Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance

Nykan Mirchi, BSc

Experimental Surgery

McGill University, Montreal

July 2019

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Master of Science in Experimental Surgery

TABLE OF CONTENT

ABSTRACT

Background: Virtual reality surgical simulators are being developed with the goal of providing a safe environment for trainees to practice specific surgical scenarios and allow for self-guided learning. Artificial intelligence technology, including artificial neural networks, offers the potential to manipulate large datasets from simulators to gain insight into the importance of specific performance metrics during simulated operative tasks.

Hypothesis: Artificial neural networks are suitable for and capable of revealing and quantitating important metrics of performance for a virtual reality spinal surgery task.

Objectives: First, to develop metrics of performance for a novel virtual reality anterior cervical discectomy simulation. Second, to employ artificial neural networks to classify participants' expertise based on their performance in the simulated task. Third, to examine the ability of the neural network to outline the relative weights of specific metrics in the determination of expert performance in this virtual reality spinal procedure.

Methods: Twenty-one participants performed a simulated anterior cervical discectomy on the novel virtual reality Sim-Ortho simulator. Participants were divided into 3 groups, including 9 Post-Resident, 5 Senior, and 7 Junior participants. Data was recorded and manipulated to calculate metrics of performance for each participant. Neural networks were trained and tested and the relative importance of each metric was calculated.

Results: A total of 369 metrics spanning four categories (safety, efficiency, motion, cognition) were generated. An artificial neural network was trained on 16 selected metrics, and tested achieving a training accuracy of 100% and a testing accuracy of 83.3%. Network analysis identified safety metrics, including the number of contacts on spinal dura, as highly important.

Conclusion: Artificial neural networks classified 3 groups of participants based on expertise allowing insight into the relative importance of specific metrics of performance. This novel methodology aids in the understanding of which components of surgical performance predominantly contribute to expertise.

RÉSUMÉ

Contexte: Les simulateurs chirurgicaux de réalité virtuelle sont en cours de développement dans le but de fournir aux stagiaires en chirurgie un environnement sûr leur permettant de mettre en pratique des scénarios chirurgicaux spécifiques et de permettre un apprentissage autonome. La technologie de l'intelligence artificielle, y compris les réseaux de neurones artificiels, offre la possibilité de manipuler de grands ensembles de données à partir de simulateurs pour mieux comprendre l'importance de mesures de performances spécifiques lors des opérations chirurgicales simulées.

Hypothèse: Les réseaux de neurones artificiels sont adaptés et capables de mettre en évidence et de quantifier des mesures de performance importantes pour une opération de chirurgie rachidienne en réalité virtuelle.

Objectifs: Premièrement, de développer des mesures de performance pour une simulation innovatrice de discectomie cervicale antérieure en réalité virtuelle. Deuxièmement, d'utiliser des réseaux de neurones artificiels afin de classer les compétences des participants en fonction de leurs performances dans la tâche simulée. Troisièmement, d'examiner la capacité du réseau de neurones à définir l'importance relative des mesures spécifiques dans la détermination de la performance d'un expert dans cette procédure rachidienne en réalité virtuelle.

Méthodologie: Vingt et un participants ont réalisé une discectomie cervicale antérieure simulée sur le nouveau simulateur de réalité virtuelle *Sim-Ortho*. Les participants ont été divisés en 3 groupes, comprenant 9 participants « Post-Resident », 5 participants « Senior » et 7 participants « Junior ». Cette étude s'est concentrée sur la portion discectomie de l'opération. Les données ont été enregistrées et traitées pour calculer des mesures de performance pour chaque participant.

Les réseaux de neurones ont été formés et testés; et l'importance relative de chacune des mesures de performance a été calculée.

Résultats: Un total de 369 mesures de performance couvrant quatre catégories incluant la sécurité, l'efficacité, le mouvement et le cognitif ont été générées. Un réseau de neurones artificiels a été formé sur 16 mesures sélectionnées et testé avec une précision d'entraînement de 100% et une précision des tests de 83,3%. L'analyse du réseau a révélé que les mesures de sécurité, y compris le nombre de contacts sur la dure-mère, étaient très importantes.

Conclusion: Les réseaux de neurones artificiels ont classé 3 groupes de participants sur la base d'une expertise permettant de mieux comprendre l'importance relative de mesures de performance spécifiques. Cette méthodologie novatrice aide à comprendre quelles composantes de la performance chirurgicale contribuent principalement à l'expertise.

ACKNOWLEDGEMENTS

Foremost, I would like to express my most sincere gratitude to my supervisor, Dr. Rolando Del Maestro for his continuous support and guidance throughout my graduate studies. The work presented in this thesis would not have been possible without him. As my principle investigator, Dr. Del Maestro offered me the perfect balance between creative freedom as a graduate student, while guiding me through the complexities of conducting and reporting scientific findings. His role went far beyond that which I expected from a supervisor. He has acted as a mentor and source of inspiration not only within the context of my research, but also with regards my future academic endeavours. I could not have asked for a better supervisor through this journey.

I would like to recognise two special individuals without whom my time at the lab would not have been as enriching and enjoyable. To Miss. Nicole Ledwos and Dr. Vincent Bissonnette. I am beyond grateful for the friendship which flourished from our work at the lab and our adventures outside the lab. I could not have asked for more enthusiastic, well-rounded and had-working colleagues to share my graduate adventures with. To Nicole, I wish you the best of luck in your future. You will make a great doctor one day. And to Vincent, I wish you best with your Orthopaedic Surgery residency at Queen's University.

I would like to thank Dr. Alexander Winkler-Schwartz and Dr. Recai Yilmaz for their help which have allowed me to accomplish this work. I am grateful for the interesting discussions which we have had at the lab, and for the unique opportunities which you have provided me with to learn more about medicine and neurosurgery.

I thank all researchers including Dr. Bekir Karlik, Dr. Hamed Azarnoush, Miss. Samaneh Siyar, and others who have visited the lab and with whom I have collaborated and learned from over the year.

I would also like to thank the members of my research advisory committee: Dr. Jean Ouellet, Dr. Maria Petropavlovskaya, Dr. Greg Berry and Dr. Roy Dudley for their encouragements and insightful comments.

I thank the AO Foundation for financially supporting the work presented in my thesis, as well as the Di Giovanni Foundation for their continuous support of the lab over the years.

To my friends and family who have been by my side throughout my studies, I would like to thank you for making this past year as enjoyable as could be.

Finally, to my parents, I could not have done everything which I have achieved so far if not for their continuous love and support. Thank you for being my role-models and for inspiring me to follow my dreams.

PREFACE AND CONTRIBUTION OF AUTHORS

The thesis presented herein is structured in a manuscript-based manner. The study is currently under review by the editorial board at Neurosurgery. This study was built on established work from the Neurosurgical Simulation and Artificial Intelligence Learning Centre.

The Candidate led the study and performed all aspects of the study including data collection, development of machine learning methodology, data analysis, result interpretation and writing of the manuscript.

Miss Nicole Ledwos and Dr. Vincent Bissonnette assisted in the recruitment of participants for the study, and running the trial.

Dr. Recai Yilmaz and Dr. Vincent Bissonnette contributed their machine learning knowledge.

Dr. Alexander Winkler-Schwatz was essential in providing a surgical perspective on the realism of the scenario to provide feedback to OSSimTech prior to the start of the study. He also provided knowledge on educational theories.

Dr. Bekir Karlik assisted in the development and optimization of the neural network.

Dr. Rolando Del Maestro was primarily in charge of the overall direction and planning of the study. He also greatly contributed to the development of performance metrics and the discussion of the educational applications of the neural networks presented herein.

ABBREVIATIONS

ACDF: anterior cervical discectomy and fusion

AI: artificial intelligence

ANN: artificial neural network

ML: machine learning

VR: virtual reality

THESIS INTRODUCTION

Virtual reality surgical simulation has become a rapidly evolving field of research in the last decade.[1,2] This trend follows the desire for more objective and competency-based training methods for surgery.[3] The complexity and implications of spine procedures makes spine training a paradigm of interest for novel simulation-based methods.[2,4] The anterior cervical discectomy and fusion (ACDF) is one of the most commonly performed spine procedures, making it an ideal candidate for simulation-based training. The ACDF procedure requires proficiency in multiple areas including an understanding of the critical anatomical structures, along with gaining an appreciation of how different structures react to manipulations and instrument usage.[5]

Virtual reality simulation typically relies on powerful computers that can record an enormous amount of data about how a surgeon is interacting with a simulated task. Studies in surgical simulation have developed methodologies to exploit these large datasets to develop validated metrics of performance which can be used by surgical educators to enhance performance.[4,6] These metrics are generally developed in such a way that they can be understood and taught by surgical educators. As no direct supervision is required to calculate individual metrics, virtual reality simulation may pave the way for more objective methods of assessment and may offer room for more self-guided learning amongst surgical trainees.

Combining this technology with artificial intelligence (AI), the surgical community may be able to gain further insight into specific components of surgical performance that can differentiate levels of expertise. Artificial intelligence is a broad term used to describe a set of algorithms that can make seemingly intelligent decisions.[7,8] Machine learning, is a subset of artificial intelligence, where algorithms are able to identify and learn from hidden patterns in multivariate datasets, without the need for explicit programming.[8] Artificial neural networks, are

a deeper subset of machine learning and inspired from neuronal connectivity in the brain.[9] They are sets of interconnected nodes (referred to as neurons in this paper) which can communicate with each other through connections with different weights. These weights are essentially analogous to neuromodulatory signals that influence how neurons communicate with one another. When combined with virtual reality simulation, artificial neural networks can be designed to classify participants and discover specific metrics that differentiate surgical performance. This information can provide an objective assessment of surgical psychomotor performance providing insight into the components that underpin surgical expertise.

Current literature contains few studies which employ artificial intelligence in surgical simulation.[10-23] The vast majority which do, limit their analysis to the classification of different groups of surgeons, and fail to explore the underlying reasons for classification by investigating the relative importance of metrics of performance.[15]

The hypothesis tested in this study is that artificial neural networks, a type of artificial intelligence, will be able to differentiate three group of varying expertise performing an anterior cervical discectomy in a virtual reality surgical simulator using metrics of performance relevant to safety, motion, efficiency, and cognition. This hypothesis incorporates three primary objectives:

1. To develop metrics of performance for a novel virtual reality ACDF simulation.
2. To employ artificial neural networks to classify participants' expertise based on their performance in the simulated task.
3. To examine the ability of our neural network to outline the relative weights of specific metrics in the determination of expert performance in this virtual reality spinal procedure.

This novel methodology has the potential to aid in the understanding of components of surgical expertise and contribute to the paradigm shift towards competency-based surgical training. In the context of surgical simulation, using artificial intelligence to understand and extract important components of expertise may help alter the purpose of simulators from a training tool allowing acquisition of skill to an educational tool able to communicate knowledge and information about one's performance.[1]

BACKGROUND

A Brief History of Surgical Training

Surgical education and training have relied upon the apprenticeship model. However, as medical knowledge and our understanding of surgical practice evolved, so have training methods.[24] A paradigm shift from the apprenticeship model, to one that is more competency based is presently taking place.[1] To understand this shift, it is important to understand the origins of the original model of surgical training.

In its early years, surgery and medicine were regarded as two separate disciplines. To many physicians and surgeons today, this is revealed through clues in the Hippocratic Oath which states: "I will not use the knife, and certainly not those suffering from stone, but I will cede to men practitioners of this activity".[25] As opposed to today, surgeons were not regarded as medical doctors, but were rather comprised of "barber-surgeons", individuals skilled in manual crafts, but did not require a medical degree.[24] In the early years of surgical training therefore, medical and surgical training developed along different paths.

With the lack of highly selective medical training, the apprenticeship model was utilized to create a path for the training of surgeons. Since surgeons' performance involved defined technical skills utilized by a multitude of different craftsmen, the apprenticeship model was considered appropriate. At its essence, this model relies on a trainee observing and learning from a master in the craft.[24] Over time, the trainees are given the opportunity to practice under gradually decreasing supervision until they are deemed capable and skilled.[24] In some craft-like and skill-dependent fields such as pottery or barberry, this model is highly useful. The trainees would learn over time, and the subject, whether it be a piece of clay for potters or a customer's beard for barbers, was never truly in a state of "high risk". A mistake by the trainee would only

result in an unhappy customer. In surgical training, the trainee observes an experienced surgeon (i.e. a master of their own craft), and is gradually given more and more responsibility during surgeries until they are deemed capable of practicing alone.[24] However, surgery is a high-risk craft and unlike some other crafts, mistakes can cause serious harm to patients. In the surgical apprenticeship model, training relies on the subjective decisions of a master surgeon or series of surgeons who deems whether a trainee is capable of performing surgical procedures adequately. The combination of these two concerning issues has driven an evolution in the apprenticeship model over time.

The original approach of the apprenticeship model relies on "see one, do one, teach one", whereby the trainees are expected to observe their master and mimic their skill, and eventually teach these skills to other students. This model results in a lack of standardization amongst trainees.[24] Each master acted as their own "training program" where individuals may choose to teach different methods and knowledge.[26] This led to one of the most significant modifications to the apprenticeship model. In the hope of creating a more standardized method of surgical training in the United States, Dr. William Halstead proposed a model inspired by the German training philosophy.[27] This model relied on "graduated responsibility".[27] This entailed that trainees would gain increasing responsibility over a defined series of years as they advanced through their training.[27] This formed the foundations of the modern surgical training model by providing both standardization and structure to surgical training.[24]

Although Halstead's model was a modification to the original apprenticeship model, it is now facing a series of significant issues. Concerns of the usefulness of this model have been expressed related to the shift from the apprenticeship to a competency based model. There are several underlying reasons for this concern including a rise in patient expectations, evidence-

based medicine and student numbers. A large part is attributed to the restriction in training hours for residents thereby reducing their practice and exposure.[28,29] In the United States, work-hours have been restricted to 80 hours per week in hopes of reducing resident fatigue and burnout.[30] However, recent psychological studies revealed no significant effect on burnout.[31] These new restrictions reduce trainee exposure to operative procedures.[30] Surgical simulators are being developed in an attempt to mitigate this issue.

Simulation in Surgery

Compared to other high-risk industries such as aviation or the military, simulation is rather new to surgery.[32] A simulation is any attempt to mimic or replicate a real life scenario, immersing users into a simulated scenario. Simulators are commonly used to train those who work in high-risk situations, as is the case for pilots.[32] They are also utilized to expose trainees or experts to complicated situations and help in the assessment of actions and reactions, with the expectation that they will be more prepared for possible unlikely events.[32] This is the case with NASA, where simulation is a vital part of exposing astronauts to conditions similar to those experienced in space before they have the opportunity to venture beyond the atmosphere.[33]

Simulation platforms have a central goal: allowing room for failure.[32,34] By replicating real surgical procedures, simulators can immerse trainees and allow them to perform complex surgical tasks without the risk of causing harm to patients. Through these systems trainees also have the opportunity to learn from their mistakes. They can visualise, experience and understand the risks and potential complications of their actions. Literature on simulation has shown that trainees appreciate the opportunity to experience the consequences of their actions in a safe environment.[34] Some trainees have reported an increase in anxiety if they performed poorly in a

simulated scenario.[34] To address this issue, it is important for trainees to not only be exposed to simulators, but also to obtain structured feedback, allowing them to reflect on their performance, while receiving guidance on how to improve. This task has proved challenging and remains a barrier for the wide adoption of simulation in surgical training.[35]

Several types of surgical simulators are available on the market today. They can be broadly divided into two categories: physical (or benchtop) simulator, and virtual reality or augmented reality simulators. Physical simulators are any simulator with which the users can directly interact. These include manikins or laparoscopic boxes. The modern manikin was first introduced to anaesthesia training in the 1960s, followed by the first development of a high-fidelity simulator in the early 1980s.[32,36,37] These combined intricate designs that looked like a patient, but also contained computer chips able to re-create vital signs that could respond to disruption caused by the user.[32] Virtual reality simulators on the other hand, were first introduced in the 1990s.[32,38] These are entirely computer-based systems which allow the user to feel immersed in a particular surgical scenario. Users typically hold tool handles and look at a screen where they can see a simulated version of their instrument interacting with anatomical structures. Some of the first VR simulators were for relatively simple tasks, such as suturing, cholecystectomy and minimally-invasive surgery.[32] As computers became more powerful at the beginning of the 21st century, more complex and more realistic VR simulations could be developed. One such example is the NeuroVR (originally known as the NeuroTouch).[2] The NeuroTouch platform resulted in an interdisciplinary collaboration amongst Canadian researchers and industry, to produce the most advanced virtual reality simulator for neurosurgery.[2] Using a finite-element method, the platform can create highly realistic 3D images able to deform and respond to physical manipulations. The system also incorporates advanced

haptic feedback, so that users can feel different tissue as they are performing a task on the simulator. However, only a very limited number of scenarios for spine surgery are available on the platform.

As the development of virtual reality simulators began to grow, efforts to develop scenarios for spine surgery simulation have been limited.[39] One reason for this is the fact that the spine is composed of multiple different anatomical components with different visual appearance and tissue densities. An effective virtual reality simulator for spine surgery would need to offer variable haptic feedback able to simulate both the softer and more malleable muscle, nerve and spinal tissue, as well as the more rigid vertebrae. In light of these challenges, attempts have been made to create a VR simulator without haptic feedback. One such example involved a VR simulator for orthopaedic surgery. However, due to the lack of haptic feedback, the primary aim of this simulator was shifted to pre-operative planning rather than surgical training.[40] The ImmersiveTouch platform which incorporates haptic feedback, was later introduced with 3 spinal procedure scenarios: lumbar puncture, pedicle screw placement, and vertebroplasty.[41] However, the spine procedures on this system are relatively simple, only requiring a single step or action to complete. A systematic review exploring VR spine surgery simulators from 2005 to 2016 identified 19 studies relevant to the field.[42] The majority studied relatively simple tasks such as pedicle screw placement or lumbar puncture on the ImmersiveTouch. Of the 19, only 2 sought to introduce a new simulator.[43,44] However, both relied on augmented reality. Augmented reality is similar to VR in the sense that it requires computers to create a visual component of the simulation.[42] Augmented reality works by superimposing visual components on top of a physical stimulus such as a mannequin.[42] The advantage of this method is that it can bypass the challenge of simulating complex haptic feedback as the user is partly interacting with physical models.

There is a need for the development of advanced VR simulators incorporating haptic feedback for more complex spine surgery procedures. The anterior cervical discectomy is an ideal candidate as it requires users to interact with a variety of soft and rigid structures, in several steps, while using a multitude of instruments, each with their unique ability to deform anatomical structures. This procedure is the basis of a novel VR simulator by OSSimTech™ (Montreal, Quebec, Canada) a Canadian start-up.

A New Virtual Reality Simulator for Spine Surgery

The Sim-Ortho simulator is a novel virtual reality simulator co-developed by OSSimTech™ (Montreal, Quebec, Canada) and the AO Foundation (Bienne, Switzerland). The platform is based on a gaming-engine, and incorporates 3 modes of mimicking real surgical scenarios: visual, auditory and haptic. The simulator replicates intra-operative environments in a realistic manner by creating a highly realistic 3D interface.[28] The auditory component is accomplished by simulating the sounds typical in the operating room as well as the sound of the various instruments when interacting with anatomical structures. The haptic components provide a variety of forces and resistance dependent on the tissue interacted with. This advanced technology also allows for diversity in the feel of tissues. For example, the intervertebral disc is relatively soft and flexible as opposed to the C4 and C5 vertebra which are harder and more rigid. Overall, it is the combination of these three sensory components that make the Sim-Ortho platform a useful system to assess the potential impact of virtual reality spinal surgical simulation.

The computer-nature of virtual reality simulators offers the potential to track many aspects of how an individual interacts with a specific simulated scenario. Previous simulation

studies in neurosurgical tasks have utilized methods to deal with such large datasets by creating a series of tiers of psychomotor performance metrics based on performance benchmarks.[4,6,45] However, current methods of analysing surgical simulation data remain limited. A large number of studies employed traditional statistical methods (such as t-tests or ANOVAs) to determine significant differences in individual metrics between groups of varying surgical expertise.[46,47] These methods consider each metric of performance independently. Surgical expertise, on the other hand, is generally regarded as an interrelated combination of measures (or metrics). Hence, a novel method of data analysis, powered by artificial intelligence is deemed well suited for dealing with large datasets extracted from surgical simulators.

Artificial Intelligence: What is it? How can we use it?

Artificial intelligence (AI) is simply a branch of computer science which revolves around giving computers and machines the ability to perform tasks intelligently. As illustrated in Figure 1, artificial intelligence can be subdivided into three main categories: Deep Learning, Machine Learning and Language Processing (Chatbots). The largest branch of artificial intelligence is machine learning. This technology allows computers to find hidden patterns in very large datasets, learn from these patterns, and essentially make decisions without the need for explicit programming. Many subsets of algorithms fit under the umbrella of machine learning. On one end, there are simple algorithms such as support vector machines, k-nearest neighbours, and decision trees, all of which have relatively similar methods of approaching large datasets. On the other end, there are more complex algorithms known as artificial neural networks. Artificial neural networks are inspired from the neuronal connectivity in the brain, as they are composed of interconnected nodes (analogous to neurons). These neurons are interconnected in such a manner

to regulate the impact of each input in the decision-making process to produce an output. This logic is analogous to the neuromodulatory signals in the brain, whereby during learning, certain neurons may be downregulated whereas others may be upregulated. Deep learning is a branch which has drawn the largest attention in recent years with significant advancements in design and mathematical underlying.[48,49] Two Canadians and a French scientist, Yoshua Bengio, Geoffrey Hinton, and Yann LeCun are often regarded as the founding fathers of deep learning and were recently awarded the Turing Prize, the highest distinction in computer science, for their work in the field.[48,50] Deep learning is the newest of the three branches and hence remains the least understood. Its applications in surgery remain limited. In addition, deep learning generally faces a "black box" problem as it is extremely difficult to understand how these algorithms make decisions.[51,52] As medicine and surgery are highly evidence-based practices, medical professionals may remain sceptical of deep learning programs which cannot reveal exactly how they made a particular decision. Another branch of artificial intelligence is known as natural language processing, or chatbots. Simply, these are able to comprehend speech and respond appropriately. Its applications in surgery are rather limited at the moment but several groups have attempted to use this technology in other fields of medicine. Some examples include a Canadian start-up, WinterLight Labs, which uses natural language processing to identify speech patterns which may be indicative of dementia in patients.[53] Another American start-up uses the technology as an automated medical scribe during physician-patient interactions, thereby significantly improving the amount of time physicians can spend with patients rather than with their computer.[54]

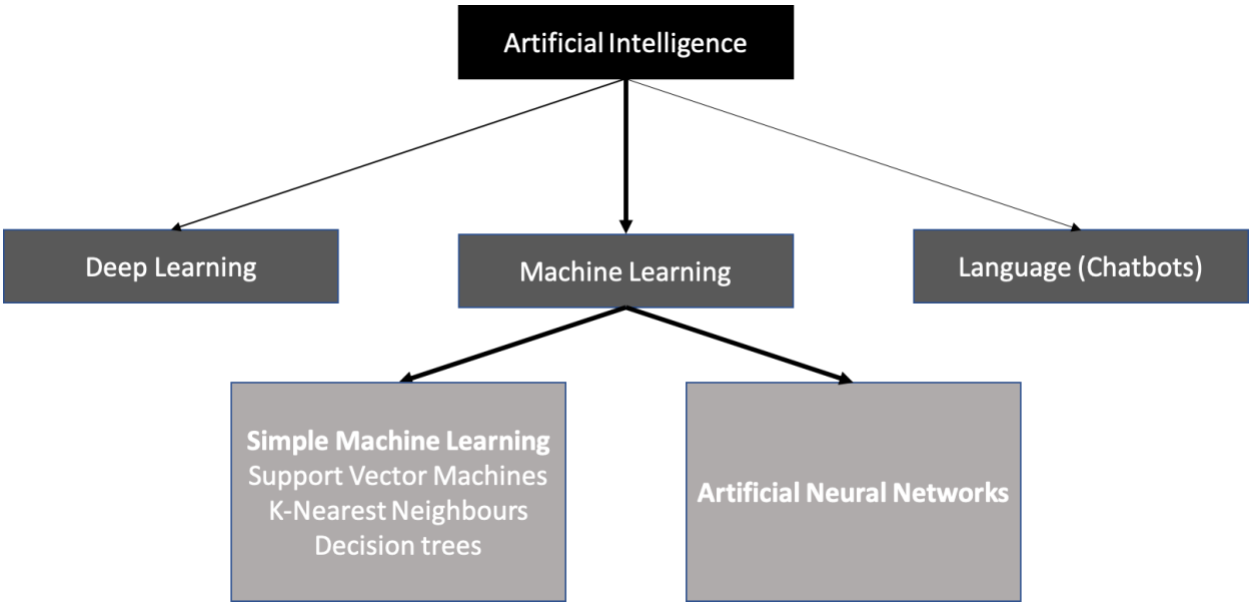

Figure 1: Three subsets of the artificial intelligence. Artificial intelligence can be broken down into: Deep Learning, Machine Learning and Chatbots. Machine learning can be further broken down into the simpler algorithms such as support vector machines, or the more complex artificial neural networks.

The work presented herein employs machine learning. There are two primary methods to employ machine learning: supervised and unsupervised. Supervised machine learning is the most common, and involves providing labelled data to the algorithm from which it can learn.[55] This means that during training, the algorithm is aware of which group each data point belongs to. It then uses this information to form a general rule that is able to differentiate two or multiple groups. For example, in the context of surgical training, a supervised approach to differentiate levels of expertise in a surgical task would involve initially providing a large dataset, where each datapoint is labelled as either belonging to a novice or an expert individual. The algorithm learns from this pre-defined grouping and is then able to classify a new surgeon as either a novice or expert based on what it learned previously. This is the approach employed in this study. The

unsupervised approach requires no labels, although a much larger dataset is generally required. In this method, the algorithm is fed a large dataset without prior knowledge of which group each datapoint belongs to. It can then automatically decide how to group datapoints according to hidden patterns in the data. In the context of surgical training, this could be accomplished by providing an algorithm with a large amount of data from a variety of different surgeons from multiple institutions. The algorithm would have no knowledge about which surgeon each datapoint belongs to. It would then independently identify groups in the data, based on how similar certain groups of surgeons performed in a scenario. Such methodology may offer insight on the existence of multiple differentiable groups of expert surgeons. However, this requires very large datasets and is therefore not employed in this study.

There are several advantages of using machine learning to analyse data from virtual reality surgical simulators. First, machine learning allows for the automated classification of individuals into two or more groups of predefined expertise levels. Second, machine learning allows for analysis of the relationship between different metrics of performance rather than simply assessing metrics individually. Hence, this provides a more holistic overview of expertise as opposed to more traditional statistical methods of differentiating expertise. Finally, machine learning can rank the importance of different metrics in differentiating expertise. For example, although a set of 10 metrics may be essential to differentiate expertise in a particular surgical task, one or two metrics may have a significantly larger influence on the algorithm's decision-making process. As such, this may inspire the development of future surgical training platform and programs. Overall, the advantages of artificial intelligence in surgical simulation has not gone unnoticed in the medical community.

Artificial Neural Networks to Assess Surgical Competency

As collaboration amongst medical professionals and computer scientists has grown over recent years, many applications of artificial intelligence have been proposed from clinical-decision support systems, to assistance in diagnosis.[56,57] A field which has remained underdeveloped for artificial intelligence is that of surgical and medical training. A recent systematic review identified only 69 articles across eight databases where machine learning was employed to assess physician competency.[58] Half of these were published in the last six years. A more focused systematic review was also performed along with members of the Neurosurgical Simulation and Artificial Intelligence Learning Centre where 12 articles were found to be relevant to the use of artificial intelligence to differentiate expertise in virtual reality surgical simulation. The candidate third author in this review which has been submitted to the Journal of Surgical Education. From the combination of both reviews, the majority of studies employed support vector machines, a simpler type of machine learning algorithm. The particular advantage of such an algorithm is that it is easier to understand how these algorithms make their decision. Only ten studies used artificial neural networks. Interestingly, the use of neural networks to assess competency in medicine seems to expand far beyond surgery, with numerous studies comparing the effectiveness of neural network-powered tools for radiologists to identify lesions.[59-64] Little attempt has been made to use artificial neural networks to assess surgical performance. In 2010, Richstone et al employed artificial neural networks to assess surgical skill by tracking the eye movement of surgeons during simulated and live surgeries.[15] Although the results established an important proof of concept for the use of artificial neural networks to assess surgical skill, this was accomplished with metrics that cannot easily be taught to trainees. This study had significant limits related to the usefulness of neural networks for formative

assessment in surgery. The complexity of eye movement renders it difficult to specifically infer how the network made its classification decisions. In 2015, Yost et al conducted a study involving 7 US surgical training centres and employed artificial neural networks for an assessment of performance.[65] However, instead of using simulation, these networks were based on the results of questionnaires assessing behavioural style and motivators. The focus of this study was on the psycho-social components which may be able to differentiate expertise in surgeons. As professional surgical practice is a holistic profession, there is no doubt that psychosocial metrics are important to assess. However, this study only provided a partial picture of factors that can differentiate expertise. An important next step would be to employ artificial neural networks to assess technical skills during simulated surgical procedures. The utilization of artificial neural networks to assess and quantitate psychomotor skills may be an important adjunct to surgical education since the skills identified can be employed in a training paradigm.

Interestingly, artificial neural networks have been used to not only classify performance in medicine and surgery (e.g. novice vs. expert), but also to gain insight into the underlying factors that led to classification. A recent study employed this technology in an attempt to understand which components are most important in the differentiation of medical students and specialists in clinical diagnostic simulations.[66,67] This was accomplished by studying ANNs in greater depth. Not only are the networks able to differentiate performance, but the authors attempted to crack open the network to gain insight on its decision-making process. This effort to render the network more transparent is vital for the use of artificial neural networks in medical education.[68] As both parties are able to gain insight into the specific components which may have led to their classification, this creates a sense of trust and ensures a successful connection between trainers and trainees in the learning process.

The study presented herein builds from previous work on artificial neural networks to assess surgical competency by not only addressing technical surgical skill in virtual reality simulation but also employing a transparent approach to help lay the foundations for real-life educational applications.

RATIONALE FOR STUDY

This project aims to offer a novel approach to learn about surgical performance by employing artificial intelligence and simulation for a spinal task. Primarily, it aims to expand on the surgical community's understanding of factors that differentiate performance in a cervical spine discectomy using virtual reality simulation.

STUDY

Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance

Nykan Mirchi, BSc, Vincent Bissonnette, MD, Nicole Ledwos, BA, Alexander Winkler-Schwartz, MD, Recai Yilmaz, MD, Bekir Karlik, PhD, Rolando F. Del Maestro, MD, PhD

The preceding work incorporates supplemental material to provide a more thorough overview of the methodologies used and results obtained.

Manuscript under review in Neurosurgery.

ABSTRACT

Background: Virtual reality surgical simulators provide a safe environment for trainees to practice specific surgical scenarios and allow for self-guided learning. Artificial intelligence technology, including artificial neural networks, offers the potential to manipulate large datasets from simulators to gain insight into the importance of specific performance metrics during simulated operative tasks.

Objective: This study aims to distinguish performance in a virtual reality simulated anterior cervical discectomy scenario, uncover novel performance metrics, and gain insight into the relative importance of each metric using artificial neural networks.

Methods: Twenty-one participants performed a simulated anterior cervical discectomy on the novel virtual reality Sim-Ortho simulator. Participants were divided into 3 groups, including 9 Post-Resident (consulting spine surgeons and spine fellows), 5 Senior (PGY 4-5 orthopaedic residents and PGY 4-6 neurosurgery residents), and 7 Junior (PGY 1-3 residents in neurosurgery and orthopaedic surgery) participants. This study focussed on the discectomy portion of the task. Data was recorded and manipulated to calculate metrics of performance for each participant. Neural networks were trained and tested and the relative importance of each metric was calculated.

Results: A total of 369 metrics spanning four categories (safety, efficiency, motion, cognitive) were generated. An artificial neural network was trained on 16 selected metrics, and tested achieving a training accuracy of 100% and a testing accuracy of 83.3%. Network analysis identified safety metrics, including the number of contacts on spinal dura, as highly important.

Conclusion: Artificial neural networks classified 3 groups of participants based on expertise allowing insight into the relative importance of specific metrics of performance. This novel

methodology aids in the understanding of which components of surgical performance predominantly contribute to expertise.

INTRODUCTION

Virtual reality surgical simulation is an evolving field of research which has the potential to complement more objective and competency-based surgical training methods.[1-3] The incidence, complexity and implications of the anterior cervical discectomy and fusion (ACDF) makes this procedure a good candidate for simulation-based training.[5,69] The ACDF procedure requires proficiency in multiple areas including an understanding of the critical anatomical structures, along with gaining an appreciation of how different structures react to manipulations and instrument usage.[5] Virtual reality simulators record an enormous amount of data concerning psychomotor performance during a simulated task. Studies in surgical simulation have developed methodologies to exploit these large datasets to develop validated metrics of performance which can be used by surgical educators to enhance performance.[4,6]

Artificial intelligence is a broad term used to describe a set of algorithms that can make seemingly intelligent decisions.[7,8] Machine learning is a subset of artificial intelligence, where algorithms are able to identify and learn from hidden patterns in multivariate datasets, without the need for explicit programming.[8] Artificial neural networks, are a deeper subset of machine learning, inspired from the brain's neuronal connectivity.[9] They are sets of interconnected nodes (referred to as neurons in this paper) which can communicate with each other through connections of different weights. These weights are essentially analogous to neuromodulatory signals that influence how neurons communicate. When combined with virtual reality simulation, artificial neural networks can be designed to classify participants and discover specific metrics that differentiate surgical performance. This information can provide an objective assessment of surgical psychomotor performance providing insight into the components that underpin surgical expertise.

Artificial intelligence has been utilized to assess surgical expertise during virtual reality performance but the majority of these studies limit their analysis to the classification of different participant groups.[10] These systems fail to explore the underlying reasons for classification by investigating the relative importance of the individual metrics of performance.[15]

The three objectives of this study were: 1) To develop metrics of performance for a novel virtual reality ACDF simulation. 2) To employ artificial neural networks to classify participants' expertise based on their performance in the simulated task. 3) To examine the ability of our neural network to outline the relative weights of specific metrics in the determination of expert performance in this virtual reality spinal procedure. This novel methodology has the potential to aid in the understanding of components of surgical expertise and contribute to the paradigm shift towards competency-based surgical training. To our knowledge, this is the first study to employ artificial neural networks to gain insight into the relative weights of teachable performance metrics in a virtual reality surgical simulation.

METHODS

Participants

Twenty-seven participants were recruited to perform a virtual reality ACDF utilizing the Sim-Ortho platform. No participants had previous experience with this ACDF scenario on the Sim-Ortho platform. This simulator is only optimized for right-handed users which excluded three left-handed participants. One fellow and two neurosurgeons were also excluded as their practice was not primarily spinal focused. The demographics of the 21 remaining participants are outlined in Table 1. The participants were divided into three groups: Post-Resident group (4 spine surgeons and 5 spine fellows), Senior group (3 PGY 4-6 neurosurgery and 2 PGY 4-5

orthopaedics residents), and Junior group (3 PGY 1-3 neurosurgery and 4 PGY 1-3 orthopaedics residents). All participants signed consent forms approved by the McGill University Ethics Review Board.

Virtual Reality Surgical Simulator

The virtual reality simulator employed is the Sim-Ortho platform (Figure 2A) co-developed by OSSimTech™ (Montreal, Quebec, Canada) and the AO Foundation (Bienne, Switzerland).[28] The platform offers a variety of tool handles (Figure 2B), each used to simulate different surgical instruments utilized by participants for the simulated procedure (Figure 2C). The platform relies on voxel-based gaming graphics to create a hyper-realistic 3D intra-operative environment mimicking real surgical procedures (Figure 2D). The participant wears 3D glasses experiencing visual and auditory feedback when employing instruments (Figure 2E), while the haptic feedback allows multiple tissue manipulations by the instruments utilized.

Simulated Surgical Scenario

The scenario simulated in this study was an anterior cervical discectomy and fusion (ACDF). The simulation was divided into four steps: cutting the disc annulus to gain disc access, cervical discectomy with excision of disc annulus and nucleus, removal of posterior osteophytes and excision of the posterior longitudinal ligament. The neck incision and bone fusion were automatically completed by the simulator. Prior to the start of the scenario, participants were asked to provide information concerning their knowledge of the ACDF procedure on a 5-point Likert scale. All participants were provided written instruction for all steps and made

knowledgeable concerning all instruments available to complete the procedure. No time limit was imposed and the simulated task was performed in an environment devoid of distractions.

Once satisfied with their performance in a given step, participants moved on to the next step and were not allowed to return to a previous step. For standardization purposes, each step was accompanied by a restricted list of simulated instruments. The scenario began with a pre-open surgical cavity revealing the simulated patient's spine. The first step involved performing a 2cm transverse box incision exposing the disc annulus (between C4 and C5) using a No.15 scalpel. The participants then performed the cervical discectomy. For the discectomy, participants could choose between a simulated bone curette, a 2mm 45˚ pituitary rongeur or a disc rongeur. Any combination of these three instruments could be utilized for the discectomy. A 3mm diamond burr was then used to remove the posterior osteophytes from the inferior C4 and superior C5 vertebrae. In the fourth step, a nerve hook was employed to lift the posterior longitudinal ligament and a 1mm 45˚ Kerrison was used to remove it.

The focus of this manuscript is the discectomy step. This component of the ACDF was chosen due to its complex nature requiring participants to choose between three different instruments to adequately perform the discectomy. This allowed for a comprehensive evaluation of both psychomotor performance and cognitive decision making.

Raw Data Acquisition

The study methodology is illustrated in Figure 3. The simulator recorded a series of data pertaining to participant use of individual instruments throughout the procedure. This resulting information was divided into 66 variables for each tool, including time, position and angles of the simulated instruments, forces applied on specific anatomical structures and volume of any

anatomical structure removed. The raw data from every participant was analysed in Matlab (Version R2018b, The MathWorks Inc., Natick, Massachusetts, United States).

Metric Generation

Metrics of performance were developed by combining available raw data to develop a smaller and more understandable set of metrics. For example, velocity can be assessed by combining position and time. Metrics were generated in three ways: 1) Consultation with expert spine surgeons to identify the components of the discectomy surgery they believed important in performing a safe procedure. The engineers involved with the development of the Sim-Ortho platform also consulted with spine surgeons to decide which raw data could be adequately provided on the platform. 2) Metrics were derived from published work involving lumbar discectomy.[70] 3) Novel metrics were created by the authors based on different components of surgical skill.

Metric Selection

Although the metric generation step reduces the amount of data by calculating a narrow set of metrics, many of these metrics may not be useful to distinguish different levels of expertise between participants. Feeding a large number of irrelevant metrics to a neural network would introduce noise, thereby affecting the network's performance. Hence, metric selection is employed to identify and filter out the non-differentiating metrics.

In this study, we perform metric selection through stepwise regression with the built-in *stepwisefit* function in Matlab (Version R2018b, The MathWorks Inc., Natick, Massachusetts, United States).[71] The full set of metrics as well as the group labels for each set (Post-Resident,

Senior or Junior) are fed to the function. Upon completion, the function returns the optimal set of metrics. Each metric was then normalized by calculating the z-score.

Artificial Neural Network

With the aim to create a system able to assess the complex components of surgical performance, an artificial neural network was trained. A number of different artificial neural network algorithms were tested with preliminary data to select the optimal algorithm. The Bayesian Regularization Backpropagation which consists of multi-layered perceptron displayed optimal performance with preliminary data and was therefore employed for the study.

The data consisting of 21 participants with the final metrics was split into two sets where 70% was used for training (15 participants: 6 Post-Residents, 4 Seniors, 5 Juniors) and 30% for testing (6 participants: 3 Post-Residents, 1 Senior, 2 Juniors). The training group was used to train the neural network (Figure 4) in a supervised manner and the remaining data was used to test the model. This means that the algorithm is provided labels for each set of metrics, such that it knows which set of metrics belongs to which groups (e.g. Post-Resident, Senior, Junior).[8] The training and testing process is more explicitly explained in Figure 5. Following training, the remaining data was used to test the accuracy of the model. The neural network was then optimized.

Neural Network Optimization

The neural network was designed with 16 neurons in the hidden layer, corresponding to the number of input neurons. Preliminary tests revealed optimized network performance with this number. Following evidence of optimized performance from previous literature, a tan

sigmoid transfer function was employed for the hidden layer and output layer.[72] In addition, three primary parameters were manually altered to prevent saturation of the model performance. The Marquardt adjustment parameter (μ) as well as its decrease (μ_dec) and increase factors (μ_inc) was altered in an iterative manner where the final values were 0.01, 0.95 and 10, respectively. These values influence the learning rate of the algorithm in order to ensure that a global minimum error is consistently reached.[73]

Metric Importance Calculation

The neural network classifies individuals by assigning a weight to each metric, as well as the hidden neurons. Interestingly, the magnitude of each metric and neuron alters the sensitivity of each metric on the algorithm's decision making process. The Connection Weights Algorithm (further explained in Supplemental Information) was employed to determine the relative importance of each metric of performance for each group (Post-Resident, Senior, Junior).[74] A detailed explanation and rationale for the Connection Weights Algorithm are discussed in the Supplemental Information.

In a simple model without hidden layers, a larger weight means that the metric will have a higher impact on the final decision. However, in more complex models such as neural networks which incorporate hidden neurons and where each metric can theoretically have over 100 interconnected weights, specific methodologies have been developed to extract their importance. One of these is the Connection Weights Algorithm (Equation 1) which literature has shown to be superior compared to other methods.[74] This method calculates the sum of the product of the weights of each metric to hidden neuron (weights w) and hidden neuron to output

(weights v). The relative importance of each metric used by our model was calculated following this equation.

$$CWP_x = \sum_{y=1}^{m} w_{xy} v_{yz}$$

Equation 1: Connection weight product (CWP) indicates the relation importance of inputs for an artificial neural network's decision making process.

Generally, a relative importance can be calculated for each combination of inputs (metrics of performance) and outputs (groups), as each input may have a different importance for Juniors, Seniors or Post-Residents. In addition, the magnitude of the relative importance allows metrics to be ranked. The sign of the Connection Weight Product (CWP) indicates whether the input should be most positive or more negative to increase the likelihood of a specific classification. For example, if the CWP of an input is positive for the Post-Resident group, a more positive value for the corresponding input increases the probability of Post-Resident classification. However, if the CWP of an input is negative for the same group, a more negative value for the corresponding input increases the probability of Post-Resident classification. Importantly, as all inputs have been normalized by z-score calculations, a more positive input indicates a metric above the mean and a more negative input indicates a metric below the mean.

RESULTS

Metrics of Performance

Performance metrics developed for the discectomy components of the simulated ACDF were divided into four categories: safety, efficiency, motion and cognitive and are outlined in

Table 2. A total of 369 metrics were calculated for each participant. Following removal of metrics that contained a value of zero for all participants, 333 remained. Following metric selection with the *stepwisefit* function, 13 significant metrics were selected, and they are described in Table 3. Some of these metrics are specific to certain instruments used for the discectomy. The authors deemed it important for the model to consider the tool choice of participants as this would influence one's metric score for this particular instrument. Hence, three binary metrics were added, one for each tool (bone curette, pituitary rongeur and disc rongeur) where a value of 1 corresponds to the use of an instrument and 0 corresponds to no use of the respective instrument, for a total of 16 metrics.

Surgical Performance Classification

The dataset was divided into a training set (15 participants, 16 metrics) and testing set (6 participants, 16 metrics). The artificial neural network was then optimized (Marquardt adjustment parameter (μ)=0.01, μ decrease ratio=0.95, μ increase ratio=10) and trained over 10,000 iterations. A training accuracy of 100% and a testing accuracy of 83.3% were achieved. A breakdown of the networks training and testing performance are displayed in confusion matrices in Figure 6A and 6B, respectively.

Metric Importance

The decision-making process of the artificial neural network is more sensitive to alterations in certain metrics of performance for each group. Using the Connection Weights Algorithm, the relative importance of each metric of performance was calculated. The ranked metrics for the Post-Resident, Senior and Junior groups are displayed Tables 4, 5 and 6,

respectively. Figure 7 illustrates a visual comparison of the Connection Weight Products and Figure 8 displays the relative importance of each metric. Interestingly, the number of contacts with the spinal dura is in the top three most important metrics for all three groups. Following the signs of the connection weight product for this metric (Post-Resident CWP: -3.01; Senior CWP: 4.83; Junior CWP = -2.71), a decrease in the contacts with the spinal dura increases the likelihood of being classified in the Post-Resident group, while an increase in the number of contacts on the spinal dura increases the likelihood of classification in the Senior group. The maximum amount of force applied on the left posterior longitudinal ligament (Post-Resident CWP = 1.91; Senior CWP = 4.17; Junior CWP = -5.24) is also highly ranked across all groups. A larger force application increases the likelihood of Post-Resident or Senior classification, while a lower force application increases the likelihood of Junior group classification.

DISCUSSION

Participant Classification with Artificial Neural Networks

The artificial neural network used in this study was able to classify the training group correctly with 100% accuracy and was 83.3% accurate in classifying the testing group. This suggests that the model has the ability to differentiate participants in the preselected testing groups. The Bayesian Regularization Backpropagation algorithm employed is particularly advantageous as it creates robust models which are less likely to overfit.[75] Overfitting occurs when a network decision-making is too closely fit to its training data and does not generalize to new participants.[55]

Patterns in Relative Metric Importance

An important study finding is the ability of the network to rank the importance of a specific metric in the final assessment of expertise in a virtual reality procedure. Generating this data allows surgical educators to address a number of new questions. Should surgical educational paradigms predominantly focus on making sure that specific metrics that contribute extensively to expertise take precedence in any surgical training system? The analysis of the neural network uncovered some general patterns in some performance metrics. For example, our network would preferentially classify a new participant in the Senior group as opposed to the Post-Resident or Junior groups if they had large numbers of instrument contacts with the dura. The Senior group contacted the dura more frequently than either the Junior or the Post-Resident group with the Post-Resident group having the least number of dural contacts. One explanation for these findings may be lack of ACDF experience in the Junior group caused more hesitation when approaching the dura or other structures with instruments, thus explaining their low number of dural contacts. The Post-Resident group, possibly associated with their greater appreciation for this safety component of ACDF procedures, appears to have modulated their behaviour after completing residency resulting in decreased instrument dural contact when compared to the Seniors group. A different pattern was observed with the maximum force applied to the left posterior longitudinal ligament. For this metric the network associates higher forces with the Senior and Post-Resident groups. Maximum force application for the Post-Resident group lies in an intermediate range compared to the Senior and Junior groups. This could also be due to the Post-Resident group altering their behaviour since residency resulting in decreased instrument force application in the posterior longitudinal ligament region associated with safety concerns associated with high force application in this area. The results of this study appear to be consistent with previous finding from our research group and with a virtual reality tumour

resection model where both safety and efficiency were found to be hallmarks of expert performance.[76]

The authors do not believe that virtual reality surgical training combined with neural networks replaces present methods of training for an anterior cervical discectomy. However, the information from this study may provide surgical educators with new perspectives on critical aspects of expert performance during cervical discectomies, as well as providing newer metrics for self-guided learning through an artificial intelligence-powered feedback platform.

Application of Neural Network in Education

Education for complex tasks has become a growing application of interest for artificial neural networks.[77] Attention is focused on the development of tools that employ neural networks to breakdown and better understand the factors that differentiate learner performance. Unlike traditional teaching methods which may weigh all components of a task relatively equally, the neural networks allow for each component (or metric of performance) of a task to be weighed individually, offering a more holistic understanding of expertise. However, some questions remain unanswered. Should the training of the junior residents performing the ACDF scenario on this simulator be focused on training to the Senior level of performance, or that of the Post-Resident group? Should significant time be spent on training all metrics, including those which are less important (i.e. less likely to influence a participant being classified into a particular group as defined by the neural network), or should training follow best practice in adult learning theory such as cognitive load theory and focus only on a small set of critical metrics (those most likely to influence participant classification as defined by the neural network) at any given time?[78] Besides the safety metrics discussed, other metrics vary in importance between groups.

The average pitch of the bone curette while in contact with the disc nucleus is a metric of high importance (ranked 2nd) for both Junior and Senior groups. However, this metric is ranked as one of the lowest of importance for the Post-Resident group (ranked 15th). This poses a similar question, should educators focus on the teaching of junior residents about the proper angles of their instruments even though it is ranked low in the Post-Resident group? The fact that this metric is important for the intermediate group (Seniors) however, may indicate that angles of the bone curette is a part of the arch of learning to mastery level for safe instrument use. A recent paper in surgical simulation supports PGY-specific benchmarks.[79] However, this study only assessed time metrics on a simpler Fundamentals of Laparoscopic Surgery simulator and may not be directly applicable to more complex metrics in a virtual reality environment. More research is needed to address this important issue.

In the future, the network presented in this study will be employed to develop an automated and more personalized feedback platform for virtual reality surgical training. Once this platform is in place, we will be able to determine whether feedback on the selected metrics has the ability to determine and truly improve performance.

LIMITATIONS

The Sim-Ortho virtual reality surgical simulator incorporates an advanced gaming engine, but fails to represent the continually changing operating room environment. First, the Sim-Ortho platform's ACDF involves 4 distinct components and this study was only focused on the cervical discectomy component. This was done to develop and assess the potential of artificial neural networks before applying them to other components of the ACDF procedure. Ongoing studies are now investigating the other 3 components of the ACDF procedure which

will allow a more comprehensive assessment of the components of expert performance. Second, the simulator is one-handed which does not allow for the quantification of bimanual skills which have been shown to be important in differentiating expertise level in previous studies and an important component of a proposed model for virtual reality surgical performance.[80,81] The simulator is also only applicable for right-handed users limiting the ability to assess left-handed participants. Left and right handed ergonomics have been shown to be different in virtual reality trials.[20] Third, multiple variables were controlled to simplify the interpretation of participants' surgical performance, including the specific instruments to be used in each component of the scenario. Fourth, the study involved a small a priori defined sample size from a single institution. Hence, it is difficult to confidently extend our results to larger populations. Prospective testing of the neural network with a large sample size from multiple institutions is required to assess its accuracy and generalizability. We believe that these further studies will improve the ability of the network to correctly predict an individual's surgical psychomotor skills and then be useful in training that individual to a defined level of expertise. Lastly, the Sim-Ortho platform is not physics-based, unlike other simulators such as the NeuroVR.[2] Hence, the simulated tissue may not respond to deformations as accurately.

CONCLUSION

This study achieved all three of our objectives: to develop performance metrics, to employ artificial neural networks to classify participants' expertise, and to outline the relative weights of specific metrics. A new virtual reality simulator, based on a gaming engine, was employed to develop novel performance metrics for an ACDF procedure. A robust artificial neural network was designed to classify three groups of participants based on expertise. Insight

into the relative importance of specific metrics of performance was outlined. The novel methodology and results presented have the potential to aid in the understanding of components of surgical expertise and contribute to the paradigm shift towards competency-based training for surgery.

THESIS CONCLUSIONS

Summary

The thesis describes the use of artificial intelligence employing artificial neural networks to differentiate three levels of expertise amongst a sample of surgeons and residents performing a cervical discectomy on a virtual reality simulator. The three thesis objectives were achieved. First, 369 metrics of performance related to safety, motion, efficiency, and cognitive decision-making were successfully developed for the simulated scenario. Second, an artificial neural network was designed and trained to achieve 83.3% classification accuracy with three groups of participants: Post-Residents (consulting spine surgeons and spine fellows), Seniors (PGY 4-5 orthopaedic residents and PGY 4-6 neurosurgery residents), and Juniors (PGY 1-3 residents in neurosurgery and orthopaedic surgery). Third, the neural network was examined to reveal the importance of specific metrics of performance.

Although further real-life testing with a large dataset of residents and surgeons from multiple institutions is required to validate the potential of the neural network for training, two primary applications of the network exist: summative and formative assessment of surgical trainees. Summative assessment involves making a single determination of one's skills. By providing a new individual's classification into one of three groups, the network may perform well in a summative assessment role. Formative assessment on the other hand is a longitudinal process where the goal is to improve one's skills. The network may also be useful for this sort of assessment by providing information on the importance and performance of individual metrics.

The research presented also raises questions about the manner in which surgical trainees should be trained. The network revealed interesting patterns in some important metrics as has been presented in the discussions component of the submitted paper. The issue of whether to

train residents in the Junior group to the metrics performance of the Senior group or to train directly to the Post-Resident group needs to be further studied. These studies will not only help understand the underpinning factors associated with surgical expertise but may allow surgical expertise to be gained earlier in residency training thus adding to the educational armamentarium of surgical competency.

It is also important to consider the social and ethical implications of employing artificial intelligence in surgical education. Surgical trainers and trainees must be able to both understand and trust artificial intelligence systems involved in the evaluation process. The AI paradigms developed will need to be integrated into the ongoing apprenticeship model to be successful. The paradigm shift that will be associated with the development of new artificial intelligence models will need to occur over time with the understanding that these systems will need to be rigorously tested. The author believes that various artificial intelligence systems may play an important role in complementing the current methods of teaching. For example, an automated AI-powered platform for surgical simulation could be used to teach basic technical skills to residents in environments where access to surgical trainers is limited. Alternatively, these could be used by experienced surgeons who wish to regain familiarity with less common surgeries. Psychomotor skills are only part of a holistic set of skills required to be a competent surgeon, but it is a part where effective teaching methods are lacking. This study advocates for the surgical community to regard AI as “Augmented Intelligence” rather than “Artificial Intelligence”, as the system will aid, and augment the knowledge and understanding of surgical educators, rather than replace them.

In summary, it is evident that the neural networks offer a new method of understanding surgical expertise, while at the same time engaging the surgical education community to question the current model of teaching.

Future Directions

As previously mentioned, the study involves a small sample size with participants from a single institution. Hence, the trained network may not be representative of a much larger population of spine surgeons and residents. Although this network is an important first step for the use of artificial intelligence in surgical training, further testing of the neural network with a larger sample size from multiple institutions is required to assess its generalizability.

Additionally, this study solely focusses on the discectomy component of the ACDF. Future studies will aim to develop artificial neural networks to reveal important performance metrics for the other steps of the procedure, including cutting of the disc, removal of the osteophytes with the burr, removal of the posterior longitudinal ligament with the nerve hook and Kerrison.

The network developed in this study needs further development to play a role in the teaching of the technical skills involved in the cervical discectomy component of the ACDF procedure. Further investigation is required to develop an automated teaching platform powered by the network developed in this study. Working alongside members of the Neurosurgical Simulation and Artificial Intelligence Learning Centre, the author has designed a framework for the development of an automated feedback platform for virtual reality neurosurgical simulation powered by artificial intelligence. The candidate is first author on this manuscript which is currently under the review of the PLoS One editorial board. As such, future work will focus on

applying this framework to the neural network presented herein followed by a validation study to test its effectiveness in improving surgical performance.

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. *Annals of surgery.* 2005;241(2):364.
2. Delorme S, Laroche D, DiRaddo R, Del Maestro RF. NeuroTouch: A Physics-Based Virtual Simulator for Cranial Microneurosurgery Training. *Operative Neurosurgery.* 2012;71(suppl_1):ons32-ons42.
3. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. *Medical teacher.* 2010;32(8):638-645.
4. Azarnoush H, Alzhrani G, Winkler-Schwartz A, et al. Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. *International journal of computer assisted radiology and surgery.* 2015;10(5):603-618.
5. 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.
6. 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). *Surgical innovation.* 2015;22(6):636-642.
7. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine.* 2006;27(4):12.

8. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering.* 2007;160:3-24.

9. Haykin S. *Neural networks.* Vol 2: Prentice hall New York; 1994.

10. Vedula SS, Ishii M, Hager GD. Objective assessment of surgical technical skill and competency in the operating room. *Annual review of biomedical engineering.* 2017;19:301-325.

11. 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. *Journal of Surgical Education.* 2019.

12. Ershad M, Rege R, Fey AM. Meaningful Assessment of Robotic Surgical Style using the Wisdom of Crowds. *International Journal of Computer Assisted Radiology and Surgery.* 2018;13(7):1037-1048.

13. Kerwin T, Wiet G, Stredney D, Shen H-W. Automatic scoring of virtual mastoidectomies using expert examples. *International Journal of Computer Assisted Radiology and Surgery.* 2012;7(1):1-11.

14. Rhienmora P, Haddawy P, Suebnukarn S, Dailey MN. Intelligent dental training simulator with objective skill assessment and feedback. *Artificial intelligence in medicine.* 2011;52(2):115-121.

15. Richstone L, Schwartz MJ, Seideman C, Cadeddu J, Marshall S, Kavoussi LR. Eye metrics as an objective assessment of surgical skill. *Annals of surgery.* 2010;252(1):177-182.

16. Sewell C, Morris D, Blevins NH, et al. Providing metrics and performance feedback in a surgical simulator. *Computer Aided Surgery.* 2008;13(2):63-81.

17. Huang J, Payandeh S, Doris P, Hajshirmohammadi I. Fuzzy classification: towards evaluating performance on a surgical simulator. *Studies in health technology and informatics.* 2005;111:194-200.

18. Loukas C, Georgiou E. Multivariate autoregressive modeling of hand kinematics for laparoscopic skills assessment of surgical trainees. *IEEE Transactions on Biomedical Engineering.* 2011;58(11):3289-3297.

19. Liang H, Shi MY. Surgical skill evaluation model for virtual surgical training. Paper presented at: Applied Mechanics and Materials2011.

20. Jog A, Itkowitz B, Liu M, et al. Towards integrating task information in skills assessment for dexterous tasks in surgery and simulation. Paper presented at: 2011 IEEE International Conference on Robotics and Automation2011.

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

22. Megali G, Sinigaglia S, Tonet O, Dario P. Modelling and evaluation of surgical performance using hidden Markov models. *IEEE Transactions on Biomedical Engineering.* 2006;53(10):1911-1919.

23. Murphy TE, Vignes CM, Yuh DD, Okamura AM. Automatic motion recognition and skill evaluation for dynamic tasks. Paper presented at: Proc. Eurohaptics2003.

24. Franzese CB, Stringer SP. The evolution of surgical training: perspectives on educational models from the past to the future. *Otolaryngologic Clinics of North America.* 2007;40(6):1227-1235.

25. Miles SH. *The Hippocratic Oath and the ethics of medicine.* Oxford University Press; 2005.

26. Hamdorf JM, Hall JC. Acquiring surgical skills. *BJS.* 2000;87(1):28-37.

27. Nguyen L, Brunicardi FC, DiBardino DJ, et al. Education of the Modern Surgical Resident: Novel Approaches to Learning in the Era of the 80-Hour Workweek. *World Journal of Surgery.* 2006;30(6):1120-1127.

28. Ruikar DD, Hegadi RS, Santosh K. A systematic review on orthopedic simulators for psycho-motor skill and surgical procedure training. *Journal of medical systems.* 2018;42(9):168.

29. Kneebone R. Simulation in surgical training: educational issues and practical implications. *Medical education.* 2003;37(3):267-277.

30. Jarman BT, Miller MR, Brown RS, et al. The 80-hour work week: will we have less-experienced graduating surgeons? *Current surgery.* 2004;61(6):612-615.

31. Gelfand DV, Podnos YD, Carmichael JC, Saltzman DJ, Wilson SE, Williams RA. Effect of the 80-hour workweek on resident burnout. *Archives of surgery.* 2004;139(9):933-940.

32. Satava RM. Historical review of surgical simulation—a personal perspective. *World journal of surgery.* 2008;32(2):141-148.

33. Martin S. Simulators Give Astronauts Glimpse of Future Flights. 2016; https://www.nasa.gov/feature/simulators-give-astronauts-glimpse-of-future-flights. Accessed April 4, 2019.

34. Ziv A, Ben-David S, Ziv M. Simulation based medical education: an opportunity to learn from errors. *Medical teacher.* 2005;27(3):193-199.

35. Stefanidis D, Sevdalis N, Paige J, et al. Simulation in surgery: what's needed next? *Annals of surgery.* 2015;261(5):846-853.

36. Denson JS, Abrahamson S. A computer-controlled patient simulator. *Jama.* 1969;208(3):504-508.

37. Rehder R, Abd-El-Barr M, Hooten K, Weinstock P, Madsen JR, Cohen AR. The role of simulation in neurosurgery. *Child's Nervous System.* 2016;32(1):43-54.

38. Satava RM. Virtual reality surgical simulator. *Surgical endoscopy.* 1993;7(3):203-205.

39. Mabrey JD, Reinig KD, Cannon WD. Virtual reality in orthopaedics: is it a reality? *Clinical Orthopaedics and Related Research®.* 2010;468(10):2586-2591.

40. Tsai M-D, Hsieh M-S, Jou S-B. Virtual reality orthopedic surgery simulator. *Computers in biology and medicine.* 2001;31(5):333-351.

41. Alaraj A, Charbel FT, Birk D, et al. Role of cranial and spinal virtual and augmented reality simulation using immersive touch modules in neurosurgical training. *Neurosurgery.* 2013;72(suppl_1):A115-A123.

42. Pfandler M, Lazarovici M, Stefan P, Wucherer P, Weigl M. Virtual reality-based simulators for spine surgery: a systematic review. *The Spine Journal.* 2017;17(9):1352-1363.

43. Sutherland C, Hashtrudi-Zaad K, Sellens R, Abolmaesumi P, Mousavi P. An augmented reality haptic training simulator for spinal needle procedures. *IEEE Transactions on Biomedical Engineering.* 2013;60(11):3009-3018.

44. Abe Y, Sato S, Kato K, et al. A novel 3D guidance system using augmented reality for percutaneous vertebroplasty. *Journal of Neurosurgery: Spine.* 2013;19(4):492-501.

45. Winkler-Schwartz A, Bajunaid K, Mullah MAS, et al. Bimanual Psychomotor Performance in Neurosurgical Resident Applicants Assessed Using NeuroTouch, a Virtual Reality Simulator. *Journal of Surgical Education.* 2016;73(6):942-953.

46. Gallagher AG, Satava R. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. *Surgical Endoscopy and Other Interventional Techniques.* 2002;16(12):1746-1752.

47. Gélinas-Phaneuf N, Choudhury N, Al-Habib AR, et al. Assessing performance in brain tumor resection using a novel virtual reality simulator. *International journal of computer assisted radiology and surgery.* 2014;9(1):1-9.

48. LeCun Y, Bengio Y, Hinton G. Deep learning. *nature.* 2015;521(7553):436.

49. Schmidhuber J. Deep learning in neural networks: An overview. *Neural networks.* 2015;61:85-117.

50. Metz C. Turing Award Won by 3 Pioneers in Artificial Intelligence. *The New York Times.* March 27, 2019, 2019: B3.

51. Zhang Q-s, Zhu S-C. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering.* 2018;19(1):27-39.

52. Lipton ZC. The mythos of model interpretability. *arXiv preprint arXiv:160603490.* 2016.

53. Strickland E. The digital fingerprints of brain disorders [News]. *IEEE Spectrum.* 2018;55(5):12-13.

54. Girouard M. AI in the Exam Room: Combatting Physician Burnout and Improving Clinical Care. 2018; https://rctom.hbs.org/submission/ai-in-the-exam-room-combatting-physician-burnout-and-improving-clinical-care/. Accessed April 4, 2019.

55. Deo RC. Machine learning in medicine. *Circulation.* 2015;132(20):1920-1930.

56. Amato F, López A, Peña-Méndez EM, Vaňhara P, Hampl A, Havel J. Artificial neural networks in medical diagnosis. In: Elsevier; 2013.

57. Benrimoh D, Fratila R, Israel S, et al. Aifred Health, a Deep Learning Powered Clinical Decision Support System for Mental Health. In: *The NIPS'17 Competition: Building Intelligent Systems.* Springer; 2018:251-287.

58. Dias RD, Gupta A, Yule SJ. Using machine learning to assess physician competence: A systematic review. *Academic Medicine.* 2019;94(3):427-439.

59. Ashizawa K, MaCMahon H, Ishida T, et al. Effect of an artificial neural network on radiologists' performance in the differential diagnosis of interstitial lung disease using chest radiographs. *AJR American journal of roentgenology.* 1999;172(5):1311-1315.

60. Ikeda M, Ito S, Ishigaki T, Yamauchi K. Evaluation of a neural network classifier for pancreatic masses based on CT findings. *Computerized medical imaging and graphics.* 1997;21(3):175-183.

61. Jesneck JL, Lo JY, Baker JA. Breast mass lesions: computer-aided diagnosis models with mammographic and sonographic descriptors. *Radiology.* 2007;244(2):390-398.

62. Li G, Kim H, Tan JK, et al. Semantic characteristics prediction of pulmonary nodule using Artificial Neural Networks. Paper presented at: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)2013.

63. Matsuki Y, Nakamura K, Watanabe H, et al. Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: evaluation with receiver operating characteristic analysis. *American Journal of Roentgenology.* 2002;178(3):657-663.

64. Nakamura K, Yoshida H, Engelmann R, et al. Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks. *Radiology.* 2000;214(3):823-830.

65. Yost MJ, Gardner J, Bell RM, et al. Predicting academic performance in surgical training. *Journal of surgical education.* 2015;72(3):491-499.

66. Stevens RH, Najafi K. Artificial Neural Networks as Adjuncts for Assessing Medical Students' Problem Solving Performances on Computer-Based Simulations. *Computers and Biomedical Research.* 1993;26(2):172-187.

67. Stevens RH, Lopo AC, Wang P. Artificial neural networks can distinguish novice and expert strategies during complex problem solving. *Journal of the American Medical Informatics Association.* 1996;3(2):131-138.

68. Conati C, Porayska-Pomsta K, Mavrikis M. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. *arXiv preprint arXiv:180700154.* 2018.

69. Fountas KN, Kapsalaki EZ, Nikolakakos LG, et al. Anterior cervical discectomy and fusion associated complications. *Spine.* 2007;32(21):2310-2317.

70. Riffaud L, Neumuth T, Morandi X, et al. Recording of surgical processes: a study comparing senior and junior neurosurgeons during lumbar disc herniation surgery. *Operative Neurosurgery.* 2010;67(suppl_2):ons325-ons332.

71. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology.* 1992;45(2):265-282.

72. Karlik B, Olgac AV. Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems.* 2011;1(4):111-122.

73. Mahapatra S, Sood AK. Bayesian regularization-based Levenberg–Marquardt neural model combined with BFOA for improving surface finish of FDM processed part. *The International Journal of Advanced Manufacturing Technology.* 2012;60(9-12):1223-1235.

74. Olden JD, Jackson DA. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling.* 2002;154(1-2):135-150.

75. Burden F, Winkler D. Bayesian regularization of neural networks. In: *Artificial neural networks.* Springer; 2008:23-42.

76. AlZhrani G, Alotaibi F, Azarnoush H, et al. Proficiency performance benchmarks for removal of simulated brain tumors using a virtual reality simulator NeuroTouch. *Journal of surgical education.* 2015;72(4):685-696.

77. Livieris IE, Drakopoulou K, Pintelas P. Predicting students' performance using artificial neural networks. Paper presented at: 8th PanHellenic Conference with International Participation Information and Communication Technologies in Education2012.

78. Sweller J. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction.* 1994;4(4):295-312.

79. Hoops HE, Haley C, Kiraly LN, An E, Brasel KJ, Spight D. PGY-specific benchmarks improve resident performance on Fundamentals of Laparoscopic Surgery tasks. *The American Journal of Surgery.* 2018;215(5):880-885.

80. Azarnoush H, Siar S, Sawaya R, et al. The force pyramid: a spatial analysis of force application during virtual reality brain tumor resection. *Journal of neurosurgery.* 2017;127(1):171-181.

81. Sawaya R, Alsideiri G, Bugdadi A, et al. Development of a performance model for virtual reality tumor resections. *Journal of neurosurgery.* 2018;1(aop):1-9.

APPENDIX

TABLES

Table 1. Demographics information for three groups of participants performing the virtual reality surgical task.

	Post-Resident (9)	Senior (5)	Junior (7)
Age (years)			
Mean, SD	44.2 ±13.2	30.6 ±2.3	27.4 ±1.4
Sex			
Male	9	4	5
Female	0	1	2
Level of Training			
Neurosurgery Resident			
PGY 1-3	-	-	3
PGY 4-6	-	3	-
Orthopaedic Resident			
PGY 1-3	-	-	4
PGY 4-5	-	2	-
Spine Fellows	5	-	-
Spine Surgeons			
Neurosurgeons	2	-	-
Orthopaedic Surgeons	2	-	-
Surgical knowledge of an ACDF (self-rated, Likert scale, 1 to 5)			
Median (range)	5 (4-5)	3 (3-4)	3 (1-3)

Table 2. Metrics of performance for virtual reality simulation of anterior cervical discectomy scenario.

Metric Category	Metric List
Safety	Number of voxels (volume) removed from an anatomical structure Average force applied on anatomical structure Maximum force applied on anatomical structure
Motion	Velocity of instrument while in contact with the disc nucleus and disc annulus Acceleration of instrument while in contact with the disc nucleus and disc annulus Angles of instruments while in contact with disc nucleus and disc annulus Angular velocity of instruments while in contact with disc nucleus and disc annulus
Efficiency	Number of instrument contacts on an anatomical structure Number of instances a volume or an anatomical structure is removed with an instrument Amount of time an instrument spends in contact with an anatomical structure Total path travelled by an instrument while in contact with an anatomical structure
Cognitive	Instrument choice

Table 3: Selected metrics of performance for simulated discectomy.

Category	Label	Description
Safety	Contacts_Dura	Number of contacts with the spinal dura during discectomy
	VolumeRemoved_PLL_Right	Volume of right posterior longitudinal ligament removed
	ForceMax_PLL_Left	Maximum force applied on the left posterior longitudinal ligament
	ForceMax_DiscAnnulus_BoneCurette	Maximum force applied on the disc annulus by the bone curette
	ForceMax_LVA_PitRongeur	Maximum force applied on the left vertebral artery region by the pituitary rongeur
Motion	VelocityMean_DiscAnnulus_PitRongeur	Average velocity of the pituitary rongeur while in contact with the disc annulus
	AccelerationNumZ_DiscAnnulus_PitRongeur	Number of accelerations of the pituitary rongeur along the anterior-posterior axis while in contact with the disc annulus
	AccelerationMaxY_DiscNucleus_BoneCurette	Maximum acceleration of the bone curette along the anterior-posterior axis while in contact with the disc nucleus
	PitchMean_DiscNucleus_BoneCurette	Average pitch of the bone curette while in contact with the disc nucleus. Pitch is the rotation of the curette in up and down (scooping) motion.
Efficiency	ContactNumber_C5	Number of contacts on the C5 vertebra over the entire procedure
	CuttingNumber_C5_BoneCurette	Number of contacts on the C5 vertebra using the bone curette
	ContactTime_LVA	Amount of time spent in contact with the left vertebral artery region
	TTPLperStrokeY_DiscNucleus_DiscRongeur	Total length of individual strokes along the medial-lateral axis with disc rongeur while in contact with the disc nucleus

Table 4: Metrics of performance ranked by their relative importance for the Post-Resident group.

Rank	Category	Metric	Connection Weight Product	Relative Importance (%)
1	Safety	Contacts_Dura	-3.01	17.18
2	Safety	ForceMax_DiscAnnulus_BoneCurette	2.56	14.60
3	Cognitive	ToolChoice_BoneCurette	1.92	10.97
4	Safety	ForceMax_PLL_Left	1.91	10.93
5	Efficiency	TTPLperStrokeY_DiscNucleus_DiscRongeur	1.14	6.51
6	Efficiency	ContactTime_LVA	1.02	5.84
7	Cognitive	ToolChoice_PituitaryRongeur	0.89	5.11
8	Motion	VelocityMean_DiscAnnulus_PitRongeur	0.87	4.98
9	Efficiency	ContactNumber_C5	-0.81	4.63
10	Motion	AccelerationNumZ_DiscAnnulus_PitRongeur	0.78	4.43
11	Cognitive	ToolChoice_DiscRongeur	0.71	4.03
12	Efficiency	CuttingNumber_C5_BoneCurette	-0.66	3.79
13	Motion	AccelerationMaxY_DiscNucleus_BoneCurette	-0.38	2.16
14	Safety	VolumeRemoved_PLL_Right	-0.35	1.98
15	Motion	PitchMean_DiscNucleus_BoneCurette	0.34	1.96
16	Safety	ForceMax_LVA_PitRongeur	0.16	0.89

Table 5: Metrics of performance ranked by their relative importance for the Senior group.

Rank	Category	Metric	Connection Weight Product	Relative Importance (%)
1	Safety	Contacts_Dura	4.83	15.90
2	Motion	PitchMean_DiscNucleus_BoneCurette	-4.65	15.31
3	Safety	ForceMax_PLL_Left	4.17	13.74
4	Efficiency	TTPLperStrokeY_DiscNucleus_DiscRongeur	-3.16	10.39
5	Cognitive	ToolChoice_BoneCurette	-3.03	9.97
6	Safety	ForceMax_DiscAnnulus_BoneCurette	-2.68	8.82
7	Cognitive	ToolChoice_PituitaryRongeur	-2.06	6.77
8	Efficiency	ContactTime_LVA	-1.90	6.25
9	Efficiency	CuttingNumber_C5_BoneCurette	-1.42	4.68
10	Motion	AccelerationNumZ_DiscAnnulus_PitRongeur	-1.23	4.04
11	Efficiency	ContactNumber_C5	-0.42	1.39
12	Motion	AccelerationMaxY_DiscNucleus_BoneCurette	-0.28	0.93
13	Cognitive	ToolChoice_DiscRongeur	0.26	0.86
14	Safety	ForceMax_LVA_PitRongeur	0.14	0.46
15	Safety	VolumeRemoved_PLL_Right	-0.13	0.42
16	Motion	VelocityMean_DiscAnnulus_PitRongeur	-0.02	0.07

Table 6: Metrics of performance ranked by their relative importance for the Junior group.

Rank	Category	Metric	Connection Weight Product	Relative Importance (%)
1	Safety	ForceMax_PLL_Left	-5.24	21.55
2	Motion	PitchMean_DiscNucleus_BoneCurette	5.14	21.13
3	Safety	Contacts_Dura	-2.71	11.14
4	Efficiency	TTPLperStrokeY_DiscNucleus_DiscRongeur	1.75	7.18
5	Cognitive	ToolChoice_PituitaryRongeur	1.56	6.41
6	Cognitive	ToolChoice_BoneCurette	1.33	5.46
7	Efficiency	ContactNumber_C5	1.31	5.40
8	Cognitive	ToolChoice_DiscRongeur	-1.06	4.36
9	Motion	AccelerationMaxY_DiscNucleus_BoneCurette	-0.89	3.65
10	Efficiency	CuttingNumber_C5_BoneCurette	0.80	3.29
11	Safety	VolumeRemoved_PLL_Right	-0.60	2.47
12	Efficiency	ContactTime_LVA	0.54	2.19
13	Safety	ForceMax_LVA_PitRongeur	-0.48	1.98
14	Safety	ForceMax_DiscAnnulus_BoneCurette	0.34	1.41
15	Motion	VelocityMean_DiscAnnulus_PitRongeur	-0.33	1.35
16	Motion	AccelerationNumZ_DiscAnnulus_PitRongeur	-0.25	1.02

FIGURES

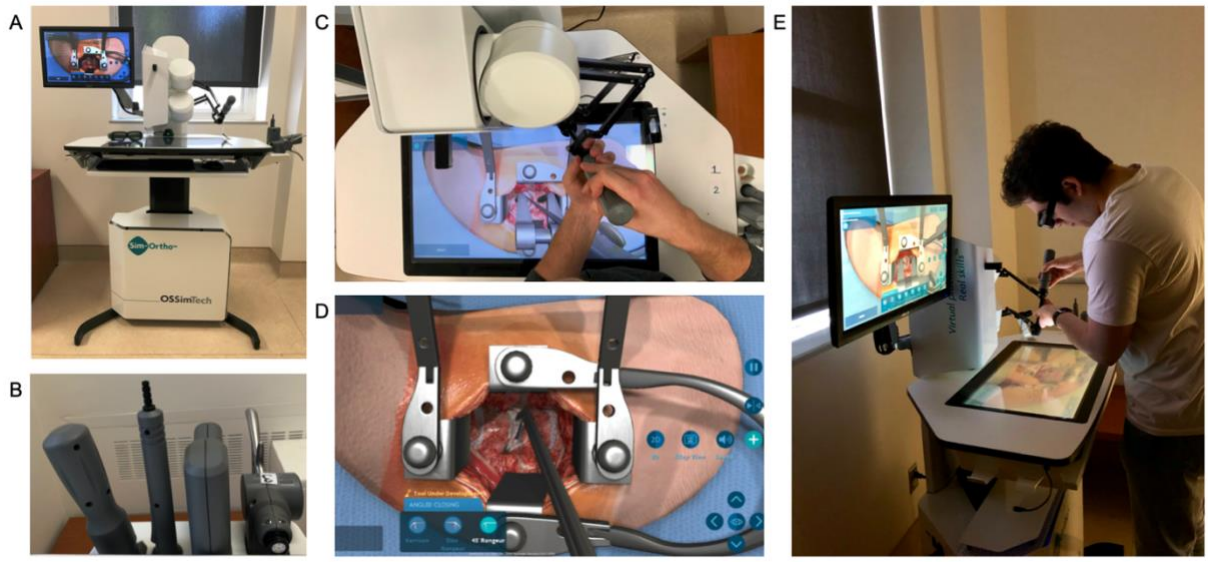

Figure 2: The virtual reality platform used to simulate an anterior cervical discectomy. (A) The Sim-Ortho platform, co-developed by OSSimTech™ and the AO Foundation is designed to simulate a number of surgical procedures. (B) A variety of instruments are available accompanied by a variety of different handles to simulate the feel of each instrument. (C) The participant holds the instrument in their dominant hand, receiving haptic feedback when interacting with anatomical structures. (D) The platform is built on a gaming engine providing very realistic 3D graphics. (E) The participant wears 3D glasses while interacting with the platform.

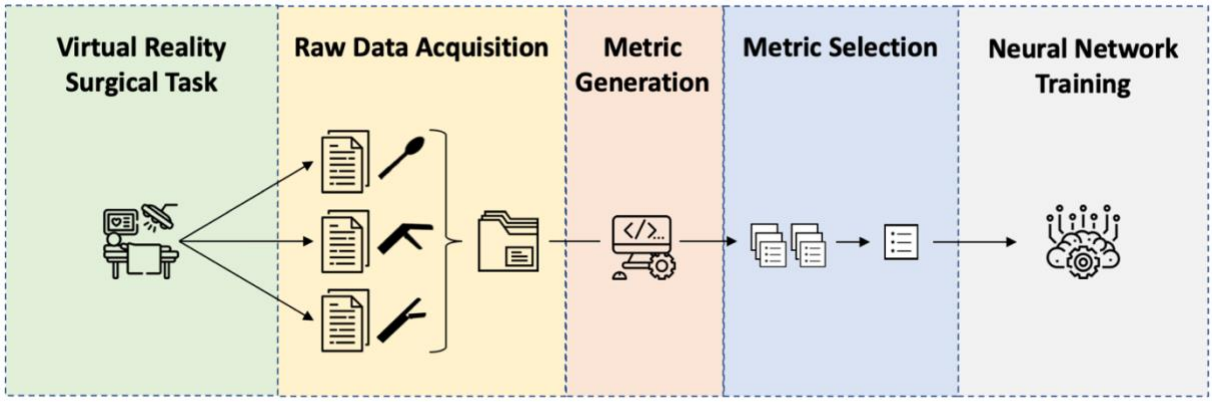

Figure 3: Methodology for the use of neural networks to assess expertise in a virtual reality surgical simulator. Users begin by performing the surgical task on the virtual reality platform. Raw data acquisition occurs as the platform creates large datasets for each instrument employed. All instrument datasets are combined into a single dataset. The large dataset can be used to generate metrics of performance for each participant. The new set of metrics can then undergo metric selection to narrow down a group of metrics able to differentiate levels of surgical expertise. The final metrics are then fed to the neural network for training and testing.

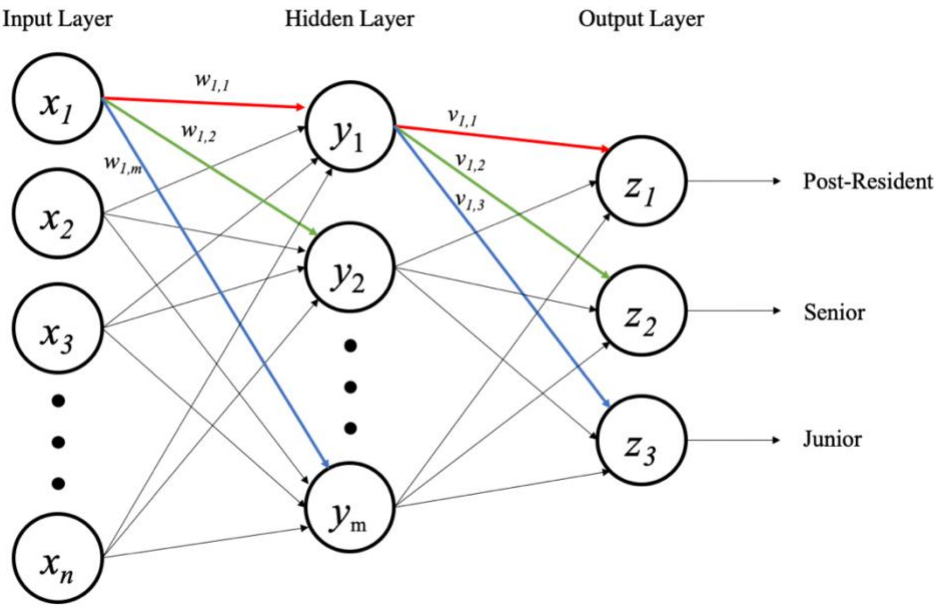

Figure 4: Simplified illustration of multi-layered perceptron used by the artificial neural network. Metrics of performance are inputs of the neural network represented by *x*. Each and every input is connected to a number of neurons in the hidden (middle) layer of the network, represented by *y*. The connection between each input and each neuron is supported by a weight (w) where a large magnitude of the weight means that the hidden neuron will be more sensitive to alterations in this specific input (*x*). Each neuron of the hidden layer is then connected to the three possible outputs, Post-Resident, Senior and Junior, represented by z1, z2 and z3, respectively. Similarly to the input layer, each neuron of the hidden layer is assigned a weight (v) to influence the output neurons. The output neuron with the value closest to 1 will be the neural network's final decision.

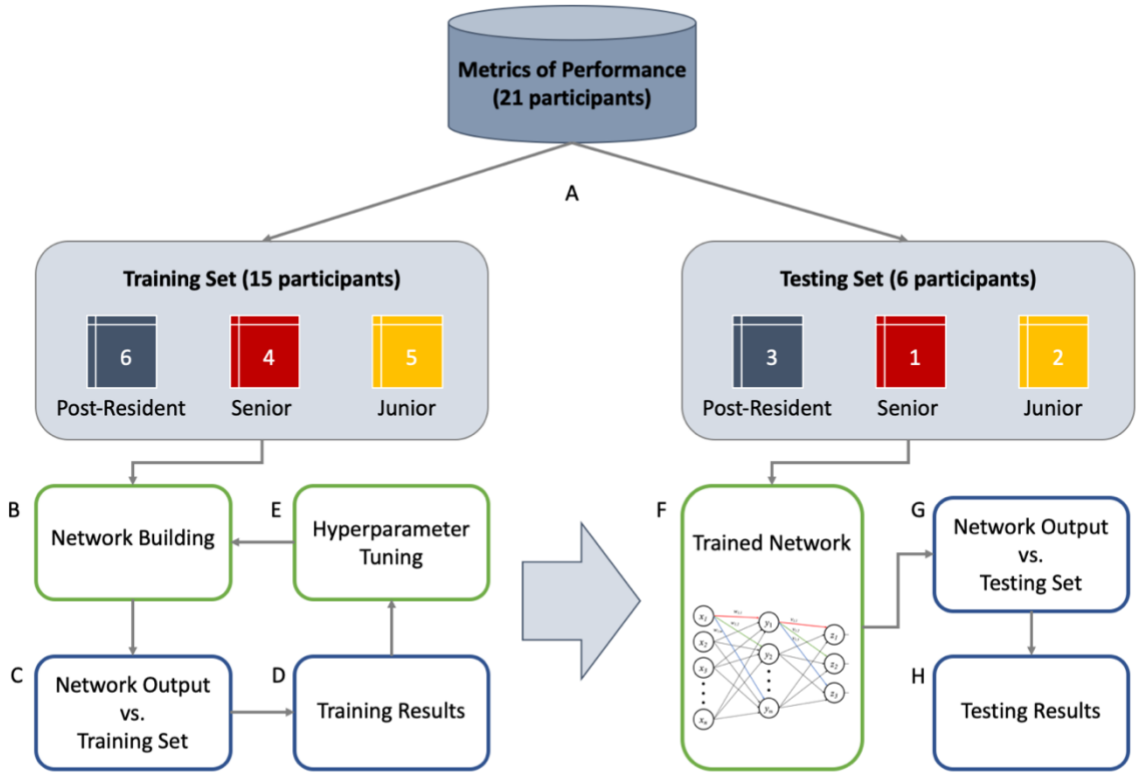

Figure 5: Training and testing of artificial neural network. (A) The final set of metrics of performance from 21 participants is split into a training and testing set. (B) The training set is used to build an initial neural network. (C) The network predicts a group for each participant, and this prediction is compared to the true grouping of each respective participant. (D) A training accuracy is calculated based on how well the network output resembles the true training set grouping. This is based on calculating a cost function for the network, where a high cost represents poor performance. (E) If the neural network performed poorly, its hyperparameters are tuned according to a set of predefined rules in an attempt to reduce cost. This process repeats until the network's cost reaches a minimum. (F) Following the iterative training process, the final trained network can now be exported and tested. The testing set, containing previously unseen

data, is fed to the network. (G) The network predicts a group for each participant in the testing set, and this prediction is compared to the true grouping of each respective participant. (H) The testing accuracy is calculated based on how well the network's output resembles the true testing test grouping. Unlike training, this process is only done once.

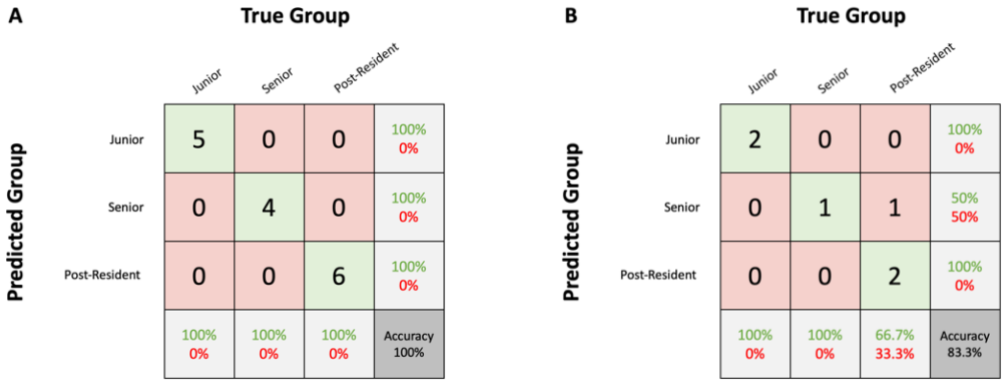

Figure 6: Confusion matrices of the artificial neural network. (A) The training group correctly classified all participants into their respective groups reaching a 100% accuracy. (B) The testing group correctly classified 5 of the 6 participants reaching an accuracy of 83.3%. One participant belonging to the Post-Resident group was incorrectly classified in the Senior group.

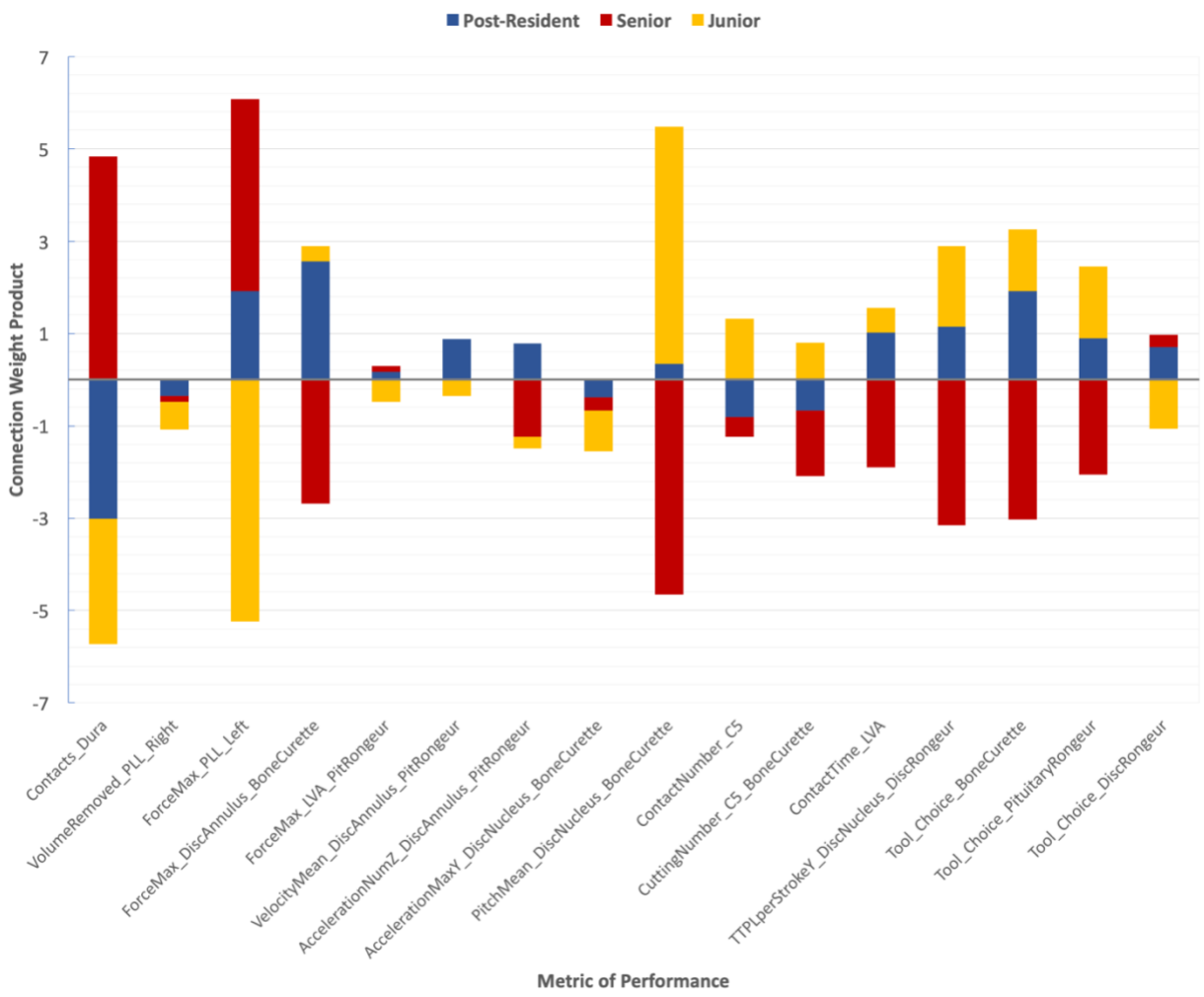

Figure 7: Connection weight products of each group for metrics of performance for the virtual reality surgical simulation. The magnitude of the connection weight product represents the relative importance of the corresponding metric in the neural network's decision to classify as participant in the corresponding group: Post-Resident (blue), Senior (red) and Junior (yellow). The sign of the connection weight product indicates whether a metric's z-score value should be positive (if sign if positive) or negative (if sign is negative) to increase the likelihood of classification in the corresponding group.

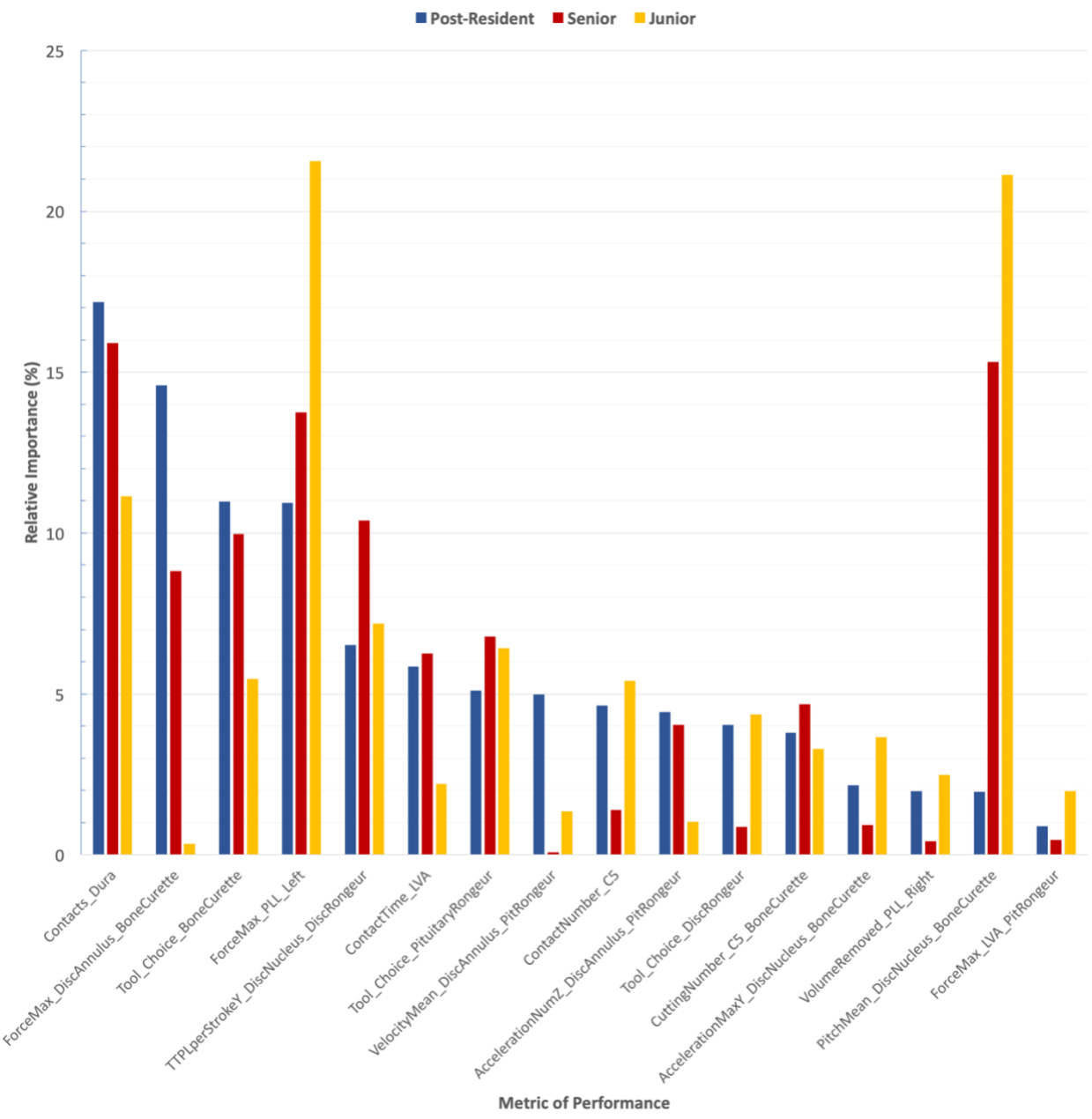

Figure 8: Relative importance of each metric of performance in the classification of participants between three groups. The relative metric importance corresponds to the magnitude of the connection weight product for each metric of performance and its corresponding group: Post-Resident (blue), Senior (red) and Junior (yellow). A higher importance indicates that fluctuations in the respective metric have a larger influence on the

neural network's decision to classify a participant in the respective group. The metrics are ranked from left to right according to descending importance for the Post-Resident group.