] While the task was very simple and the sample size limited, they reported an accuracy of up to 81.31% for classifying expertise. Huang et al. used number of errors, economy of movement and time to train fuzzy algorithms to classify individuals performing a basic laparoscopic virtual task on the MIST-VR simulator (Wolfson Centre and VR Solutions, United Kingdom) into 3 groups of different skill level.[40] Their results were inconclusive due to the very limited sample size. Megali et al. trained hidden markov models to analyze kinematic data and predict whether new individuals were experts or novices with respect to laparoscopic instrument handling in virtual reality.[41] They described that one of the four individuals defined as novice was closer to the expert group in terms of motion skills. Again, their sample size was limited (2 experts and 4 novices) and the task

was very basic (touch a sphere with right and left instrument). In 2007, Hajshirmohammadi et al.[42] did a follow-up study of Huang et al.[40] study. They trained fuzzy algorithms to classify between novices, intermediates and experts on a suturing and knot-tying scenario on the MIST-VR platform. They reported 29 to 44% of correct predictions utilizing number of errors, tissue deformation, thread overstretch and time as metrics of performance. In 2010, Sewell et al. used naïve bayes, hidden markov models and logistic regression to classify experts and novices performing a virtual mastoidectomy achieving 87.5% accuracy.[43] In 2010, Richstone et al. used eye metrics to train linear discriminant analysis algorithms and neural networks to predict whether surgeons performing on the LapMentor system (Simbionix, Cleveland, Ohio) were experts or non-experts.[44] These algorithms achieved 91.9% and 92.9% accuracy, respectively. In 2011, Jog et al. used support vector machine and decision tree algorithms to classify novices and experts performing a virtual robotic surgical task.[45] They achieved an accuracy of 87.5% based on motion data. Loukas et al. used support vector machines to distinguish between 11 novices (PGY-2 and PGY-3) and 11 intermediates (PGY-5 and PGY-6) performing needle driving and intracorporal knot tying on the LapVR platform (CAE Healthcare, Montreal, Canada).[46] They achieved sensitivity/specificity ranging from 86% to 96%. Liang et al. used hidden markov models to perform binary classification of expertise level.[47] They achieved 85% accuracy by looking at force and position of tools. Rhienmora et al. reported a 100% accuracy with five-fold cross-validation utilizing hidden markov models to distinguish between experts and novices performing on a virtual reality dental simulator.[48] In 2012, Kerwin et al. used decision trees to classify individuals performing a virtual mastoidectomy into experts and non-experts by looking at the end product of the procedure.[49] They described accuracies ranging from 45% to 89%. In 2018, Ershad et al. used positional, motion and physiological data to train naïve bayes and support vector machines to

distinguish between 4 levels of expertise (expert, fellow, intermediate or novice) in a basic virtual robotic surgery task.[50] They reported accuracies ranging from 75% to 89% using cross-validation techniques.

These articles were all analyzed utilizing the Machine Learning to Assess Surgical Expertise (MLASE) checklist. The MLASE checklist was developed by a group of engineers, specialists in artificial intelligence and physicians interested in surgical education, with the aim to improve the quality of reporting studies involving the assessment of surgical skill in virtual reality with machine learning methodology. The checklist is comprised of 20 elements divided into 4 components: Study Design, Data Structure, Supervised Machine Learning and Discussion.

Six of the elements were reported in less than 60% of the studies. In the data structure, only a few authors (n=6) described normalizing metrics to put every metric on the same scale. This is a crucial step prior to using machine learning algorithms. In addition, a small proportion (n=7) described the metrics they used to train algorithms to recognize skill level. Failure to do so does not allow readers to understand and critique the machine learning algorithms' decision-making process. From an educational perspective, it is important for residents to know on which metrics they are assessed so they can improve. Only seven of the twelve manuscripts provided a clear explanation of their algorithms' training and testing process. This step is one of the most important aspects of research involving machine learning as it allows readers to understand and critique the authors' methods. If the authors used cross-validation or hold out validation techniques, it should be explicitly mentioned and assumptions on the generalizability of the models generated should not be made. While most of the authors reported their algorithms' accuracy, only six included

sensitivity and specificity. These two measures give a better understanding of the misclassifications made by the algorithms. This step is especially important in studies involving unequal numbers of individuals in the groups of different expertise. For example, in an hypothetical study involving 100 participants (10 experts and 90 non-experts), one could report a 90% accuracy if every participant gets classified as a non-expert. This very good result does not represent the algorithms' poor sensitivity (0% capability of identifying experts correctly). Merely 42% of the articles (n=5) described the educational rationale of their metrics (i.e. how these metrics would be utilized to train surgical residents). If the algorithm is to be used for formative assessments to help surgical residents improve, it is important for the algorithm to make its decisions based on metrics that are intuitive and teachable. Finally, only five articles included a description of the limitations of their machine learning methodology.

The following study was carried out and written with the intention to incorporate every key element from the MLASE checklist.

MANUSCRIPT

Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task

Vincent Bissonnette, MD *(1,2), Nykan Mirchi, BSc *(1), Nicole Ledwos, BA(1), Ghusn Alsidieri, MD, MSc(1), Alexander Winkler-Schwartz, MD(1), Rolando F. Del Maestro, MD, PhD(1) on behalf of the Neurosurgical Simulation & Artificial Intelligence Learning Centre† *co-first authors

This study was conducted at the Neurosurgical Simulation & Artificial Intelligence Learning Centre, (formerly Neurosurgical Simulation Research and Training Centre), Montreal Neurological Institute and Hospital, McGill University.

1. Neurosurgical Simulation & Artificial Intelligence Learning Centre, Department of Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, Canada.
2. Division of Orthopaedic Surgery, Montreal General Hospital, McGill University, Montreal, Canada.

Corresponding author:

Vincent Bissonnette

Neurosurgical Simulation & Artificial Intelligence Learning Centre, Department of Neurology & Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, 3801 University Street, Room E2.89, Montreal, Quebec, Canada, H3A 2B4

Email: vincent.bissonnette@mail.mcgill.ca Phone: 514-893-8606

† Recai Yilmaz, MD, Samaneh Siyar, MSc, Hamed Azarnoush, PhD, Bekir Karlik, PhD, Robin Sawaya, MSc, Fahad E Alotaibi, MD, MSc, Abdulgadir Bugdadi, MD, MSc, Khalid Bajunaid, MD, MSc, MMgmt, Jean Ouellet, MD and Greg Berry, MD, MSEd.

This manuscript has been published by the Journal of Bone and Joint Surgery.
(Date of publication: December 4th, 2019)
J Bone Joint Surg Am. 2019;101:e127(1-8)

INTRODUCTION

With the shift toward competency-based curricula, surgical educational paradigms are evolving to include new methods of assessment and training. Whereas current assessments rely on subjective methods, new technologies offer the potential for more objective approaches to surgical skill evaluation.[51] Simulation has become important in surgical education with many programs implementing courses involving animal models, cadavers, benchtop models and virtual reality simulators.[7] Virtual reality simulators provide opportunities for repeat practice in risk-free environments and can quantify multiple aspects of psychomotor performance during surgical procedures.[52] The large amount of data collected from an individual's technical performance during a simulated task can be distilled into specific metrics. Metrics can be considered standards of reference to quantitate performance, efficiency and progress.[17,53] Individual metrics are often incapable of effectively assessing surgical expertise since many procedures involve multiple complex psychomotor skills. The requirement to efficiently combine multiple metrics has resulted in the need to assess systems capable of analyzing extensive amounts of information from multivariate datasets.

Artificial intelligence employs machine learning algorithms, giving computers the ability to identify patterns and perform tasks without explicit programming when sufficient data is provided.[21,26] Different types of machine learning algorithms exist. Supervised algorithms, including support vector machines, are utilized most commonly. These algorithms are trained with examples of labelled data and learn patterns associated with each label, giving them the ability to label new data. [26] In surgical simulation, supervised algorithms could be trained utilizing sets of

metrics labelled as Senior or Junior, thereby allowing them to classify new individuals' metrics as Senior or Junior. This is referred to as two-class learning. One-class learning (training algorithms to identify individuals belonging to one group, e.g. experts) and multi-class learning (training algorithms to classify individuals in more than two groups, e.g. junior residents, senior residents, and staff surgeons) could also be employed but would require large participants numbers in each group to adequately train the algorithms. As such, these techniques have not been widely utilized to assess psychomotor skills in this context.[30] The purpose of this study was to evaluate the potential of artificial intelligence as an assessment tool in virtual reality spine surgery simulation. We aimed to provide a preliminary proof of concept that could act to introduce artificial intelligence as a mechanism to objectively assess surgical skill level. We addressed three questions in this investigation: (1) Can artificial intelligence uncover novel metrics of surgical performance that differentiate between two groups of different training levels? (2) Can support vector machine algorithms be trained to recognize whether an individual executing a virtual reality hemilaminectomy is of senior or junior level? (3) Can other algorithms achieve a good classification performance (accuracy above 75%)?

MATERIAL AND METHODS

Spine surgeons, spine fellows, orthopaedic and neurosurgery residents, and medical students from four Canadian universities were recruited.

As this investigation aimed to provide an initial proof of concept of the utility of machine learning as an assessment tool, we employed simple two-class learning algorithms. Thus, two groups of different expertise level had to be a priori defined. Participants were divided into senior (PGY-4

and above) and junior (PGY-3 and below) groups because our group of surgeons considered that the procedure required basic burr and suction handling skills that should be acquired by the fourth year of orthopaedic and neurosurgery training.

All participants signed a consent approved by McGill University Health Center Research Board before entering the study. The NeuroVR (CAE Healthcare, Montreal, Canada) virtual reality platform which incorporates a microscopic view and haptic feedback was employed to perform a left L3 hemilaminectomy.[54] This platform includes numerous simulated surgical scenarios which have been extensively studied.[18,55–57]As demonstrated in Video 1, the virtual hemilaminectomy required participants to remove the L3 lamina with a simulated burr in their dominant hand while controlling bleeding with a simulated suction in their non-dominant hand (Figs. 1-A, 1-B, and 1-C). Participants were given verbal and written instructions to remove the L3 lamina without damaging surrounding tissues. Subjects had five minutes to complete the task since this was found to be adequate in preliminary studies. Each participant performed the task once without prior practice. Individuals participated in the trial at a single time point without follow up**.** The trial was conducted in an experimental setting void of distractions.

Artificial intelligence methodology was applied through a series of steps including raw data acquisition, metric extraction, metric normalization, metric selection, machine learning algorithms and model selection (Fig. 2). These methods follow guidelines to utilize machine learning algorithms to assess surgical expertise in simulation previously established by our group.[38]

Raw data acquisition

The position, angle and force of both simulated instruments along with the removed volume of all simulated tissues were captured at twenty millisecond intervals and exported to a file.

Metric extraction

A metric is an input used to train a machine learning algorithm to predict whether a participant belongs to the senior or junior group. The accuracy of an algorithm can be defined as the number of good predictions out of the total number of predictions made. To obtain the best accuracy and to reduce computational cost, metrics given to algorithms must be carefully processed.[58]

The raw variables provided by the NeuroVR can be combined to generate more complex metrics. For instance, by combining tooltip position and time, velocity can be assessed. A series of functions was developed to extract metrics from the raw data using MATLAB R2018a (Natick, MA, USA). Metrics were divided into four categories including safety, efficiency, coordination and motion.[17,57] Since metrics of varying scales were generated, data normalization was performed with z-scores.

Metric selection

Metric selection is an important step in machine learning which attempts to find the combination of metrics that most accurately differentiates between the two groups.[59] This step is vital to prevent the algorithm from receiving irrelevant input, thereby avoiding the training of algorithms that are too closely "fitted" to a specific dataset and tend to generalize poorly to new subjects.[60]

Here, metric selection was performed in two parts. First, to capture metrics that are clinically relevant, two spine surgeons selected metrics they felt could differentiate between the two groups through a questionnaire (Appendix A). Second, since these metrics may not all adequately discriminate between the two groups in this scenario, a backward selection algorithm from PRtools (http://prtools.org/) was employed. This backward algorithm started with all the metrics chosen by spine surgeons and removed them sequentially while iteratively training a machine learning algorithm and testing its accuracy using 10-fold cross-validation.[59] The backward algorithm stopped when a combination of metrics provided the highest accuracy of classifying senior individuals as senior and junior individuals as junior. Metrics that were not selected were not further analyzed.

Machine learning algorithms

Support vector machines are suited for small sample size and multivariate data necessary for global evaluation of surgical skill, thereby making it a prime candidate for virtual reality surgical simulation.[26,60,61] Furthermore, their decision-making process is explainable. In a manner similar to the coefficients in a linear logistic regression, these algorithms attribute a weight to each metrics and make their classification based on an equation that considers every metric and their respective weights. This is interesting from an educational perspective because it could help juniors understand what they need to improve to achieve the senior level. These factors lead us to focus on this algorithm. Four other algorithms (k-nearest neighbors, linear discriminant analysis, naive bayes and decision tree) were also trained to assess whether the selected metrics could achieve a similar accuracy with diverse classification methods. The mechanism of each algorithm is explained in Table I. Additional information is available in the literature.[25,26,60–63]

Since our sample size was relatively limited, leave-one-out cross-validation was employed to train and test the algorithms.[25] Leave-one-out cross-validation trains the algorithm with all but one of the participants and subsequently tests the trained algorithm on the one participant left out of the training set. This process is repeated with every participant, hence, in our case, the process was repeated 41 times. As algorithms are built according to various parameters, these were adjusted in an iterative manner to optimize classification accuracy.

Metric Analysis

To analyze the performance of senior and junior participants, the ratio of the average metric score for senior and junior participants (fold difference) was calculated for each metric.

SOURCE OF FUNDING

This work was supported by the Di Giovanni Foundation, the Montreal Neurological Institute and Hospital and the McGill Department of Orthopaedics.

RESULTS

Twenty-two senior participants (6 spine surgeons, 3 spine fellows and 13 senior residents) and 19 junior participants (11 junior residents and 8 medical students) were recruited. The distribution of the participants' training level and specialty is presented in Table II. Number of assisted laminectomy cases by each resident are outlined in Table III. Forty-one metrics were generated. Of these, 36 metrics were selected by spine surgeons and are presented in Table IV. The backward algorithm identified twelve final metrics listed in Table V. Eight metrics relate to tool motions and

four relate to safety, efficiency and coordination. The maximum force applied on dura is lower in the senior group (fold difference: 0.56). The amount of time spent while using simultaneously the burr and suction was higher for senior participants (fold difference: 1.73). The senior participants touched adjacent structures more with their suction while removing L3 with the burr (fold difference: 2.18). The ratio of the amount of time spent removing L3 on the total time of the procedure was similar in both groups (fold difference: 0.96). Finally, senior participants displayed slower deceleration overall, showed higher delays between two consecutive accelerations while removing L3 and exhibited less variance in the pitch angle of the burr when they remove L3.

Using leave-one-out cross-validation, five algorithms were assessed. The support vector machine achieved the highest accuracy at 97.6%. The k-nearest neighbors, linear discriminant analysis, decision tree and naive bayes, had 92.7, 87.8, 70.7 and 65.9% accuracy, respectively (Fig. 3).

A confusion matrix was produced for the support vector machine algorithm (Fig. 4). Only one junior surgeon was misclassified.

DISCUSSION

Machine learning algorithms have defined novel metrics of surgical performance in a virtual reality spinal task. This addresses our first research question.

The four areas of surgical skill identified were represented in the twelve metrics selected. From a safety perspective, the senior group restricted the force applied on the dura. This is an important metric to teach considering that applying high forces on the dura may increase the risk of dural

tear. Senior participants also used their tools simultaneously more often than the junior participants. This shows the importance of the acquisition of bimanual skills in spine surgery. Furthermore, senior participants displayed less angle variance with the burr when removing L3 and higher delays between two acceleration peaks which provides new insights on the consistency of their movements. These results support that surgical skill is multifaceted and may be benefitted by teaching based on metrics that embody different aspects of surgical performance.

An automated feedback system was created with these metrics. Future participants will be able to see their scores on each of the metrics, as well as a global classification of the surgical training level (junior or senior). In addition, they will individually be guided to improve their skills through video-based and auditory feedback, which attempts to mimic current training in the operating room whereby surgeons explain what to improve and demonstrate how to do it.

We addressed the second question by training a support vector machine algorithm with twelve metrics to classify senior and junior participants performing a virtual reality spine procedure. The advantage of applying machine learning to our multivariate dataset is that it provides a more objective and holistic assessment of psychomotor performance.

As a support vector criterion was employed to select metrics, the final metrics were likely to best perform with support vector machine algorithms. To evaluate the ability of these metrics to differentiate training level, other algorithms were trained with the same metrics. Two other algorithms displayed accuracies above 75%, thereby addressing the third research question outlined.

The subject misclassified by the support vector machine algorithm was a PGY-2. Although we cannot be certain that this misclassification is attributed to a higher set of skills, we analyzed this individual's metrics to understand this result. This individual applied less force on the dura, spent more time using both tools simultaneously and displayed more consistency with the burr (less variance in pitch angle and larger distance between two acceleration peaks) than juniors. These results suggest that this individual's performance was more consistent with the expected performance of the senior group.

Participants were from multiple institutions and two specialties (neurosurgery and orthopaedics) making this data more representative of different training paradigms. Incorporating residents from both specialties allows the platform to have the potential to improve the standardization of spine training. However, this study is only an initial step to incorporate these technologies in residency training. It only acts as a proof of concept, and generalizability testing in a new population is required to ensure the algorithm is not overfitted and to evaluate the platform's potentials in training. This algorithm was trained according to residency training levels without explicit knowledge of surgical skill and has yet to be tested on an independent dataset. Thus, it cannot be used to certify the proficiency of residents prior to practice independently, nor can it assess surgical skill level with certainty, but it may help with psychomotor skills acquisition.

There are limitations to employing machine learning in this simulated procedure. First, a simulated burr and sucker are not representative of the many instruments and bimanual psychomotor skills employed during spine operations. Second, the visual and haptic complexities of the simulated

procedure, task duration and need to use a microscopic view may not adequately discriminate operator performance. More complex and realistic scenarios involving use of multiple instruments are presently being studied to address these issues. Third, although participants were asked to remove only the lamina, the lamina was not segmented separately from the spinous process and facets. Therefore, the volume of lamina removed could not be determined. The new spine scenarios being developed are designed to segment all surrounding structures. Defining participants' surgical skill level is difficult.[64,65] Numbers of surgical cases assisted are often biased when reported by residents and may not reflect the skills acquired throughout their residency.[66] It was implied that senior residents had acquire the basic skills of using a burr and suction. Since spine training varies from one program to another and PGY-4 are in a pivotal year in terms of surgical skills acquisition, efforts were made to understand whether the PGY-4 individuals should be included in the senior group. Thus, the study was repeated without incorporating PGY-4. The support vector machine algorithm achieved a 100% accuracy with 10 metrics, 6 of which are part of the 12 final metrics previously described. This is consistent with the concept that psychomotor skills of PGY-4 in this study are more aligned with the senior group. However, assessment tools, such as the Objective Structured Assessment of Technical Skill, to evaluate the residents' skills a-priori may help to provide a better division of groups in the future.[67] Furthermore, if large numbers of spine surgeons are recruited, one-class learning could be used to train algorithms to recognize expert performances and assess participants according to expert standards. This could provide a more robust evaluation of trainees' technical skill level.

To our knowledge, this is the first investigation employing machine learning to assess surgical expertise in a virtual reality spine procedure. Methods outlined in this study could be applied to

any surgical simulation scenario, provided that data on individual's performance is collected. As virtual reality simulation becomes more realistic and more widely utilized, algorithms will become more robust. One could envision that once algorithms are rigorously validated to recognize expert surgeons, surgical accreditation bodies could employ these techniques to ensure their members' technical competency. The significance of this study lies in the potential of combining virtual reality simulation and artificial intelligence to provide safer training and objective assessment of surgical skills, which could lead to improved patient care.

DISCUSSION OF THESIS

The importance of this manuscript lies in the potential of utilizing artificial intelligence in the context of virtual reality simulation to improve spine surgical training. The 12 metrics of performance identified through machine learning methodology quantify safety, efficiency, coordination and motion during an individual's performance in a virtual reality hemilaminectomy scenario. These metrics have been utilized to create an automated feedback platform that allows participants to assess their standing in terms of each of these metrics as compared to the senior group as well as a global evaluation of their skill level. Such platforms could allow for regular formative assessment of surgical skills and self-guided learning. In addition, this system could help residents perform deliberate practice and target the specific psychomotor skills to improve their surgical performance. In the scenario studied, the 12 performance metrics differentiate between Senior (PGY-4 and above) and Junior (PGY-3 and below). Junior residents could perform the virtual hemilaminectomy, compare their performance to benchmarks established by more senior individuals score with respect to specific surgical skills and repeat the virtual procedure as needed to improve their performance.

While the accuracy of the support vector machine algorithm reported in this study is encouraging, it is important to understand that this study constitutes a preliminary proof of concept. The high performance of the algorithm may be attributed to the small sample size of the study or the relative simplicity of the task performed by the machine learning algorithms (dividing a junior group of inexperienced trainees from a more senior group). 14 of the 41 participants were on the

extreme spectrum of both groups (medical students and attending spine surgeons), this considerable difference in expertise may be simple to establish by machine learning algorithms. However, differentiating between the PGY-3 and PGY-4 may be more complex. If there were less differences between the participants' levels of training, the algorithms' may have been less accurate in differentiating between the two groups. In addition, the final metrics were selected using a backward metric selection algorithm with the goal to optimize the support vector machine algorithms' accuracy. Although efforts were made to reduce the risk of overfitting—notably by reducing the number of metrics, the algorithms were not tested on an independent dataset. Thus, the results presented in this study represent an average of multiple algorithm models (a total of 41 models trained on 40 participants and tested on 1 participant), and may not be an exact representation of the accuracy of the model to predict new participants' level of expertise. Future studies could aim to test the proposed algorithm model on a dataset of new participants to evaluate better its generalizability.

With advancements in virtual reality and haptics technologies, simulated procedures will become more realistic and new metrics of performance that discriminate between individuals of different expertise levels may be identified. Moreover, while the simulated scenario presented in this study required basic burr and suction handling skills, more complex spine surgery scenarios are currently being developed. This may lead to better differentiation of multiple groups of different expertise level. Our group, utilizing a complex virtual reality neurosurgical procedure, has been able to segregate individual technical performance into 4 levels of expertise demonstrating that

algorithms may be capable of classifying surgical expertise with greater granularity and precision than has been previously demonstrated in surgery. [68]

Besides its use in formative assessments, artificial intelligence methodology has the potential to help provide summative assessments. An example would be to train algorithms to recognize individuals that have mastered certain skills and use these to evaluate residents before allowing them in the operating room. Such a system could improve the safety of surgical procedures. In the future, the Royal College of Surgeons could also potentially utilized artificial intelligence methodology combined with virtual reality surgical simulators to objectively assess their members' psychomotor skills and proficiency at the end of residency training.

While the study presented in this thesis focused on psychomotor skills, it is evident that surgeons should not solely be evaluated on their technical abilities. Cognitive skills, judgement, ability to cope with acute stress and social skills (i.e. leadership, teamwork and communication skills) are all essential to becoming a good surgeon. As technologies improve, new systems may give us the ability to collect more objective data on these other critical skills, which could potentially allow for the training of machine learning algorithms that take into account multiple aspects of surgical competency.

Future directions

While the future may involve the incorporation of virtual reality platforms utilizing artificial intelligence methodology to assess surgical skills, further studies are required to ensure these systems perform adequately. First, the models (trained algorithms) outlined in this study will need to be validated with completely independent test datasets incorporating new individuals from multiple institutions not previously seen by the algorithms. This will evaluate the generalizability of the models and the performance metrics utilized. If the models do not accurately predict training levels of individuals from the testing set, adjustments will need to be made. These could include increasing the training sample size (recruiting more participants), investigating new methods to select performance metrics (using other selection algorithms) and using new definitions of expertise level to split the groups a priori. New definitions of level of expertise may require the utilization of global rating scales of residents performing laminectomies in the operating room or using crowd-sourcing evaluation systems. Other machine learning algorithms not employed in this study–such as artificial neural networks and deep learning methodologies, could also be assessed to evaluate their capacity of distinguishing groups of different expertise level. A study is currently being performed comparing support vector machines to artificial neural networks in the scenario investigated in this thesis. Second, studies evaluating the optimal feedback platforms which lead to improved learning curves will need to be defined. Third, if the feedback platform improves surgical skill on a virtual reality platform, the transferability of these skills in the operating room will need to be confirmed. Once all of these questions are answered and algorithms as well as virtual reality simulated scenarios

are rigorously validated, these systems may be implemented into surgical residency curricula, providing new learning and assessment opportunities.

CONCLUSION OF THESIS

Summary

In this investigation, three objectives were outlined to test the hypothesis that machine learning could help differentiate between two groups of different training levels in a virtual reality hemilaminectomy simulation. First, new complex metrics of performance were generated to quantitate psychomotor performance of individuals during a spine surgery simulation task. Second, a support vector machine algorithm was trained to predict whether individuals performing a virtual reality hemilaminectomy on the NeuroVR (CAE, Montreal, Canada) belonged to a Senior or Junior group. Utilizing leave-one-out cross-validation and 12 metrics of performance quantifying safety, efficiency, coordination, and motion, a 97.6 % accuracy was achieved. Third, four other machine learning algorithms were also trained to perform the same task utilizing the same metrics of performance and two of these achieved accuracies above 75% using leave-one-out cross-validation. This suggests that machine learning algorithms have the potential to be trained to recognize expertise levels of individuals performing a virtual reality hemilaminectomy. The importance of these findings lies in the potential of combining artificial intelligence methodology and virtual reality simulators to provide an objective assessment of surgical skills in a safe environment, thereby giving surgical trainees an opportunity to improve their surgical skills before operating on patients.

REFERENCES OF THESIS

1. Franzese CB, Stringer SP. The Evolution of Surgical Training: Perspectives on Educational Models from the Past to the Future. *Otolaryngol Clin North Am*. 2007. doi:10.1016/j.otc.2007.07.004
2. Allen RW, Pruitt M, Taaffe KM. Effect of Resident Involvement on Operative Time and Operating Room Staffing Costs. *J Surg Educ*. 2016;73(6):979-985. doi:10.1016/J.JSURG.2016.05.014
3. Damadi A, Davis AT, Saxe A, Apelgren K. ACGME Duty-Hour Restrictions Decrease Resident Operative Volume: A 5-Year Comparison at an ACGME-Accredited University General Surgery Residency. *J Surg Educ*. 2007;64(5):256-259. doi:10.1016/J.JSURG.2007.07.008
4. Deyo RA, Martin BI, Kreuter W, Goodman DC, Jarvik JG. Trends , Major Medical Complications , and Charges Associated With Surgery for Lumbar Spinal Stenosis in Older Adults. 2010;303(13).
5. Royal College of Physicians and Surgeons of Canada. Competence by Design (CBD): What you need to know A Resident ' s Guide. 2017:1-4.
6. Borgersen NJ, Naur TMH, Sørensen SMD, et al. Gathering Validity Evidence for Surgical Simulation: A Systematic Review. *Ann Surg*. 2018;267(6):1063-1068. doi:10.1097/SLA.0000000000002652
7. Reznick RK, MacRae H. Teaching Surgical Skills — Changes in the Wind. *N Engl J Med*. 2006;355(25):2664-2669. doi:10.1021/jf052913o

8. Alaraj A. *Comprehensive Healthcare Simulation: Neurosurgery*.; 2018. doi:10.1007/978-3-319-75583-0

9. Badash I, Burtt K, Solorzano CA, Carey JN. Innovations in surgery simulation: a review of past, current and future techniques. *Ann Transl Med*. 2016;4(23):453-453. doi:10.21037/atm.2016.12.24

10. Ray WZ, Ganju A, Harrop JS, Hoh DJ. Developing an Anterior Cervical Diskectomy and Fusion Simulator for Neurosurgical Resident Training. *Neurosurgery*. 2013;73(suppl_1):S100-S106. doi:10.1227/NEU.0000000000000088

11. Harrop J, Rezai AR, Hoh DJ, Ghobrial GM, Sharan A. Neurosurgical Training With a Novel Cervical Spine Simulator: Posterior Foraminotomy and Laminectomy. 2013;73(4):94-99. doi:10.1227/NEU.0000000000000103

12. Ferguson D, Agyemang K, Barrett C, Mathieson C. A low cost dural closure simulation model for tomorrow's spinal neurosurgeons A low cost dural closure simulation model for tomorrow's spinal neurosurgeons. *Br J Neurosurg*. 2018. doi:10.1080/02688697.2018.1540765

13. Tanner G, Vojdani S, Komatsu DE, Barsi JM. Development of a saw bones model for training pedicle screw placement in scoliosis. *BMC Res Notes*. 2017;10(1):1-5. doi:10.1186/s13104-017-3029-3

14. Pfandler M, Lazarovici M, Stefan P, Wucherer P, Weigl M. Virtual reality-based simulators for spine surgery: a systematic review. *Spine J*. 2017;17(9):1352-1363. doi:10.1016/j.spinee.2017.05.016

15. Burdea GC. Proceedings of International Workshop on Virtual prototyping, Laval, France,

pp. 87-96, May 1999. *Proc Int Work Virtual prototyping, Laval, Fr*. 1999;(May):87-96.

16. Alotaibi FE, Alzhrani GA, Sabbagh AJ, Azarnoush H, Winkler-Schwartz A, Del Maestro RF. Neurosurgical assessment of metrics including judgment and dexterity using the virtual reality simulator NeuroTouch (NAJD Metrics). *Surg Innov*. 2015;22(6):636-642. doi:10.1177/1553350615579729

17. Azarnoush H, Alzhrani G, Winkler-Schwartz A, et al. Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. *Int J Comput Assist Radiol Surg*. 2015;10(5):603-618. doi:10.1007/s11548-014-1091-z

18. Sawaya R, Alsideiri G, Bugdadi A, et al. Development of a performance model for virtual reality tumor resections. *J Neurosurg*. 2018:1-9. doi:10.3171/2018.2.JNS172327

19. Alzhrani G, Alotaibi F, Azarnoush H, et al. Proficiency performance benchmarks for removal of simulated brain tumors using a virtual reality simulator neurotouch. *J Surg Educ*. 2015;72(4):685-696. doi:10.1016/j.jsurg.2014.12.014

20. Winkler-Schwartz A, Marwa I, Bajunaid K, et al. A Comparison of Visual Rating Scales and Simulated Virtual Reality Metrics in Neurosurgical Training: A Generalizability Theory Study. *World Neurosurg*. 2019:2-7. doi:10.1016/j.wneu.2019.03.059

21. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence. *AI Mag*. 2006;27(4):12-14. doi:http://dx.doi.org/10.1609/aimag.v27i4.1904

22. Senders JT, Zaki MM, Karhade A V., et al. An introduction and overview of machine learning in neurosurgical care. *Acta Neurochir (Wien)*. 2018;160(1):29-38. doi:10.1007/s00701-017-3385-8

23. Wang S, Summers RM. Machine learning and radiology. *Med Image Anal*. 2012;16(5):933-951. doi:10.1016/j.media.2012.02.005

24. Jones LD, Golan D, Hanna SA, Ramachandran M. Artificial intelligence, machine learning and the evolution of healthcare. *Bone Joint Res*. 2018;7(3):223-225. doi:10.1302/2046-3758.73.BJR-2017-0147.R1

25. Bishop CM. *Pattern Recognition and Machine Learning*. New York: Springer Science+Business Media; 2006.

26. Kotsiantis SB. Supervised Machine Learning: A Review of Classification Techniques. *Informatica*. 2007;31:249-268. doi:10.1115/1.1559160

27. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2(10):719-731. doi:10.1038/s41551-018-0305-z

28. Rajkomar A, Dean J, Kohane I. Machine learning in Medicine. *N Engl J Med*. 2019;380(14):1347-1358. doi:10.1007/978-94-007-5824-7

29. Wang S, Summers RM. Machine learning and radiology. *Med Image Anal*. 2012. doi:10.1016/j.media.2012.02.005

30. Vedula SS, Ishii M, Hager GD. Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. *Annu Rev Biomed Eng*. 2017;19(1):301-325. doi:10.1146/annurev-bioeng-071516-044435

31. Dias RD, Gupta A, Yule SJ. *Using Machine Learning to Assess Physician Competence*. Vol 94.; 2019. doi:10.1097/acm.0000000000002414

32. Hurwitz EL, Randhawa K, Yu H, Côté P, Haldeman S. The Global Spine Care Initiative : a summary of the global burden of low back and neck pain studies. *Eur Spine J*.

2018;27(s6):796-801. doi:10.1007/s00586-017-5432-9

33. Kobayashi K, Ando K, Nishida Y, Ishiguro N, Imagama S. Epidemiological trends in spine surgery over 10 years in a multicenter database. *Eur Spine J*. 2018;27(8):1698-1703. doi:10.1007/s00586-018-5513-4

34. Rajaee SS, Delamarter RB. Spinal Fusion in the United States. 2012;37(1):67-76. doi:10.1097/BRS.0b013e31820cccfb

35. Dvorak MF, Collins JB, Murnaghan L, et al. Confidence in Spine Training Among Senior Neurosurgical and Orthopedic Residents. 2006;31(7):831-837.

36. Daniels AH, Ames CP, Smith JS, Hart RA. Variability in Spine Surgery Procedures Performed During Orthopaedic and Neurological Surgery Residency Training. 2014;196:1-7.

37. Alsideiri G. Validating A Spinal Simulation Model Using NeuroVR. *McGill Univ*. 2017;(August).

38. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. *J Surg Educ*. 2019;In Press:1-13.

39. Murphy TE, Vignes CM, Yuh DD, Okamura AM. Automatic Motion Recognition and Skill Evaluation for Dynamic Tasks. In: *Eurohaptics*. ; 2003:363-373.

40. 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://www.ncbi.nlm.nih.gov/pubmed/15718726. Accessed September 23, 2018.

41. Megali G, Sinigaglia S, Tonet O, Dario P. Modelling and evaluation of surgical performance using hidden Markov models. *IEEE Trans Biomed Eng*. 2006;53(10):1911-1919. doi:10.1109/TBME.2006.881784

42. Hajshirmohammadi I, Payandeh S. Fuzzy set theory for performance evaluation in a surgical simulator. *Presence-Teleoperators Virtual Environ*. 2007;16(6):603-622.

43. Sewell C, Morris D, Blevins NH, et al. Providing metrics and performance feedback in a surgical simulator Providing metrics and performance feedback in a surgical simulator. 2010;9088. doi:10.3109/10929080801957712

44. 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(1):177-182. doi:10.1097/SLA.0b013e3181e464fb

45. 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*. ; 2011:5273-5278. doi:10.1109/ICRA.2011.5979967

46. Loukas C, Georgiou E. Multivariate autoregressive modeling of hand kinematics for laparoscopic skills assessment of surgical trainees. *IEEE Trans Biomed Eng*. 2011;58(11):3289-3297. doi:10.1109/TBME.2011.2167324

47. Liang H, Shi MY. Surgical Skill Evaluation Model for Virtual Surgical Training. *Appl Mech Mater*. 2011;40-41:812-819. doi:10.4028/www.scientific.net/AMM.40-41.812

48. Rhienmora P, Haddawy P, Suebnukarn S, Dailey MN. Intelligent dental training simulator with objective skill assessment and feedback. *Artif Intell Med*. 2011;52(2):115-121.

doi:10.1016/j.artmed.2011.04.003

49. 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):1-11. doi:10.1007/s11548-011-0566-4

50. 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(7):1037-1048. doi:10.1007/s11548-018-1738-2

51. Leong JJH, Leff DR, Das A, et al. Validation of orthopaedic bench models for trauma surgery. *J Bone Jt Surg - Br Vol*. 2008;90-B(7):958-965. doi:10.1302/0301-620X.90B7.20230

52. Bartlett JD, Lawrence JE, Stewart ME, Nakano N, Khanduja V. Does virtual reality simulation have a role in training trauma and orthopaedic surgeons? *Bone Jt J*. 2018;100B(5):559-565. doi:10.1302/0301-620X.100B5.BJJ-2017-1439

53. 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(2):364-372. doi:10.1097/01.sla.0000151982.85062.80

54. Delorme S, Laroche D, Diraddo R, F. Del Maestro R. NeuroTouch: A physics-based virtual simulator for cranial microneurosurgery training. *Neurosurgery*. 2012;71(SUPPL.1):32-42. doi:10.1227/NEU.0b013e318249c744

55. Alotaibi FE, Alzhrani GA, Sabbagh AJ, Azarnoush H, Winkler-Schwartz A, Del Maestro RF. Neurosurgical assessment of metrics including judgment and dexterity using the virtual reality simulator NeuroTouch (NAJD Metrics). *Surg Innov*. 2015;22(6):636-642.

doi:10.1177/1553350615579729

56. Bajunaid K, Abu M, Mullah S, et al. Impact of acute stress on psychomotor bimanual performance during a simulated tumor resection task. 2017;126(January):71-80. doi:10.3171/2015.5.JNS15558.

57. Alotaibi FE, AlZhrani GA, Mullah MAS, Sabbagh AJ, Winkler-schwartz A, Maestro RF Del. Assessing Bimanual Performance in Brain Tumor Resection With NeuroTouch, a Virtual Reality Simulator. 2015;11(1):89-98. doi:10.1227/NEU.0000000000000631

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

59. Ladha L et al. Feature Selection Methods and Algorithms. *Int J Comput Sci Eng*. 2011;3(5):1787-1797.

60. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593

61. Noble WS. What is a support vector machine? *Nat Biotechnol*. 2006;24(12):1565-1567. doi:10.1038/nbt1206-1565

62. Ye J, Janardan R, Li Q. Two-Dimensional Linear Discriminant Analysis. *Adv Neural Inf Process Syst*. 1980;17(60):1569-1576. doi:10.4135/9781412983938

63. McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI/ICML-98 Work Learn Text Categ*. 1998:41-48. doi:10.1.1.46.1529

64. Gélinas-Phaneuf N, Del Maestro RF. Surgical expertise in neurosurgery: Integrating theory into practice. *Neurosurgery*. 2013;73(SUPPL. 4):30-38. doi:10.1227/NEU.0000000000000115

65. Bhatti NI, Cummings CW. Viewpoint: Competency in surgical residency training: Defining and raising the bar. *Acad Med*. 2007;82(6):569-573. doi:10.1097/ACM.0b013e3180555bfb

66. McPheeters MJ, Talcott RD, Hubbard ME, Haines SJ, Hunt MA. Assessing the accuracy of neurological surgery resident case logs at a single institution. *Surg Neurol Int*. 2017;8:206. doi:10.4103/sni.sni_83_17

67. Martin JA, Reznick R, Regehr G, Murnaghan J, Macrae H, Hutchison C. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273-278. doi:10.1046/j.1365-2168.1997.02502.x

68. Winkler-Schwartz A, Yilmaz R, Mirchi N, Bissonnette V, Ledwos N, Siyar S, Azarnoush H, Karlik B, Del Maestro RF. Machine Learning Identification of Surgical and Operative Factors Associated with Expertise in Virtual Reality Simulation. *JAMA Network Open*. 2019 [In Press]

TABLES

TABLE I Description of the mechanisms of the five machine learning algorithms employed

Machine Learning Algorithm	Mechanism*
Support Vector Machines	Use a hyperplane to separate data in two or more groups and maximize the distance between the closest points from both groups and the hyperplane.
Linear Discriminant Analysis	Projects multidimensional data (many metrics) on a single dimension to maximize the distance between the means of the groups and minimize the variance within each group.
k-Nearest Neighbors	Use distance functions such as the Euclidean distance to determine the closest neighbors to a point. A parameter (k) corresponds to the number of neighbors considered. The class of a participant is determined based on their relationship with the nearest participants in a multidimensional space.
Naive Bayes	Classify participants based on probabilities that the chosen metrics belong to experts or novice surgeons. It assumes that all the chosen metrics are independent from each other.
Decision Trees	Classify individuals by building a series of nodes whereby subjects are divided according to the value of a certain metric. The algorithm finds the optimal values to divide subjects in classes.

*The mechanism of every algorithm is further discussed in the literature. [23,25,26,28]

TABLE II Distribution of the sample of population studied in regards to training level and specialty

Training Level	**Orthopaedic Surgery (*counts*)**	**Neurosurgery (*counts*)**	**Total (*counts*)**
Spine surgeons	N/A	N/A	6
Spine fellows	2	1	3
PGY-6*	N/A	2	2
PGY-5*	3	1	4
PGY-4*	3	4	7
Total Senior	**8**	**8**	**22**
PGY-3*	1	1	2
PGY-2*	3	2	5
PGY-1*	2	2	4
Medical students	N/A	N/A	8
Total Junior	**6**	**5**	**19**

*PGY stands for Post-Graduate Year

TABLE III Number of laminectomy cases assisted for each resident

Junior Orthopaedics	Number of laminectomy cases assisted
PGY-1*	0
PGY-1*	0
PGY-2*	3
PGY-2*	6
PGY-2*	N/A
PGY-3*	25
Median	3

Junior Neurosurgery	Number of laminectomy cases assisted
PGY-1*	3
PGY-1*	15
PGY-2*	N/A
PGY-2*	N/A
PGY-3*	3
Median	3

Senior Orthopaedics	Number of laminectomy cases assisted
PGY-4*	4
PGY-4*	30
PGY-4*	20
PGY-5*	50
PGY-5*	10
PGY-5*	N/A
Median	20

Senior Neurosurgery	Number of laminectomy cases assisted
PGY-4†	50
PGY-4†	60
PGY-4†	80
PGY-4†	100
PGY-5†	75
PGY-6*	30
PGY-6*	40
Median	60

*University A
†University B

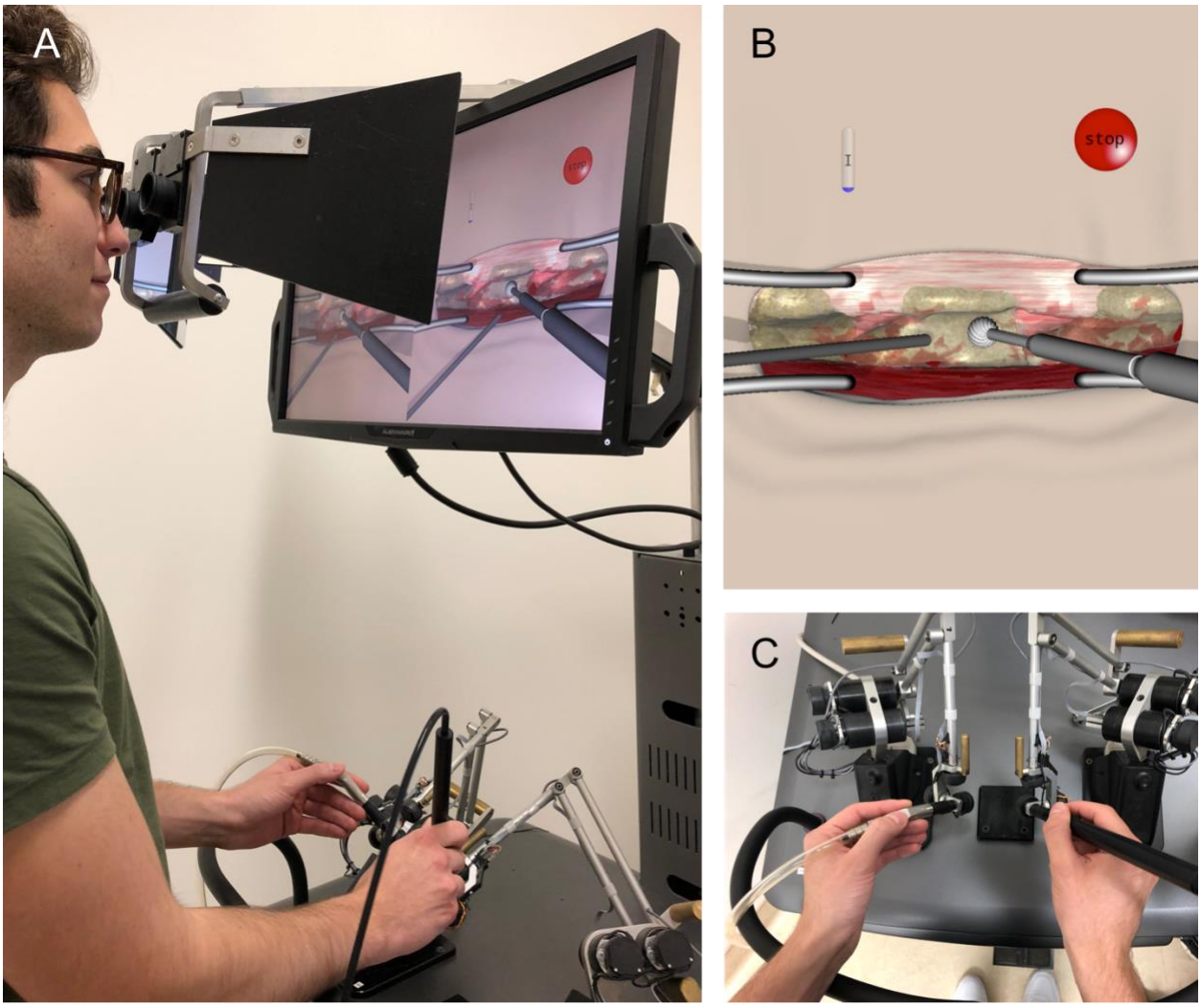
TABLE IV Initial metrics selected by 2 spine surgeons

Safety
Mean force applied on ligamentum flavum
Maximum force applied on ligamentum flavum
Mean force applied on dura
Maximum force applied on dura
Volume of ligamentum flavum removed
Number of times dura was touched with an active burr
Minimum and maximum position of the burr in the cephalad-caudad axis while removing L3
Minimum and maximum position of the burr in the medial-lateral axis while removing L3
Efficiency
Position of the burr when the first removal of L3 occurs
Idle time (amount of time no force is applied by any tool on any structure)
Total tip path length of the burr (sum of every change in position)
Total tip path length of the suction (sum of every change in position)
Amount of time spent removing L3/ Total time to completion
Time to completion
Coordination
Volume removed while simultaneously using the suction and the burr
Mean velocity of the suction while simultaneously using the burr
Number of times structures are touched with suction while using the burr
Amount of time spent while simultaneously using the suction and the burr
Mean distance between the tip of the burr and the tip of the suction
Motion of the tools
Variance of angles of the burr when removing L3
Consistency of movements (distance between two acceleration peaks for both tools when removing L3)
Mean acceleration of the burr over the whole procedure
Mean acceleration of the suction over the whole procedure
Mean velocity of the burr when removing ligamentum flavum
Maximum velocity of the burr when removing ligamentum flavum

TABLE V Final metrics selected by metric selection algorithm

Safety	**Ratio Senior/Junior**
Maximum force applied on dura	0.56
Efficiency	
Amount of time spent removing L3/ Total time to completion	0.96
Coordination	
Amount of time spent while using suction and burr at the same time	1.73
Number of times structures are touched with suction while using the burr	2.18
Motion of the tools	
Distance between two acceleration peaks for the burr in the cephalad-caudad axis when removing L3 (consistency of movements of the burr)	1.48
Distance between two acceleration peaks for the suction in the medial-lateral axis when removing L3 (consistency of movements of the suction)	0.99
Mean acceleration of the burr in the anterior-posterior axis	0.61
Mean acceleration of the burr in the medial-lateral axis	0.73
Mean acceleration of the suction in the medial-lateral axis	0.46
Mean velocity of the burr when removing ligamentum flavum	1.16
Maximum velocity of the burr when removing ligamentum flavum	0.87
Variance of the pitch angle of the burr when removing L3	0.34

FIGURES



Figs. 1-A, 1-B, and 1-C. Demonstration of the NeuroVR platform. **Fig. 1-A** Individual performing the virtual hemilaminectomy scenario. **Fig. 1-B** Virtual tissues include L2, L3 and L4 vertebrae, interspinous ligament, surrounding muscles, ligamentum flavum, intervertebral disc and dura. **Fig. 1-C** The participant must hold the burr in their dominant hand and the suction in their non-dominant hand.

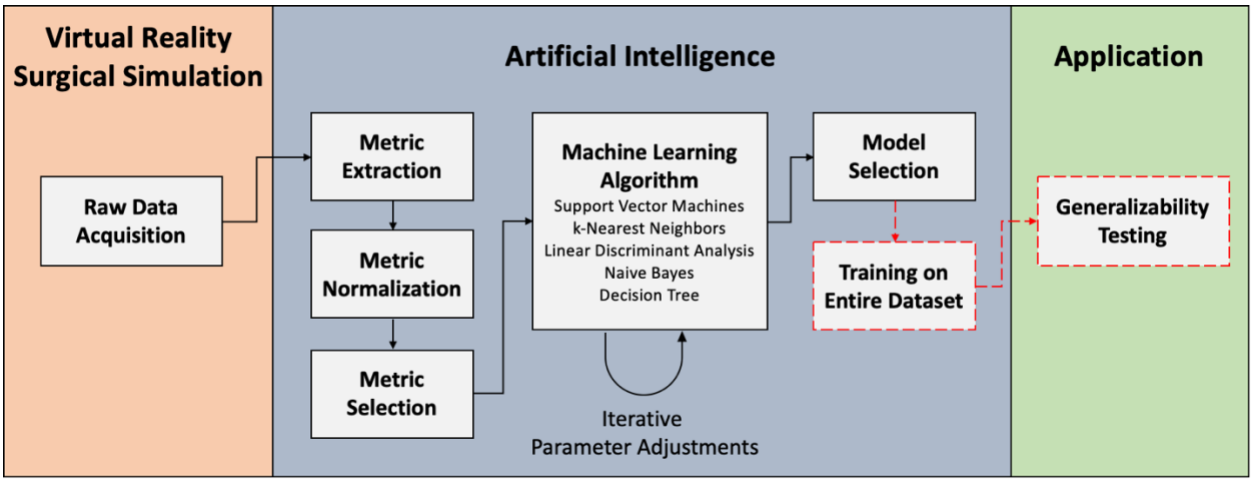


Fig. 2

A framework for integrating artificial intelligence in virtual reality surgical simulation. The virtual reality surgical simulation section involves raw data acquisition from the simulator. Machine learning methodology is followed by performing metric extraction, normalization and selection. The selected metrics are fed to a collection of machine learning algorithms and an iterative process of parameter adjustment is followed to optimize classification accuracy. This step uses cross-validation techniques to assess classification accuracy. Once the optimal algorithm and parameters are identified, a single model is trained using all data. This model can then be used for generalizability testing on new subjects.

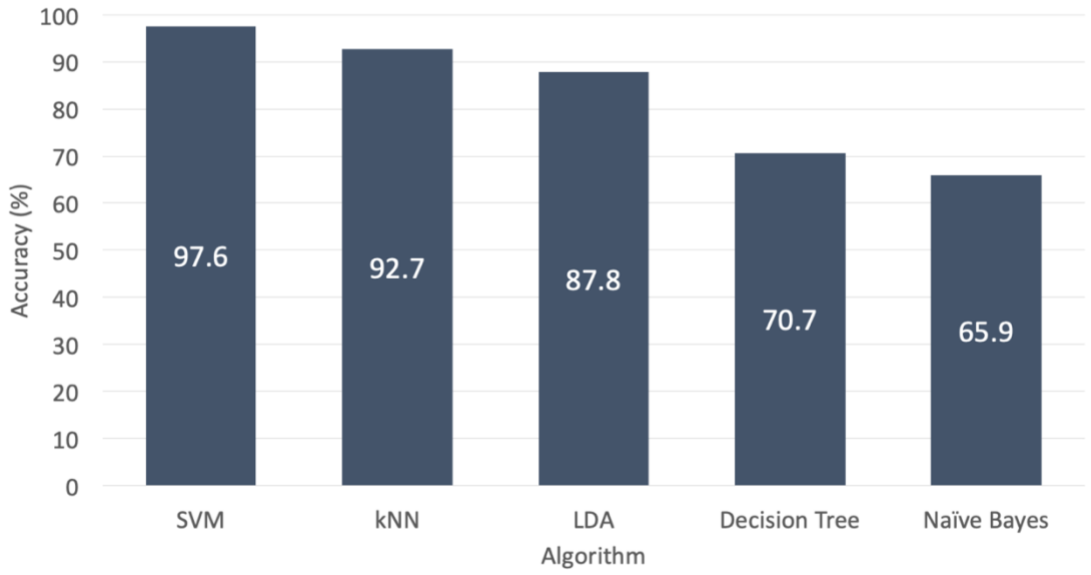


Fig. 3

The support vector machine (SVM) achieved the highest accuracy at 97.6% using leave-one-out cross-validation. The k-nearest neighbors (kNN) reached an accuracy of 92.7%. The linear discriminant analysis (LDA) achieved 87.8%. The decision tree had a 70.7% accuracy. The naive bayes reached the lowest accuracy at 65.9%.

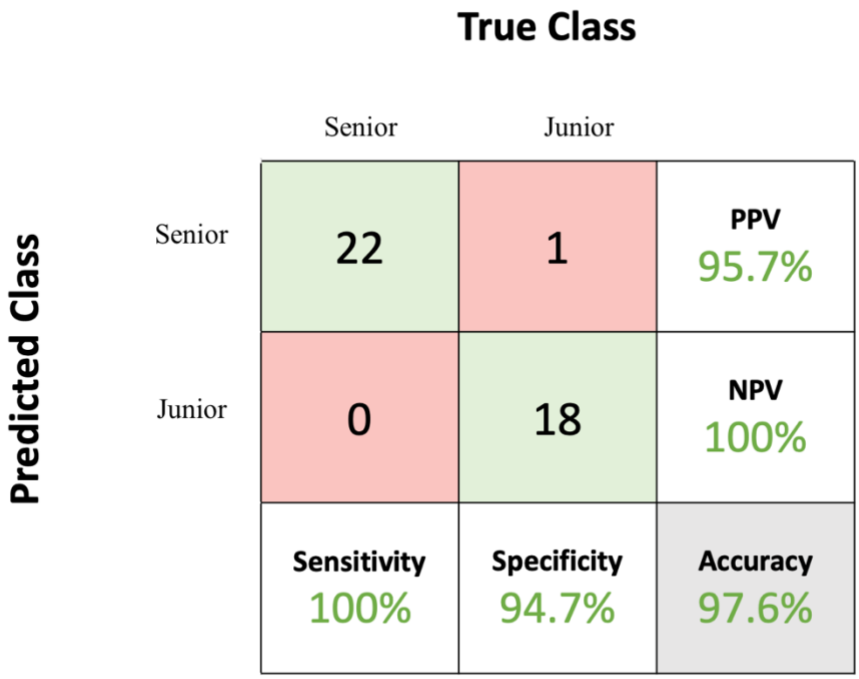


Fig. 4

Using leave-one-out cross-validation, the support vector machine classified senior participants with a sensitivity of 100% and junior participants with a specificity of 94.7%. The Positive Predictive Value (PPV) obtained was 95.7% and the Negative Predictive Value (NPV), 100%. The algorithm achieved an overall classification accuracy of 97.6%.

Video 1

This video shows an individual performing a simulated L3 hemilaminectomy on the NeuroVR platform.

APPENDIX A-Questionnaire provided to spine surgeons

Please check (√) the metrics of performance you believe are important to measure to evaluate a surgeon performing a L3 hemilaminectomy with a burr and suction and could differentiate between a Senior surgeon (PGY-4 and above training level) and a Junior surgeon (PGY-3 and below training level).

METRIC OF PERFORMANCE	Yes
Example: Relevant metric to measure	√
Example: Irrelevant metric to measure	
SAFETY	
Mean force applied on ligamentum flavum	
Maximum force applied on ligamentum flavum	
Volume of ligamentum flavum removed	
Mean force applied on dura	
Maximum force applied on dura	
Volume of dura removed	
Number of times dura was touched with an active burr	
Minimum position of the burr in the cephalad-caudad axis while removing L3	
Maximum position of the burr in the cephalad-caudad axis while removing L3	
Minimum position of the burr in the medial-lateral axis while removing L3	
Maximum position of the burr in the medial-lateral axis while removing L3	
Maximum position of the burr in the anterior-posterior axis while removing dura	
Blood loss	
Other SAFETY metrics not mentioned above?	
EFFICIENCY	
Position of the burr when the first removal of L3 occurs	
Idle time (amount of time no force is applied by any tool on any structure)	
Total tip path length of the burr (sum of every change in position)	
Total tip path length of the suction (sum of every change in position)	
Time to completion	
Amount of time spent removing L3/ Time to completion	
Other EFFICIENCY metrics not mentioned above?	
COORDINATION	
Volume removed while simultaneously using the suction and the burr	
Mean velocity of the suction while simultaneously using the burr	
Number of times structures are touched with suction while using the burr	
Amount of time spent while simultaneously using the suction and the burr	
Mean distance between the tip of the burr and the tip of the suction	

Other COORDINATION metrics not mentioned above?	
MOTION	
Variance of angles of the burr when removing L3	
Distance between two acceleration peaks for both tools (consistency of movements) when removing L3	
Mean acceleration of the burr over the whole procedure	
Mean acceleration of the suction over the whole procedure	
Mean velocity of the burr when removing ligamentum flavum	
Maximum velocity of the burr when removing ligamentum flavum	
Mean velocity of the burr when removing dura	
Maximum velocity of the burr when removing dura	
Other MOTION metrics not mentioned above?	
Other metrics ideas not described in this list?	