Development and Evaluation of a Machine Learning Approach for Application Validation of a Novel VR/AR Surgical Training Device of a Spinal Operation with Focus on Physics-Based Force Feedback

Sami Alkadri

Department of Mechanical Engineering
McGill University, Montreal

August 2024

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Doctor of Philosophy in Mechanical Engineering

© Sami Alkadri, 2024

Dedication

This thesis is dedicated to my parents, Marwan and Rim, whose endless support and sacrifices have been the cornerstone of my journey. I cannot express enough gratitude for the efforts you have made to seek a better future for us. Your unwavering belief in me has been a source of inspiration throughout this work.

I also dedicate this work to the rest of my family. Your support and faith in me have not only motivated but truly inspired me to continue learning and growing. This accomplishment is as much yours as it is mine.

Table of Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

%VAF	Percent-Variance-Accounted-For
2D	Two Dimensions
3D	Three Dimensions
ACDF	Anterior Cervical Discectomy and Fusion
AF	Anulus Fibrosus
AI	Artificial Intelligence
ANN	Artificial Neural Network
AR	Augmented Reality
CT	Computer Tomography
CWA	Connection Weights Algorithm
CWP	Connection Weights Product
$F_{Structure_{Tool}}$	Force Exerted on Structure by Tool
FEM	Finite Element Model
FPS	Frame Rate Per Second
GOALS	Global Operative Assessment of Laparoscopic Skills
IAP	Inferior Articular Process
IRB	Institutional Review Board
IVD	Intervertebral Disc
$J_{Direction_{Tool}}$	Jerk In Direction by Tool
LIF	Lumbar Interbody Fusion
MIS	Minimally Invasive Surgeries

MISS	Minimally Invasive Spinal Surgeries
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MLP	Multilayer Perceptron
MR	Mixed Reality
NP	Nucleus Pulposus
OLLIF	Oblique Lateral Lumbar Interbody Fusion
OSATS	Objective Structured Assessment of Technical Skill
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
SAP	Superior Articular Process
SFS	Sequential Forward Selection
SGD	Stochastic Gradient Descent
$sign_{a_{Direction_{Tool}}}$	Sign Changes of The Acceleration In Direction by Tool
SVM	Support Vector Machine
$T_{Structure_{Tool}}$	Torque Exerted on Structure by Tool
TACT	Technical Abilities Customized Training
TLIF	Transforaminal Lumbar Interbody Fusion
$v_{Direction_{Tool}}$	Velocity In Direction by Tool
VR	Virtual Reality

Abstract

This thesis evaluates and validates a physics-based virtual (VR) and augmented reality (AR) simulator for training in Oblique Lateral Lumbar Interbody Fusion (OLLIF) surgery, addressing the increasing demand for effective spinal surgical training. With over 80% of people experiencing back pain, leading to surgery, the precision and safety in spinal operations are paramount. Minimally Invasive Spinal Surgeries (MISS), although reducing recovery times, pose significant training challenges due to their complexity and reliance on advanced surgical skills. Traditional training methods have proven inadequate, highlighted by the limitations of the "see one, do one, teach one" model and restricted resident working hours, necessitating innovative training solutions such as VR/AR surgical simulations. However, despite the encouraging initial outcomes of VR/AR surgical training systems, thorough validation studies are essential to advance their integration into surgical education programs.

The project's main objective was achieved through a sequential validation approach, encompassing subjective assessments, machine learning analyses, and the examination of haptic feedback's role in surgical performance. The first objective established the foundational validation of the simulator – defined by face, content, and construct validity – using both subjective and objective methods. Starting with subjectively assessing the resemblance of the simulated scenario to reality, a side-by-side comparison with a cadaveric surgery is conducted to further support face and content validity of the developed simulator. Face-validity is the extent to which the developed simulation mimics reality, whereas content-validity is the extent to which it is representative of the skills required to successfully complete the real surgery. Afterwards, construct validity was established by extracting psychomotor data to derive surgical metrics that distinguish between skill levels. The second objective explored the application of artificial neural networks (ANNs) for surgical performance classification, developing a novel methodology for feature importance in multilayered ANNs and addressing challenges related to limited datasets through data augmentation and transfer learning. This objective also demonstrated the utility of combining neural network depth with traditional statistical analysis breadth for a holistic understanding of surgical expertise. The third objective focused on the significance of accurate physics-based force feedback, particularly in the "gaining access" step of the OLLIF surgery, where visual feedback is limited. It introduced a novel approach to objectively measure haptic fidelity, highlighting how realistic forces directly influence surgical simulation performance and, by extension, training outcomes.

The dissertation presents original contributions to surgical simulation validation, machine learning in surgical performance analysis, and the biomechanics of surgery, offering a comprehensive framework for future surgical simulator development and validation. The methodologies developed have broader implications, paving the way for more effective training tools in high-risk fields. This work advocates for standardized guidelines in surgical simulation validation and emphasizes the necessity of incorporating accurate haptic feedback in training to enhance the safety and effectiveness of MISS.

Résumé

Cette thèse valide un simulateur en réalité virtuelle (VR) et augmentée (AR) basé sur la physique pour la formation en Oblique Lateral Lumbar Interbody Fusion (OLLIF), répondant à la demande croissante pour une formation chirurgicale spinale efficace. Avec plus de 80 % des personnes souffrant de douleurs dorsales, menant à la chirurgie, la précision et la sécurité des opérations spinale sont primordiales. Les chirurgies spinale minimalement invasives (CSMI), bien qu'elles réduisent les temps de récupération, posent des défis de formation significatifs en raison de leur complexité et de leur dépendance à des compétences chirurgicales avancées. Les méthodes de formation traditionnelles se sont révélées inadéquates, soulignées par les limites du modèle ‘ voir un, faire un, enseigner un’ et les heures de travail limitées des résidents, nécessitant des solutions de formation innovantes telles que les simulations chirurgicales VR/AR. Cependant, malgré les résultats initiaux encourageants des systèmes de formation chirurgicale VR/AR, des études de validation approfondies sont essentielles pour faire avancer leur intégration dans les programmes d'éducation chirurgicale.

L'objectif principal du projet a été atteint grâce à une approche de validation séquentielle, englobant des évaluations subjectives, des analyses d'apprentissage automatique et l'examen du rôle de la rétroaction haptique dans la performance chirurgicale. Le premier objectif a établi la validation fondamentale du simulateur – définie par la validité de ‘face;, de ‘content’ et de ‘construct’ – en utilisant à la fois des méthodes subjectives et objectives. Commençant par une évaluation subjective de la ressemblance du scénario simulé à la réalité, une comparaison côte à côte avec une chirurgie cadavérique est réalisée pour soutenir davantage la validité de ‘face’ et de ‘content’ du simulateur développé. La validité de ‘face’ est la mesure dans laquelle la simulation

développée imite la réalité, tandis que la validité de ‘content’ est la mesure dans laquelle elle est représentative des compétences requises pour réussir la chirurgie réelle. Ensuite, la validité de ‘construct’ a été établie en extrayant des données psychomotrices pour dériver des métriques chirurgicales qui distinguent les niveaux de compétence. Le deuxième objectif a exploré l'application des Artificial Neural Networks (ANNs) pour la classification de la performance chirurgicale, développant une nouvelle méthodologie pour l'importance des caractéristiques dans les ANNs multicouches et abordant les défis liés aux ensembles de données limités grâce à l'augmentation des données et à l'apprentissage par transfert. Cet objectif a également démontré l'utilité de combiner la profondeur des réseaux neuronaux avec la largeur de l'analyse statistique traditionnelle pour une compréhension holistique de l'expertise chirurgicale. Le troisième objectif s'est concentré sur l'importance de la rétroaction haptique basée sur la physique précise, en particulier dans l'étape "d'accès" de la chirurgie OLLIF, où le retour visuel est limité. Il a introduit une nouvelle approche pour mesurer objectivement la fidélité haptique, soulignant comment les forces réalistes influencent directement la performance de la simulation chirurgicale et, par extension, les résultats de la formation.

La dissertation présente des contributions originales à la validation de la simulation chirurgicale, à l'analyse de la performance chirurgicale par apprentissage automatique et à la biomécanique de la chirurgie, offrant un cadre complet pour le développement et la validation futurs des simulateurs chirurgicaux. Les méthodologies développées ont des implications plus larges, ouvrant la voie à des outils de formation plus efficaces dans les domaines à haut risque. Ce travail préconise des directives standardisées pour la validation des simulations chirurgicales et souligne la nécessité d'incorporer un retour haptique précis dans la formation pour améliorer la sécurité et l'efficacité des CSMI.

Acknowledgments

I would like to first express my deepest gratitude to my supervisor Professor Mark Driscoll for his invaluable guidance, unwavering support, understanding, patience, and motivation throughout my doctoral studies. I have been very blessed to have Professor Driscoll as a mentor who believes in my capabilities and potential, providing me with opportunities that have allowed me to excel. I am also grateful to Dr. Rolando Del Maestro, my co-supervisor, whose support, mentorship, and assistance have been instrumental over the years. His exemplary role as a mentor and model of professional and academic excellence is something I aspire to attain in my own career. My sincere thanks also extend to my thesis committee members, Dr. Ahmed Aoude and Dr. Fiona Zhao, for their insightful feedback and guidance which have significantly contributed to the development of my work.

I am fortunate to have been part of two remarkable research groups: the Musculoskeletal Biomechanics Research Lab and the Neurosurgical Simulation and Artificial Intelligence Learning Centre. The friendship, support, and collaborative spirit of the members of these labs have been vital to my research and personal growth. Specifically, I would like to acknowledge Khaled El-Monajjed, Trevor Cotter, and Tianqi Wang from the Musculoskeletal Biomechanics Research Lab, as well as Recai Yilmaz, Abdulmajeed Albeloushi, Nicole Ledwos, Nykan Mirchi, and Aiden Reich from the Neurosurgical Simulation and Artificial Intelligence Learning Centre, who have each contributed to some of the manuscripts contained within this thesis.

I would like to extend my appreciation to our project partners: CAE Healthcare, Depuy Synthes, and the Natural Science and Engineering Research Council of Canada for their generous support throughout the development of this project. Additionally, I am profoundly grateful to

McGill University and the Vadasz Scholarship Committee for awarding me the Vadasz Meda award, which played a crucial role in funding my research.

To all the staff surgeons, fellows, surgical residents, and everyone who contributed in various capacities to the collection of materials and data, the design, construction, and execution of the simulator study, the analysis of data, and the preparation of this thesis, your contributions have been essential to the completion of this work. Your collective efforts have not only facilitated my research but have also enriched my doctoral experience in countless ways.

Lastly, beyond the professional and academic support, I must express my profound gratitude to my parents and family, whose unwavering belief in my capabilities and unconditional support have been my constant source of strength and inspiration. Their faith in my decision to pursue a PhD has been a driving force behind my endeavors, for which I am eternally thankful.

Contribution to Original Knowledge

This thesis represents the development of a novel framework encompassing innovative concepts for validating a physics-based lumbar interbody fusion simulator. The original scholarship and distinct contributions to knowledge detailed within this work include:

1- A novel validation approach by conducting a side-by-side cadaver and simulator trial. This methodology not only validates the novel minimally invasive Oblique Lateral Lumbar Interbody Fusion simulator but also contributes to the construction of novel surgical performance metrics, setting a new standard in the field of minimally invasive surgical simulation (Article 1: Face, Content, and Construct Validity of a Novel VR/AR Surgical Simulator of a Minimally Invasive Spine Operation).

2- First adaptation and validation of a novel approach for identifying feature importance in multilayered Artificial Neural Networks (ANNs). This methodology enhances the understanding of surgical performance classifications, offering a groundbreaking perspective on how machine learning can contribute to surgical excellence (Article 3: Utilizing a Multilayer Perceptron Artificial Neural Network to Assess a Virtual Reality Surgical Procedure and Article 4: Unveiling Surgical Expertise Through Machine Learning in a Novel VR/AR Spinal Simulator: A Multilayered Approach Using Transfer Learning and Connection Weights Analysis).

3- Development of a comprehensive framework to conduct and analyze performances on surgical simulations. This framework addresses the challenge of small datasets typically collected from a single study center by employing a strategic combination of machine learning techniques such as data augmentation, feature selection, and transfer learning.

Additionally, it introduces a methodology that combines the complexity of ANNs with the simplicity of traditional statistical analyses, providing a more holistic understanding of complex surgical performance profiles across different skill levels (Article 1: Face, Content, and Construct Validity of a Novel VR/AR Surgical Simulator of a Minimally Invasive Spine Operation and Article 4: Unveiling Surgical Expertise Through Machine Learning in a Novel VR/AR Spinal Simulator: A Multilayered Approach Using Transfer Learning and Connection Weights Analysis).

4- Development of a novel method for quantifying the impact of realistic, physics-based force feedback on surgical training within minimally invasive VR/AR environments. This method offers a unique approach to assessing the potential for negative training effects in surgical simulations, advancing the understanding of how haptic feedback influences surgical performances (Article 5 Impact of Physics-Based Force Feedback on Surgical Training and Performance in VR/AR Simulations).

Contribution of Authors

I, Sami Alkadri, confirm that I am the primary author and contributor of the research work contained within this thesis. At the beginning of each chapter, specific acknowledgements are given to those who assisted in the work presented. The work presented in this thesis was completed under the supervision and co-authorship of Professor Mark Driscoll and Dr. Rolando Del Maestro. I prepared and executed the face, content, and construct validation study involving surgical staff surgeons, fellows, and residents, who performed the virtual surgical scenarios on the newly developed VR/AR surgical simulator. This included the design and development of the study protocol, the face and content validity questionnaires, and the generation of novel surgical performance metrics through consultation with expert surgeons and literature reviews. Furthermore, I designed and conducted the machine learning analyses for classifying surgical performances and identifying feature importance. This included programming and coding all necessary machine learning scripts, such as the novel adaptation of the connection weights algorithm, data augmentation, feature selection, and transfer learning. Moreover, I recognized the untapped potential of a data-driven machine learning approach to objectively assess the impact of haptic fidelity on virtual surgical performance. Therefore, I initiated and conducted a final study, aiming to deepen the understanding of haptic feedback's role in enhancing surgical training within VR/AR environments. In conducting the study, I designed the new force profiles and made all the necessary modifications to the haptic algorithm.

Additional contributions are recognized from colleagues who collaborated on various aspects of this thesis:

- Khaled El-Monajjed: Collaborated to incorporate the novel surgical performance metrics into the Gaining Access step module utilized throughout Articles 1, 4, and 5. Notably, Khaled developed the main code that was altered and modified to generate new force profiles for Article 5.
- Trevor Cotter: Assisted in incorporating the novel surgical performance metrics into the Facetectomy, Discectomy, and Annulotmy steps module that were employed throughout Articles 1 and 4.
- Tianqi Wang: Co-authored the side study outlined in Article 2, where he contributed the initial conceptualization, acquisition, and interpretation of data including statistical analysis, complementing the new conceptualization, acquisition, methodology, and analyses developed by the current author.
- Abdulmajeed Albeloushi: Assisted in participant recruitment for the final study outlined in Article 5.
- Recai Yilmaz, Nicole Ledwos, Nykan Mirchi, and Aiden Reich: Co-authored the side study in Article 3, conducting participant recruitment and data acquisition, in addition to assisting in conceptualization and critical review of the manuscript.

These collaborations have enriched the research and enhanced the scope and impact of the findings presented in this thesis.

Introduction

Surgical interventions require proficient physicians capable of utmost precision in critical situations. Effective surgical training reduces risks to patients and increases the success rate of the procedure. Spinal surgical training is particularly significant due to the increasing demand and the intricate nature of such procedures. Spinal surgeries address and correct numerous conditions, the most frequent being low back pain, a notably prevalent and economically burdensome condition in the western world, which is predominantly managed through spinal surgeries [1]. Over 80% of individuals experience back pain during their lifetime, making it the primary reason for activity restriction and the third most frequent reason for surgical interventions in the US [1]. Consequently, there has been a significant rise in the number of spinal procedures in recent decades [1]. Additionally, spinal surgeries present greater intricacy than many other surgical fields due to the proximity of neurological components to the spine, which notably amplifies both the frequency and variety of complications. Surgical errors can induce both immediate and secondary complications; more specifically, adverse events during spinal surgery may lead to neural damages, pulmonary embolism, neurological impairments, or infections in the surgical area of the spine [2]. Some patients might also experience chronic back pain post-surgery, and in severe scenarios, subsequent operations may be necessary [3]. Recent innovations in surgical techniques gave rise to minimally invasive spinal surgeries (MISS), where surgeons approach the target area via small incisions as demonstrated in Figure 0-1. These procedures reduce blood loss and foster quicker recoveries [4]. While such techniques often lead to reduced post-operation pain and shorter hospitalizations compared to the classical invasive spinal surgeries, they still pose issues for both the patients and the medical professionals [1, 4]. In MISS, surgeons depend on the laparoscope[1] for visual direction and require the use of long surgical instruments through narrow openings, increasing the complexity and the potential for complications [5]. Such complications stem from challenges faced during the operation, such as diminished depth perception from the 2D view of the surgical area, tool tremors, and the pivot effect of instruments around the incision point. Such challenges are attributed to a surgeon's insufficient skill set stemming from the lack of proper training [4]. Therefore, to truly harness the benefits of MISS, surgical trainees must refine their abilities and achieve ambidexterity through rigorous surgical training methods [4].

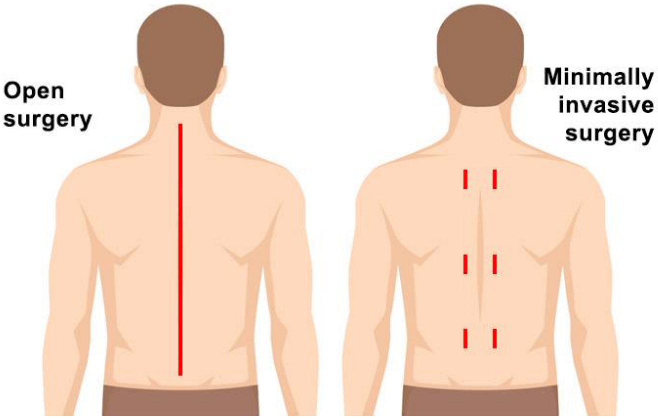


Figure 0-1 involve smaller incisions, which lead to less blood loss and enhanced recovery rates when compared to traditional open surgeries. However, the complexity and technical challenges associated with MISS can significantly raise the risk of complications (adapted from [6]).

From the late 1800s, surgical residency training programs have employed the conventional master-apprentice model known as "see one, do one, teach one". In this framework, a surgical

[1] A thin, tube-like instrument equipped with a light and camera used in minimally invasive procedures.

trainee is expected to observe a specific procedure, before being able to perform that procedure themselves, and ultimately gain the sufficient proficiency to instruct another trainee [7]. Research has highlighted the shortcomings of this model, especially for complex surgeries; it has been linked to elevated teaching costs, a higher error rate, and suboptimal patient outcomes [4]. Moreover, it has been shown that staff surgeons often hesitated to let trainees significantly participate in sophisticated surgeries, such as intricate spinal procedures, out of concern for patient safety [4]. Consequently, many trainees concluded their residencies without the necessary competence to autonomously conduct certain surgical procedures [4]. Compounding the issue, the Accreditation Council for Graduate Medical Education (ACGME) both in the US and Europe set guidelines limiting surgical resident working hours. While intended to protect patients by preventing fatigued residents from participating in surgeries, this mandate further restricted trainees' hands-on experience, leading to diminished surgical proficiency and, paradoxically, poorer patient outcomes [4]. Therefore, there's an urgent need for innovative, cost-efficient surgical training methods to adequately and safely train residents [8].

Virtual reality (VR) and augmented reality (AR) surgical simulators have been rapidly adopted as a more objective method of training and evaluating surgical technical skills, especially when compared to conventional training methods [9, 10]. These modules offer a safe and controlled environment, helping residents to sharpen their surgical competencies, especially when dealing with sensitive procedures [8]. The incorporation of automated grading systems not only created a chance to evaluate a trainee's proficiency but also to pinpoint and enhance areas needing further refinement. Furthermore, the integration of haptic technology into VR/AR setups, allowed trainees to acquire a tangible understanding of surgeries before transitioning to the operation room.

Such haptic feedback amplifies the authenticity of the training experience, especially when incorporating realistic force responses [8]. Despite the advancements of VR simulators in the surgical field, the development of surgical simulators for MISS was lagging [8]. Until recently, spinal simulation training had sparse research dedicated to create specialized spinal surgical simulators [8]. Moreover, the increasing demands for spinal surgeries led to the continuous need to refine both the surgical techniques and the skills of the surgeons. As a result, efforts were directed to design novel MISS simulators with accurate haptic feedback that prioritize patient safety and recovery [11, 12]. One such simulator is the platform developed by our group in collaboration with CAE and DePuy Synthes to train orthopaedic and neurosurgeons on a novel Oblique Lateral Lumbar Interbody Fusion (OLLIF) surgery. However, even with the promising preliminary results exhibited by VR/AR surgical training systems, proper application validation studies of the simulator systems are required to further encourage its adaptation to surgical curriculums [13]. In this thesis, the validation performed – which may be termed as application validation – encompasses face, content, and construct validation specific to VR/AR surgical simulators. Application validation ensures that the simulation is not only realistic but also effectively evaluates the intended skills and outcomes in a medical training context. In contrast, engineering validation studies generally refer to the process of verifying and validating that a product, system, or component meets the set design requirements and specifications. This type of validation is more focused on technical accuracy, performance metrics, reliability, and safety standards. It ensures that all engineering design criteria are met, and the system performs as expected under specified conditions. While both types of validation are crucial, they serve different purposes. Engineering validation is concerned with the functional and technical aspects of a system, ensuring it works correctly and reliably. Application validation, particularly in the context of surgical simulations, focuses on the educational training and evaluation efficacy, ensuring that the simulation effectively replicates real-world scenarios and accurately assesses the users' skills.

The general objective of this doctoral project is to establish the application validity of a physics-based VR/AR spinal surgical simulator in training and assessing surgical trainees on the OLLIF procedure. The main objective of the thesis was established by the sequential validation of different aspects of the novel simulator platform using varying methods that ranged from subjective assessments to machine learning analyses to haptic fidelity analysis of spinal tissue surgical forces.

The following dissertation consists of seven chapters as demonstrated in Figure 0-2. Chapter 1 includes the relevant literature that was applicable in this thesis, followed by the overall objectives and the corresponding hypotheses (Chapter 2). The main objective was attained by the conception of three main manuscripts and two additional manuscripts in side-studies as outlined in Chapter 3, Chapter 4, and Chapter 5. The themes discussed are consolidated in a general discussion in Chapter 6. The dissertation ends with conclusions and future perspectives in Chapter 7.

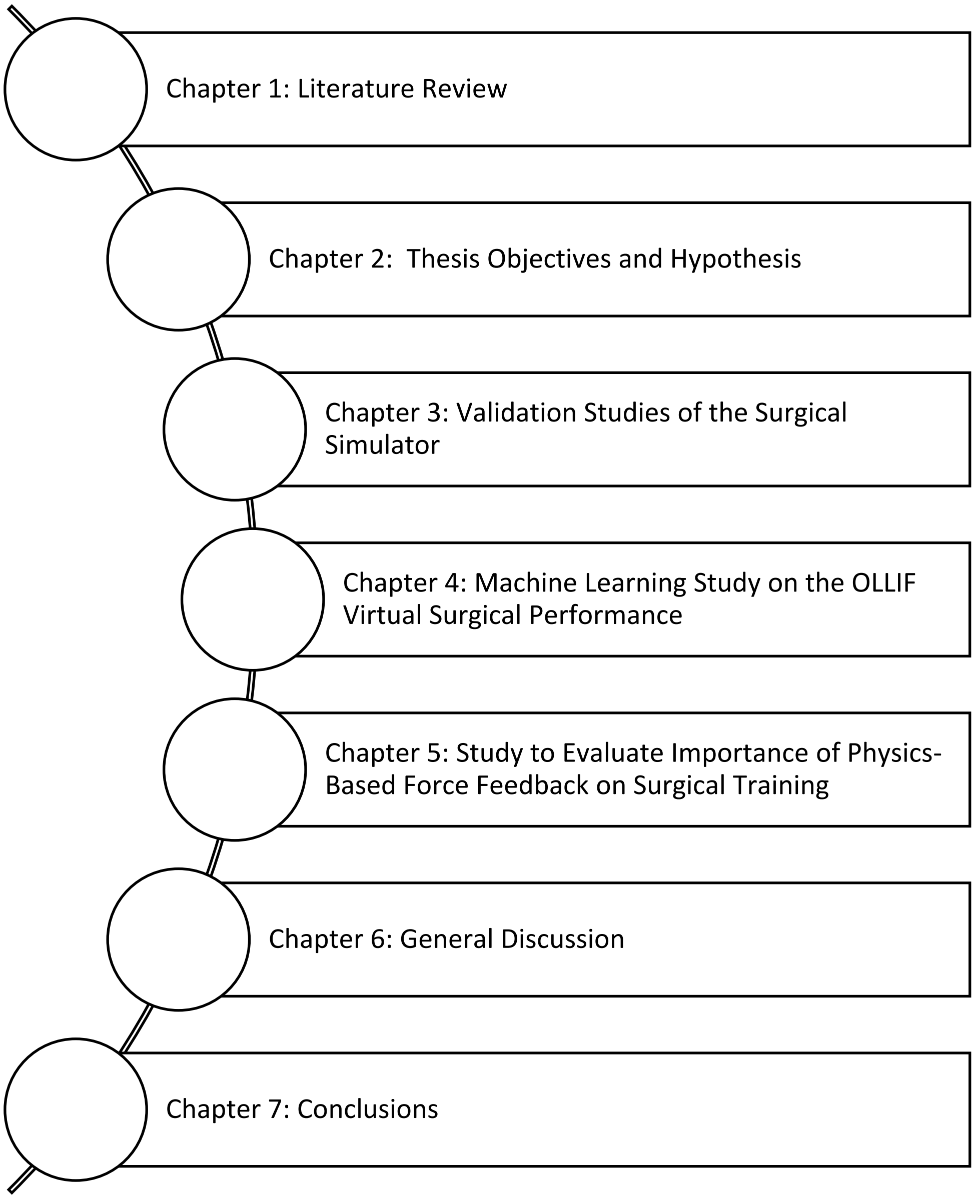

Chapter 1: Literature Review
Chapter 2: Thesis Objectives and Hypothesis
Chapter 3: Validation Studies of the Surgical Simulator
Chapter 4: Machine Learning Study on the OLLIF Virtual Surgical Performance
Chapter 5: Study to Evaluate Importance of Physics-Based Force Feedback on Surgical Training
Chapter 6: General Discussion
Chapter 7: Conclusions

Figure 0-2 Thesis List of Chapters

Chapter 1. Literature Review

1.1 VR/AR Surgical Simulators & Validation Studies

The use-case of VR/AR simulators in surgical training and evaluation has been extremely impactful in recent years. As briefly highlighted in the Introduction, their ability to simulate complex surgical scenarios underscores their potential benefits for skill development and risk-free surgical practice and evaluation. Nevertheless, rigorous validation is paramount to their integration into surgical curriculums, requiring detailed testing and assessment of the different components of the training platforms. To successfully understand and develop an effective validation process of VR/AR surgical simulators, they must be deconstructed and broken down into their primary concepts, which include theories pertaining to virtual and augmented reality, surgical training techniques, and the underlying simulation mechanisms. Each of these areas requires exploration to fully grasp their contributions and interplay within the specific context of VR/AR surgical simulators. This literature review section defines and briefly traces the historical development of the mentioned concepts, then it discusses the general principles of simulation validation, focusing on the gold standards of validating VR/AR surgical simulators: face, content, and construct validity.

1.1.1 Surgical Training: A Brief Review

Traditionally, surgery was seen as a separate discipline from general medicine. In fact, surgery was regarded as a manual craft not requiring a medical degree [14]. This difference set surgical training on a path distinct from medical training, embedding it in the master-apprentice model similar to craftsmanship training. Here, a trainee would closely observe and learn from a seasoned practitioner or a “master of the craft” [14]. Such a system had its flaws, notably, the significant risks posed to patients, given that they were often exposed to inexperienced residents-in-training [14]. Furthermore, this model inherently lacked standardization. The proficiency of surgical trainees was determined subjectively by their mentors, resulting in a diverse range of training methodologies, based on the mentor's knowledge and techniques [15]. To address these inconsistencies, Dr. William Halstead introduced the "graduated responsibility" model in the United States, aiming to standardize and structure surgical training, which created the foundations of the modern surgical training model [16]. Under this system, residents progressively gain more responsibility as they advance in their training years. While Dr. Halstead's model, known as the Halstedian approach, marked a significant transformation in surgical education, it wasn't without its shortcomings. Particularly, both safety concerns and residents completing their training without adequate hands-on experience persisted, issues that are further exacerbated by restrictions on residents' working hours as previously highlighted in the Introduction section of this thesis [8].

Moreover, another significant challenge within surgical training programs has been identifying the training primary objective: whether the aim is for trainees to achieve surgical expertise or mere competency. To address this issue, initial steps involved clarifying what defines "surgical expertise". Reports from the 1990s characterized expertise based on years of experience, specialty board certifications, and academic rankings or duties [17]. However, subsequent research challenged these criteria, noting their weak correlation with actual clinical performance. Surprisingly, a systematic review even suggested a negative relationship between years of practice and clinical performance, indicating that extended experience might inversely impact performance [18]. Such findings underscored the absence of a universally accepted definition of surgical expertise. Contemporary definitions, endorsed by major North American bodies like the Royal

College of Surgeons and Physicians of Canada and the Accreditation Council on Graduate Medical Education in the US, describe surgical expertise as achieving a combination of technical skills and "other skills". These secondary skills encompass professionalism, communication, collaboration, leadership, health advocacy, scholarship, and medical expertise, all essential for delivering "adequate" patient care [17]. Paradoxically, the use of the term "adequate" rather than "excellent" in these definitions highlights the focus towards competency-based training objectives rather than achieving expertise [17]. In fact, current surgical assessments focus on measuring surgical competency and generally adopt one of two strategies: the behaviorist approach, focusing on distinct behaviors and skills, or the holistic approach, considering a broader array of combined attributes [17]. Although future directives in surgical training strive to incorporate holistic training programs, the inherent difficulty in gauging competencies within the "other skills" category has caused many programs and surgical assessment tools to lean towards the behaviorist method. Existing tools, like the Objective Structured Assessment of Technical Skill (OSATS) and the Global Operative Assessment of Laparoscopic Skills (GOALS), strive to quantify and evaluate surgical skills. While OSATS is used to assess general open surgeries, GOALS is developed primarily for minimally invasive surgeries (MIS). For optimal validity and reliability, these evaluations typically involve robust expert reviewers for critical assessments [19, 20]. These evaluations can occur in real-time during procedures or posteriori using recorded video. However, in addition to the mentioned challenges presented by the current surgical training models – such as safety concerns and inadequate practice durations – these assessment tools struggle with inherent human biases and the absence of quantifiable feedback metrics. These challenges further catalyzed the adoption of simulation-based training in surgical education for technical skills training and evaluation. Simulation-based training addresses these challenges by providing a controlled, risk-free environment, with ample training hours and the opportunity to incorporate precise surgical performance metrics that may shed light to skills that define surgical expertise that surpass mere competency.

1.1.2 Simulation Technologies with focus on VR and AR: A Brief Overview

Simulation, at its core, represents the methodological reproduction of real-world scenarios or systems within a controlled setting, allowing for observation, training, or experimentation without the direct consequences of real-world interaction [21]. Simulations have been used historically across multiple disciplines, from aeronautics to medicine, to train, test, and improve various skills and strategies. The use of simulation can be traced back to ancient times, when scale models and figurines served as rudimentary forms of representing medical operations, warfare strategies, and architectural planning [22]. As a matter of a fact, the earliest method to safely train surgeons without risking patient safety was through the use of cadavers and synthetic models that were used as early as 600 B.C. to simulate one of the earliest recorded operations [23]. Cadaver training, which has been utilized to simulate anatomy and tissue fidelity in invasive procedures throughout history, remains in use today. However, it is associated with several limitations, such as high costs, difficulty in procurement, and a limited time of use [8].

The 20th century marked a turning point in simulation's evolution. The introduction of computers in the 1950s facilitated the development of digital simulations, allowing complex systems and phenomena to be modeled with unprecedented detail [22]. The 1960s and 1970s saw the growth of flight simulators, showcasing the potential for training in high-risk professions [22].

In the late 20th century, technological leaps, especially with the emergence of advanced microchips, led to significant enhancements in computing power, storage, graphics, and interaction design. This collective progress laid the groundwork for immersive virtual environments. By the end of the century and into the 21st, this resulted in the rise of VR and AR simulations, which integrated the physical and digital worlds, offering unparalleled levels of interactivity and realism.

In general, VR systems create a completely immersive environment, isolating the user from the actual physical surroundings [24]. Historically, VR was anchored in the concept of "presence", emphasizing the immersive feeling of being in a different environment regardless of the used hardware technology. In attaining the goal of full immersion, two dimensions were constructed to evaluate the effectiveness of VR systems: vividness (how rich the environment feels) and interactivity (the user's ability to manipulate and influence the VR space) [25]. Deriving from these dimensions, comprehensive definitions describe VR's essence as focusing on an almost-real experience characterized by a virtual world providing immersion, sensory feedback, and interactivity [22].

AR is a rapidly evolving technology that blends the physical world with virtual elements [22] . Unlike VR, which immerses users in an entirely virtual environment, AR superimposes virtual objects onto the real world, viewable through transparent screens or wearables like the Google Glass (Google, Menlo Park CA) [26]. Some definitions even suggest the possibility of viewing virtual and real objects concurrently using different eyes. Sherman and Craig [22] articulate this concept by suggesting that one eye might perceive the real world while the other visualizes virtual components. A key feature of AR is its capacity to enhance real-world perception by unveiling attributes otherwise invisible, such as subcutaneous structures, bone anatomy, and even microscopic tumor deposits [26]. This technology has found profound applications in surgery, from fluoroscopy-based guidance to 3D Computer Tomography (CT) reconstructions, enabling real-time visualizations that aid precise surgical actions like osteotomy cuts and implant placements [26]. Recognizing the convergence and overlapping functionalities of VR and AR, the combination is often categorized under Mixed Reality (MR), which has been especially useful in surgical training. In this setting, the trainee interacts with a physical entity, yet their field of view includes both real and virtual elements [26]. This blend offers more sophisticated, high-fidelity simulations, particularly beneficial for simulating complex surgical procedures.

1.1.3 Medical Simulators: A Brief Examination

Surgical and medical training has seen significant advancements over the years with the introduction of innovative simulation methods. Broadly classified, there are three primary types of surgical and medical simulators: physical (or benchtop) simulators, VR simulators, and the most recently introduced MR simulators [4, 27]. Physical simulators offer direct interaction, manifesting as manikins, laparoscopic box trainers, and other tactile models [4]. Such systems trace back to the 1960s with the introduction of the modern manikin for anesthesia training, representing a major leap in hands-on training approaches [27-30]. The early 1980s marked another milestone with the development and adoption of refined fidelity simulators within training curriculums. Notably the Comprehensive Anesthetic Simulation Environment (CASE) and the Gainesville Anesthetic Simulator (GAS) were developed [31]. Taking inspiration from aviation training models, CASE incorporated team-based realistic environments for crisis management into the anesthesia curriculum. These simulators were carefully designed, resembling patient-like appearances while integrating computer chips that could mimic and react to vital signs based on interactions with the user [31]. Simultaneous to the improvement of the fidelity of simulators, the rise of patient-based simulations, which involve learners interacting with simulated patients to replicate real-life clinical scenarios, began to gain traction in the 1970s. The Massachusetts General Hospital introduced the pioneering computer-based simulator for clinical encounters, later granting access to institutions like Ohio State University and the University of Illinois. By 1973, the University of Wisconsin further refined this concept, laying the foundation for computerized examinations that eventually led to the establishment of the Objective Structured Clinical Examinations (OSCE). Through OSCEs, institutions could simulate standardized situations, assessing student performance in both competence and confidence. Both these developments led the stage to the development of the more advanced digital-based simulations.

On the virtual frontier, the 1990s signaled the introduction of VR simulators, marking a profound shift from physical to digital. These simulators, wholly computer-based, offered users an immersive surgical experience, letting them interact with simulated anatomical structures on a screen. The initial VR simulations covered simpler tasks; however, with technological advancements in the 21st century, it was evolved to depict more complex surgical scenarios. Cutting-edge VR simulations began leveraging finite element models (FEM) to deliver visual feedback, particularly in capturing realistic deformations. Delorme, et al. [32] present a prime example of this integration through the NeuroVR, initially known as the NeuroTouch. Tailored predominantly for neurosurgical training, the NeuroVR employs finite element systems to produce incredibly detailed and dynamic 3D visuals. These graphics are capable of responding to user manipulations, enabling them to morph and adapt in real-time. Furthermore, the simulator integrates haptic feedback mechanisms, enabling users to distinguish between the tactile sensations of different tissues as they perform tasks. However, when shifting the focus to spinal surgeries, especially MISS, the scope of currently available simulations in the market is significantly limited [8]. The intricate nature of spinal anatomy, with its diverse components each having different visual appearances and tissue densities, makes it an immense task to simulate accurately. For instance, effectively replicating the spine in a virtual environment would necessitate a very complex FE modelling with variable haptic feedback to differentiate between the soft spinal tissues, such as muscles and nerves, and the hard vertebral structures. As such, early VR spinal simulators developed did not include haptic feedback and focused on simple tasks such as pedicle screw placement or lumbar puncture. Initially, to address these challenges, researchers turned to AR simulators, combining them with physical benchtop models, such as rubber mock-ups of the spine. This combination aimed to bypass the complexities of creating a high-fidelity VR spinal simulation. However, one drawback is still the low fidelity associated with the force feedback provided by such simulations. One notable attempt to embrace the full potential of VR without relying on physical benchtop models was the Sim-Ortho simulator developed by OSSimTech™. Instead of tangible models, this simulator utilized 3D glasses to enhance the immersive experience. Nevertheless, the simulator focused on open spinal surgeries providing force feedback based on gaming engines rather than physiological forces. As research progressed, the focus kept shifting towards simulating MISS. An ideal candidate is the use of MR simulations, in an attempt to exploit the best of both worlds: the precision of FE modeling from VR and the tangible interaction offered by physical benchtop models. Yet, while these advancements reflect significant improvements in surgical simulation, they are not without limitations. An obvious concern lies in the haptic forces generated by these simulators. Most are neither verified nor validated, meaning they don't necessarily replicate the actual physical forces encountered during real surgical procedures. Surgeons rely heavily on the tactile (sense of touch) and the kinesthetic (the perception of body movement and position) feedback during MISS, making the accurate replication of the force feedback crucial. Therefore, discrepancies in haptic feedback led to a disconnect between what trainees feel in simulations versus actual surgeries. To the best of the author's knowledge, there is no commercially available surgical simulator in the market that presents high visual fidelity – as defined by visual accuracy, graphical appearances, physical deformations, frame speed, and the like – while simultaneously delivering equally high-fidelity haptic feedback. Such a simulator, offering authentic tactile and kinesthetic feedback, remains a goal yet to be achieved by future surgical training tools.

1.1.4 VR/AR Surgical Simulators: Validation Studies Summarized

Broadly speaking, validity is defined as “the property of being true, correct, and in conformity with reality” [33]. In logical principles, an argument or a conclusion is valid when it relies on sound logical reasonings. The emphasis here is on the logical principles being applied to reach the conclusion. However, if the underlying assumptions are wrong, the outcome will be far from reality despite using valid logical reasonings in the process. Similarly, testing the fundamental property of any measuring instrument or test requires that it should “measures what it purports to measure” [33]. Applying the same principles to the context of simulation validity, for a simulation to serve its purpose effectively, it not only has to closely resemble the real-world scenario but also assess the intended skills with precision. As described by Cook and Hatala [34], simulations, especially in medical training, must display realism (face validity), while measuring the desired skills to be assessed (content validity). Face validity ensures that a simulation appears to be a valid way of measuring its intended purpose at a superficial level, while content validity verifies if the simulation genuinely evaluates the pertinent facets of performance [35]. Referring back to the definition of validity, a simulation that posses a high face validity with very accurate physical and graphical representation but with poor content validity falls short of its primary objective. This demonstrates the foundational aspects of face and content validity in simulations. In fact, these validations carry over to the more advanced simulations based on VR.

In addition to the mentioned dimensions of validity, VR simulations also rely on the concept of immersion. Immersion, as highlighted by Slater, et al. [36], is a pivotal factor in VR, enhancing its effectiveness by providing an engulfing experience that closely mirrors reality. Immersion contains aspects of face validity to ensure realism of the virtual environment but also consists of aspects pertaining to the technological hardware that provides the realistic graphical feedback frequency to maintain user engagement [37]. Maintaining the immersive experience in VR is contingent on the feedback frequency and responsiveness of the VR system. Any lag or discrepancies can not only impede the learning experience but also induce discomfort, such as simulation sickness [37]. Kourtesis, et al. [37] emphasize the necessity of seamless feedback, positing that technological competence and consistent feedback frequency are preconditions for an impactful VR experience. This poses a natural limitation linked to existing computing capabilities on the extent of graphical fidelity and visual realism achievable without inducing lag. Thus, for VR simulations, achieving the right balance of face validity, content validity, and immersion is paramount.

Moving towards the specific context of the validation of VR/AR surgical simulations, face and content validity are defined as the extent to which the developed simulation environment mimics the real surgery, and the extent to which the developed system is representative of the skills required to successfully complete the real surgery, respectively [38]. Moreover, additional validation metrics emerge in the context of surgical simulators, which are borrowed from surgical education assessment theories, namely: construct validity, concurrent validity, and predictive validity. Construct validity entails assessing an instrument based on how well its test items reflect the specific quality, ability, or trait they were designed to gauge [33]. For surgical simulators, this means ensuring the simulations can sufficiently and accurately distinguish between different surgical expertise levels. Specifically, there should be discernible differences in performance between novice, intermediate, and experienced surgeons. Such differentiation is pivotal not only for surgical assessment but also during training through monitoring the progress of junior surgeons, aiming for them to reach the performance levels of their more seasoned counterparts. Furthermore, this stage facilitates the creation of unique, quantifiable metrics which, when accurately developed, may serve as clear learning objectives for surgeons in training. Concurrent validity evaluates how closely the outcomes of a newly developed test align with those of an established gold standard. In surgical simulations, this could involve contrasting the performance outcomes on the simulator with those derived from established, validated surgical assessment tests such as OSATS and GOALS. On the other hand, predictive validity assesses whether skills acquired on the simulator yield better results in real surgical settings, therefore leading to better patient outcomes. While the other validity steps are crucial in establishing the utility of the simulator in surgical training and assessment, predictive validity is the most likely test to provide clinically meaningful assessment of the simulation [33]. It is the sole validation step that focus on clinical outcomes. The process of determining both concurrent and predictive validity necessitates extensive multi-center research studies. This involves evaluating surgeons of diverse expertise and specialties across multiple institutions, followed by tracking their simulation training and practical surgical outcomes [39]. It's essential to note that while these validation procedures, especially related to predictive validity, are critical for the broad integration of simulators into educational curriculums, they should be approached only after the foundational validation criteria—namely face, content, and construct validity—are solidly in place.

The current literature offers limited studies that showcase predictive and/or concurrent validity for medical and surgical simulations. A meta-analysis of 83 studies revealed that only 5% demonstrated predictive validation for surgical simulations, 24% exhibited concurrent validation, with the majority (60%) emphasizing face, content, or construct validation [40]. Even though this analysis dates back to 2010, more recent literature reviews still identify only a handful of studies focusing on predictive or concurrent validation for surgical simulators. One example is a recent study published in 2023, which examined the impact of VR simulator training on the technical thrombectomy performance of interventional radiologists [41]. The study utilized an already validated simulator with published face, content, and construct validity studies [42, 43]. This study involved interventional radiologists and residents from three distinct centers and demonstrated predictive validation. Still, recent 2022 meta-analyses on neurosurgical simulation validations emphasize the scarcity of predictive validation studies in surgical simulations, attributing this gap to the logistical challenges of long-term follow-ups, particularly in multi-center studies [44]. A 2018 systematic review on validated sinus surgery simulations revealed that the majority of studies centered on face, content, and construct validations, with only a single study focusing on predictive validation [45]. Such findings underscore the importance of foundational validation – face, content, and construct validity – for VR/AR surgical simulators, emphasizing that the development of such simulators is still in its infancy. By serendipity, conducting studies for face, content, and construct validation is logistically simpler than those for concurrent and predictive validations, and they can often be undertaken within the same recruitment process.

Face and content validity use subjective assessments as they are established using questionnaires. Both of these validation steps rely on the evaluations of the training system by expert surgeons recruited to perform the simulated surgical scenario followed by completing a Likert-scale questionnaire designed to capture the required validity [9, 38]. A statistical approach is often used to rate the consensus among experts and trainees on certain aspects of the simulator pertaining to both face and content validity [9, 38]. Comparing the consensus between the experts and trainees is often used to analyze the change in perspective with surgical experience [9]. This also allows for detailed analyses of validity that pinpoints aspects of the simulator that are adequately developed, requires further improvements, or require a complete change [9].

As discussed, construct validity refers to the ability of the simulator to distinguish between different levels of surgical expertise [9, 46, 47]. It is an objective validation step that relies on the enormous sets of data generated from the interactions of the user during the simulated task. Such data are often transformed into surgical performance metrics that play an important role in not only establishing construct validity, but also in assessing and training surgical trainees. The use of statistical analyses is the gold standard for establishing construct validity [9, 46, 47]. Statistically significant differences in the scores among experts and trainees on the generated surgical performance metrics highlight the ability of the simulator to adequately differentiate between levels of surgical expertise.

Alongside construct validity, there's an increasing emphasis on delving into analyzing aspects of surgical performance that differentiate levels of expertise [12]. This requires a detailed examination of the surgical performance metrics to capture even the subtlest differences that uniquely define surgical expertise. As such, machine learning has been recently coupled to surgical simulators for the objective of deconstructing composites of surgical performance [48-53].

1.2 Artificial Intelligence: An Overview

Artificial intelligence (AI) can be described as the field of computer science dedicated to creating systems capable of performing tasks that typically require human intelligence [54]. These tasks include learning, reasoning, problem-solving, perception, and language understanding. The objective of AI is to bridge the gap between human and machine capabilities by mimicking human cognitive functions and actions [55]. In general, AI encompasses fields such as machine learning (ML), expert systems, and robotics [55]. Indeed, the combination of these fields are pivotal in embodying the approaches to achieving AI presented by Russell and Norvig [55]. The authors argue that AI is historically developed to achieve either a human-centered or a rationalist[2] behaviour to attain one or more of the four goals of Acting Humanly, Thinking Humanly, Acting Rationally, or Thinking Rationally. For instance, developing an AI that aim to "Act Humanly" require a unique combination of the mentioned fields. To achieve that goal, ML is required for natural language processing and computer vision, as well as for pattern recognition and adaptation, facilitating efficient communication and visual interpretation. Expert systems contribute significantly with their knowledge representation capabilities, vital for information storage and retrieval, along with automated reasoning for logical problem-solving. Lastly, robotics plays a key role in enabling these systems to physically interact with their environment. These elements are

[2] The distinction between human and rational behavior does not imply that humans are irrational in terms of emotional instability or unsoundness. It simply recognizes human limitations: not all are chess grandmasters, and not everyone scores an 'A' on exams.

essential for an AI to pass the Turing Test, demonstrating human-like behavior. Similarly, the other approaches require different combinations of the distinct yet interconnected domains of ML, expert systems, and robotics to achieve artificially intelligent systems. The distinction among these domains lies in the process of acquiring and applying knowledge. In general, robotics use sensory and control systems to allow machines to interact with their environment, while expert systems apply predefined rules on large data to emulate the decision-making abilities of a human and/or "rational" expert. Both are distinct from ML in relying on predefined rules and information, while ML systems develop their understanding through learning from data.

1.2.1 Machine learning: A Review

ML is a term used to describe the ability of algorithms and statistical models to make classifications or decisions by identifying and learning from hidden patterns within datasets, without the need for explicit instructions [56]. Broadly, ML can be categorized into supervised, unsupervised, semi-supervised, or reinforcement learning [56, 57]. In supervised learning, a ML algorithm is built and trained using labelled data to generalize on new unseen datasets [56, 57]. Using supervised learning, a ML algorithm is trained to make predictions of either a continuous real number (regression), a discrete class label (classification), or a structured arbitrary object (structured prediction) [56, 57]. Conversely, unsupervised algorithms train without the explicit labeling of datapoints, allowing the algorithm to find hidden structures, patterns, or relationships in the dataset [58]. Anomaly detection and clustering are some examples of unsupervised algorithms, which are especially impactful in applications where data labels are not readily available such as in fraud detection, web mining, and social network analysis [58]. Semi-supervised algorithms are hybrids of the above-mentioned algorithms, requiring the labelling of some datapoints while the rest are left for the algorithm to distinguish with no supervision [59]. More specifically, either a supervised learning algorithm is augmented with unlabelled data resulting in what is known as semi-supervised classification, or an unsupervised algorithm is coupled with labelled data imposing a constraint on the clustering algorithm in what is known as constrained clustering [59]. Semi-supervised classification is especially useful in real-world applications where only part of the data is labelled such as in spam filtering, speech recognition, and video surveillance [59]. Reinforcement learning algorithms train using a numerical reward system for desired behaviours while punishing undesired ones; unlike other ML models, reinforcement learning algorithms are mostly suitable for interactive systems interfacing with humans (such as in games, personalized recommendations, and resource management) and/or the surrounding environment (such as in aircraft and robotic motion control) [60].

The discussion in this thesis is limited to supervised ML algorithms due to the desired scope of the work. Supervised ML classifiers include both simple linear algorithms and more complex non-linear ones [56]. Linear classifiers such as logistic regression, support vector machine (SVM), naïve bayes, and simple perceptron assume the decision boundaries of the classifications to be linearly separable [56]. Choosing the appropriate classifier highly depends on the structure and size of the dataset, including the number and type of features, as well as the strength and correctness of the assumptions made about the problem (inductive bias) [56]. It is seen that discriminative classifiers such as logistic regression – that maximize the conditional likelihood of observing a certain class given the dataset – perform better on large datasets [56]. Conversely, generative classifiers such as naïve bayes – that maximize the joint likelihood of observing a certain class with a given dataset – perform better on small datasets [56]. A main limitation of utilizing linear classifiers arises in applications where the dataspace is non-linear. As a result, deeper subsets of ML, such as artificial neural networks (ANNs), are used as they can correctly learn complex non-linear patterns within the given dataset. ANNs consist of a series of layers containing nodes or neurons. The layers are interconnected via the nodes that pass information through connections with different weights [56]. The algorithm adaptively learns the weights associated with connections between nodes in different layers to generate a better representation of the true model. A potential benefit of an ANN is the ability of classifying both large and small datasets by modifying the architecture of the network such as varying the depth of the hidden layers. In fact, it is empirically observed that increasing the depth of an ANN is an effective method to improve the performance of the classifier [56]. Moreover, deeper neural networks fall under the umbrella of deep learning, a more profound subset of ML characterized by its capacity to learn intricate, non-linear patterns. Deep learning has profound applications in image processing – such as in Convolutional Neural Networks (CNNs) – and language processing – such as in Recurrent Neural Networks (RNNs). Other recent breakthroughs in deep learning include the development of Generative Pre-trained Transformers (GPTs) that synthesize text by predicting subsequent word sequences, intelligently building upon both the immediate input and the broader contextual understanding acquired from previous data interactions.

1.2.2 Machine Learning: Basic Principles

ML algorithms fundamentally rely on probability theory, a mathematical framework that deals with uncertain statements [57]. This theory underpins the development of ML algorithms and branches into two primary methodologies: frequentist probability and Bayesian probability. Frequentist probability, also referred to as classical probability, conceptualizes probability as the long-term frequency of an event's occurrence. This interpretation uses probability to draw inferences about parameters or hypotheses by examining data from experiments or studies. Frequentist methods typically involve the computation of likelihoods and the application of test statistics to either reject or fail to reject hypotheses, without directly assigning probabilities to the hypotheses themselves. Within this framework, parameters are considered to be fixed yet unknown quantities, while data are viewed as random. In the context of ML, this approach often leads to the development of discriminative algorithms aimed at achieving a maximum likelihood estimation (MLE) of the conditional probability of observing the data given the model parameters. Conversely, Bayesian probability offers a fundamentally distinct perspective by treating probability as a subjective measure of belief regarding an event's occurrence, encompassing the uncertainty surrounding model parameters. This approach facilitates the integration of prior knowledge with new evidence to update the probability of an event or the parameters' values. In Bayesian statistics, probabilities are directly assigned to both hypotheses and parameters conversely to the common statistical approach of exploring a null hypothesis. In the context of ML, algorithms based on Bayesian principles, also known as generative algorithms, are designed to maximize the joint probability of observing the data with the model parameters. Bayesian methods excel in scenarios characterized by data scarcity or the significance of prior knowledge, providing a structured method for integrating such knowledge via the prior distribution. Furthermore, these methods offer a more detailed understanding of uncertainty, essential for decision-making and prediction in such situations. The scope of this current thesis is limited to algorithms based on the frequentist approach, and more specifically discriminative classifiers maximizing the conditional likelihood of observing the data given the model parameters.

1.2.2.1 Model Formulation & Cost Function

Generally, ML is described as the ability of a computer algorithm to learn by improving its performance (P) on a specific task (T) using certain experiences (E). The fundamental goals of ML are attained by specifying a model that represents certain beliefs about the task being solved, designing a cost function that measures how well those beliefs correspond with reality, and using a training algorithm to minimize that cost function. In fact, the learning of an algorithm can be described using the following formulation:

$$Learning = Representation + Evaluation + Optimization$$

$\boldsymbol{Representation}$: *This refers to the selection of a model based on hypotheses that define the learner's framework. It involves choosing the structure and parameters that the algorithm will use to process data and make predictions.*

$\boldsymbol{Evaluation}$: *This involves the use of objective and cost functions as metrics to assess the suitability of a model. These functions quantify the difference between the actual outcomes and the predictions made by the model, guiding the selection of the most effective model.*

$\boldsymbol{Optimization}$: *This is the process of refining the model parameters to minimize the cost function. Optimization algorithms iteratively adjust the model to find the configuration that produces the best predictions according to the evaluation criteria.*

Contrary to tasks that can be solved with fixed algorithms, such as expert systems with predefined rules, the tasks (T) addressed by ML algorithms are typically too complex for such straightforward approaches. These tasks encompass a variety of types, including but not limited to classification, regression, and the analysis of structured objects. ML algorithms are trained either in a supervised manner with labeled experiences (E) or unsupervised without labels. This discussion primarily focuses on supervised classification, although many of the principles discussed are applicable to other ML tasks as well.

In the context of a general multiclass classification problem, the objective is to predict one of C classes ($c \in [0, \dots, C]$) by providing the probabilities of each of these classes $\hat{y}_c{}^{(n)}$ given inputs with D features $x^{(n)} \in R^D$, where n represents an instance in a dataset of size N. A general formulation for a model would be:

$$\hat{y}_c^{(n)} = f\left(x^{(n)}; w\right) = \sum_{d=1}^{D} w_{d,c} \phi_d\left(x^{(n)}; \mu_d\right) \qquad \text{Equation (1)}$$

Here, $\hat{y}_c^{(n)}$ estimates the probability of the input belonging to class c, assigning the predicted class based on the highest probability outcome. This relationship can be succinctly represented as:

$$\hat{Y} = f(X; W, \mu) = \phi(X; \mu)W \qquad \text{Equation (2)}$$

where $\hat{Y} \in R^{N \times C}$ denotes the matrix of predicted probabilities for all classes across all instances, $X \in R^{N \times D}$ represents the matrix of input features, $W \in R^{D \times C}$ is the matrix of D weights for each of the c classes, ϕ is the basis function transforming the inputs into a feature space conducive to classification, and μ are bases that can either be fixed or adaptive as the case for neural networks (Section 1.2.2.3).

To convert the outputs into probabilities, a commonly employed activation function in multiclass classification tasks is the softmax function. It transforms logits (the raw model predictions) into probabilities by exponentiating each output and then normalizing these exponentials by their sum:

$$\hat{Y} = softmax(\phi(X;\mu)W) = \frac{e^{\phi(X;\mu)W}}{\sum_{c=1}^{C} e^{\phi(X;\mu)W}}, such\ that \sum_{c=1}^{C} \hat{y}_c = 1 \qquad \text{Equation (3)}$$

This ensures that each output probability lies within the [0,1] range and that the probabilities across all classes sum to 1 for any given instance, thus constituting valid probabilities. This formulation establishes our task (T): to predict the class labels y from the input features x by producing the probability distribution $\hat{y}$.

The performance (P) of a ML algorithm is commonly evaluated using a loss function or, in contexts where no regularization is applied, also known as the cost function $J(W)$. For multiclass classification tasks the cross-entropy cost is commonly used:

$$J(W) = -\sum_{c=1}^{C} Y \log\big(f(X;W,\mu)\big) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \qquad \text{Equation (4)}$$

The objective is to minimize this cost function with respect to the model parameters $W$and μ. The rationale behind employing cross-entropy cost originates from information theory, which employs entropy to quantify the uncertainty associated with a specific probability distribution P over K possible events. Entropy H(P) is defined as:

$$H(P) = -\sum_{k=1}^{K} p(k) \log\big(p(k)\big) \qquad \text{Equation (5)}$$

Cross-entropy extends this concept to measure the uncertainty associated with using a wrong probability distribution Q instead of a true distribution P:

$$H(P,Q) = -\sum_{k=1}^{K} p(k)\log\big(q(k)\big) \qquad \text{Equation (6)}$$

In the context of a multiclass classification within ML, cross-entropy acts as a cost metric to gauge the discrepancy between the model's predictions $\hat{y}$ and the true labels y:

$$J(y,\hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \qquad \text{Equation (7)}$$

An alternative perspective employs probability theory to justify the selection of the cross-entropy function. Specifically, in a multiclass classification context, maximizing the conditional likelihood of observing the data given the model parameters is mathematically represented by the Multinoulli Distribution:

$$max\big(L(y|\hat{y})\big) = max\left(\prod_{c=1}^{C} \widehat{y_c}^{\mathrm{y}_c}\right) \qquad \text{Equation (8)}$$

Taking the log of the above expression yields:

$$max\Big(\log\big(L(y|\hat{y})\big)\Big) = max\left(\log\left(\prod_{c=1}^{C} \widehat{y_c}^{\mathrm{y}_c}\right)\right) = max\left(\sum_{c=1}^{C} \mathrm{y}_c \log(\hat{y}_c)\right) \qquad \text{Equation (9)}$$

This elucidation demonstrates that maximizing the log-likelihood shown in Equation (9) is functionally equivalent to minimizing the cross-entropy (Equation (4) and Equation (7)), underscoring the function's theoretical underpinnings and its critical role in the optimization of ML models for multiclass classification tasks.

1.2.2.2 Optimization

Minimizing the cost function is a central optimization challenge in the development of ML algorithms. This process involves adjusting model parameters, such as weights W and the bases μ, to identify the most effective model configuration that minimizes error. Optimization, as a discipline, is intricate and demands a comprehensive exploration. For the purposes of this discussion, the focus is limited to concepts directly relevant to the scope of this thesis, directing readers seeking an in-depth treatment to the work of Boyd and Vandenberghe [61] for additional insight.

Typically, addressing optimization problems entails approximating objective functions with linear or higher-order polynomials to ensure the continuity of second and third-order derivatives. This approximation is facilitated by employing Taylor series expansions around the optimal or critical point $\boldsymbol{x}^* \in R^D$, such that for any positive scalar δ where $0 < ||\Delta\boldsymbol{x}|| \leq \delta$, we have $f(\boldsymbol{x}^* + \Delta\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$**.** The Taylor series expansion about $\boldsymbol{x}^* + \Delta\boldsymbol{x}$ is given by:

$$f(\boldsymbol{x}^* + \Delta\boldsymbol{x}) = f(\boldsymbol{x}^*) + \Delta\boldsymbol{x}^T\nabla f(\boldsymbol{x}^*) + \frac{1}{2}\Delta\boldsymbol{x}^{\boldsymbol{T}}\nabla^2 f(\boldsymbol{x}^*)\Delta\boldsymbol{x} \qquad \text{Equation (10)}$$

Achieving a weak minimum at the point $\boldsymbol{x}^*$ **r**equires, at the very least, $f(\boldsymbol{x}^* + \Delta\boldsymbol{x}) = f(\boldsymbol{x}^*)$, which implies for a first-order expansion that $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$**,** defining the first-order necessary conditions for an optimal point. For a strong minimum, $f(\boldsymbol{x}^* + \Delta\boldsymbol{x}) > f(\boldsymbol{x}^*)$, requiring for a second order expansion that $\nabla^2 f(\boldsymbol{x}^*) > \boldsymbol{0}$ since $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$, which outlines the second-order necessary conditions. Therefore, the optimization problem is reduced to simply finding the roots of the gradient of the objective function, employing methods such as the Bisection Method or

Newton’s Method. Within the realm of ML, gradient descent emerges as the predominant method, a line search method which defines the search algorithm to follow a simple formula:

$$\boldsymbol{x_{k+1}} = \boldsymbol{x_k} - \alpha_k \nabla f(\boldsymbol{x_k}) \qquad \text{Equation (11)}$$

Here, α_k represents the step length, a crucial hyperparameter that necessitates careful calibration. Its magnitude is pivotal, as excessively large values can lead to algorithmic overshooting, whereas overly small values can significantly delay convergence. Furthermore, in ML, Stochastic Gradient Descent (SGD) is a widely adopted method for tackling high-dimensional problems to address the computational costs of the algorithm. Unlike batch gradient descent, which calculates the gradient of the objective function across the entire dataset, SGD estimates the gradient using just a sample of instances (or a small subset termed a mini-batch) at each step. This strategy significantly lessens the computational demands for each iteration, enhancing the algorithm's speed and scalability, especially with large datasets. However, this benefit comes at the cost of increased oscillation around the optimal point, which can impede efficient convergence. To mitigate these oscillations, the integration of momentum with SGD is a common practice. The concept of momentum helps to smooth out the variations and accelerates convergence by incorporating a fraction of the previous update. The formulation of SGD with momentum is given by:

$$\boldsymbol{x_{k+1}} = \boldsymbol{x_k} - \alpha_k \Delta\boldsymbol{x}_{\mathbf{k+1}};\ \Delta\boldsymbol{x}_{\mathbf{k+1}} = \beta\Delta\boldsymbol{x_k} + (1-\beta)\nabla \mathrm{f}(\boldsymbol{x}_{\mathbf{k}}) \qquad \text{Equation (12)}$$

Here, β represents the momentum coefficient, controlling the extent to which previous gradients influence the current direction. A momentum coefficient of 0 simplifies Equation (12) back to the basic gradient descent formula (Equation (11)). Within the realm of ML optimization,

β is treated as an additional hyperparameter that requires careful tuning to balance the trade-off between convergence speed and stability.

1.2.2.3 Neural Network Framework

Consider again the general formulation of the ML algorithm described in Equation (1), it can be noted that the basis function might depend on both the input features X and basis parameters μ, determining whether the algorithm employs a fixed or adaptive basis. With a fixed basis, the count of basis functions becomes a hyperparameter optimized during training, with the potential risk of overfitting. Conversely, an adaptive basis leads to a formulation that resembles a general neural network, as described below:

$$\hat{y}_k = g\left(\sum_m W_{k,m}\, h\left(\sum_d V_{m,d} x_d\right)\right) \qquad \text{Equation (13)}$$

Here, g and h represent activation functions that introduce nonlinearity into the model, enabling it to capture complex nonlinear relationships in the data. Equation (13) describe a one-layered neural network depicted in Figure 1-1 below, and can be more succinctly expressed as:

$$\hat{Y} = g\big(Wh(XV)\big) \qquad \text{Equation (14)}$$

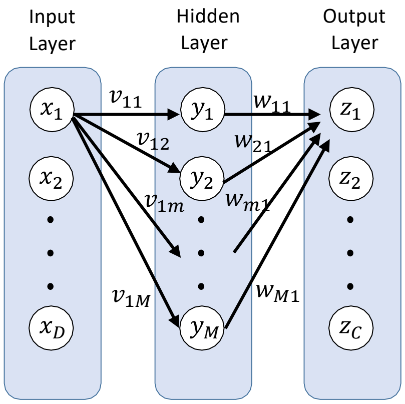


Figure 1-1 One-layered multi class classification neural network with D input features, M hidden units, and C outputs.

The choice of the activation function in the outer layer depends on the task. For multiclass classification, the outer layer activation function g would again be the softmax function. As demonstrated, the cost function can be generated from the below expression by either minimizing the cross-entropy cost or maximizing the Multinoulli likelihood:

$$\hat{Y} = g\big(Wh(XV)\big) = g(Wz) = softmax(Wz) \qquad \text{Equation (15)}$$

The choice of the inner layer activation function depends on the desired complexity of the model and the desired properties of the activation function. Nonlinearity is pivotal for learning complex data patterns, but the choice of activation function within the inner layers also significantly impacts computational efficiency and the accuracy of gradient computations during optimization. The Rectified Linear Unit (ReLU) function is widely favored for inner layers due to its computational efficiency and its ability to mitigate vanishing or exploding gradients in deep networks. ReLU is defined as:

$$h(x) = \max(0, x) \qquad \text{Equation (16)}$$

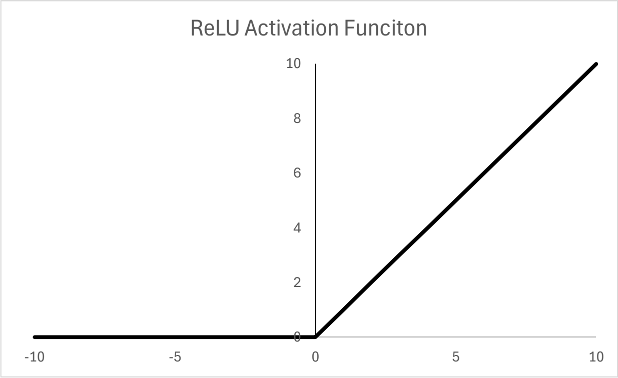


Figure 1-2 ReLU activation function is computationally efficient and solves the vanishing or exploding gradient problem. The derivative goes to zero if the function is inactive.

Beyond adjusting hyperparameters linked to the optimization algorithm, such as step length α and momentum coefficient β for SGD with momentum, neural networks also involve architectural hyperparameters. These include the number of hidden layers (depth) and the number of units within each layer (width). In scenarios involving high-dimensional inputs, hidden units can serve to reduce dimensionality, a concept we will explore further in the subsequent section. Generally, empirical observations suggest that deeper networks offer more benefits than merely wider networks with a large number of units. Nevertheless, despite careful tuning of hyperparameters related to the model's architecture and optimization, challenges like overfitting and limited generalizability can still arise, necessitating additional measures for mitigation.

1.2.3 Machine Learning: Overcoming Limitations

The primary goal of any ML algorithm is effective generalization: the ability to perform accurately on unseen data. Generalization in ML is influenced by several factors. These include the availability of a representative training dataset that accurately reflects the broader population related to the predictive task, and the sufficiency of the dataset size to adequately train the algorithm. Additionally, it's crucial to prevent overfitting, where the algorithm overly adapts to the training data, learning its noise and peculiarities instead of the underlying pattern. Furthermore, it is also imperative to maintain a balance between the sizes of the feature space and the training dataset. To mitigate these limitations, careful steps are necessary, particularly in the data collection stage. Ensuring that the dataset is both sufficiently large and representative of the population is vital. In cases of limited data availability, a combination of various techniques is applied at different stages of the ML algorithm’s training and development. For data preparation and preprocessing, this includes addressing the features through dimensionality reduction and feature selection, as well as enhancing the overall dataset size via data augmentation. During the training and model development phase, techniques such as regularization and more precisely early stopping are integral to prevent overfitting. Finally, more complex approaches like transfer learning are employed to leverage pre-existing models for superior performance on new tasks.

1.2.3.1 Dimensionality Reduction & Feature Selection

Dimensionality reduction, the process of mapping high-dimensional data into a lower-dimensional space, is particularly useful when the feature space is substantially larger than the training set [62]. A widely used technique is Principal Component Analysis (PCA), which applies linear transformations to minimize the difference between original and reconstructed data vectors [62]. However, PCA can lead to reduced feature interpretability, which is a significant drawback when it is important to understand the influence of specific features. Feature selection, in contrast, involves selecting the most impactful features for use in ML algorithm development. This can be achieved through filter methods, which score each feature based on statistical techniques relating them to the target variable, selecting the top performers. Alternatively, wrapper methods employ a more iterative process, either adding (forward selection) or removing (backward elimination) features to find the optimal subset; each generated sub-model is evaluated, and the one with the best performance is chosen.

1.2.3.2 Data Augmentation

Data Augmentation is another effective approach to addressing dataset limitations, which involves synthetically generating additional datapoints. Techniques like jittering (adding noise) or scaling to the original dataset effectively increase its size [57]. These methods enhance the robustness of models, particularly in scenarios with small datasets, by preventing overfitting and improving generalizability. Data jittering introduces minor variations to the existing data, randomly sampling from the original dataset and adding slight random noise. Data scaling, meanwhile, employs a similar technique but with a constant fixed scale. Depending on the context, data jittering might be more beneficial. For instance, jittering is particularly beneficial for neural networks, which often exhibit sensitivity to noise. Introducing controlled noise through data augmentation can thus enhance their noise robustness [57].

1.2.3.3 Regularization & Early Stopping

Regularization techniques may be defined as strategies used in ML model development and training that are explicitly designed to reduce the generalization error, possibly at the expense of increased training error [57]. The primary objective of employing regularization is to balance the model's ability to learn from the training data while averting the risk of overfitting, thereby ensuring that the model remains robust and performs well on unseen data. Among the various regularization approaches, two prevalent methods stand out: incorporating constraints within the model's cost function and implementing early stopping during the training phase.

Regularization through model constraint involves adjusting the model's complexity – sometimes simplifying it – to enhance its generalization capabilities. This form of regularization amends the cost function to include an additional penalty term, which acts to constrain the model during the optimization process, thereby fostering better generalization. Using this method, Equation (4) is modified to include a regularization term:

$$J(W) = -\sum_{c=1}^{C} Y \log\big(f(X; W, \mu)\big) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) + \lambda f(W) \qquad \text{Equation (17)}$$

In this equation, λ signifies the regularization parameter that controls the strength of the regularization effect, whereas $f(W)$ specifies the regularization type. The most common forms include weight decay regularization, which computes the L2 norm of the weights W, and the Least Absolute Shrinkage and Selection Operator (Lasso) regularization, focusing on the L1 norm of W.

Early stopping, on the other hand, is based on the observation that while training error may decrease steadily, validation set error often starts to increase after reaching a minimum. This phenomenon indicates that a model configuration with a lower validation error – and potentially a lower test error – can be identified. Early stopping returns the model configuration at the point where the validation set error is minimized. It achieves this by storing the model parameters each time there's an improvement in validation set performance. When the training terminates, the model returns these optimally stored parameters instead of the most recent ones. Training stops when there is no enhancement in the recorded best validation error for a designated number of iterations, providing a practical and effective means to prevent overfitting while securing a model that is well-tuned to generalize to new data.

1.2.3.4 Transfer Learning

Another powerful strategy is transfer learning, where knowledge from a previously trained model on a related task is utilized. Two main methods are highlighted in the literature: fine-tuning a pre-trained model and using it as a feature generator [57, 63]. Fine-tuning involves continuing the model's training phase on a new dataset. In deep learning applications, this often starts with shallow tuning of the outermost layers, progressively moving to deeper layers. This technique leverages the concept that an ANN’s early layers contain generic low-level features, whereas later layers hold task-specific high-level features [57, 63]. However, when applied to ANNs with few layers on small datasets, there's a significant risk of overfitting. The feature extractor method, in contrast, involves freezing the trained layers of the model and appending new layers to its output. This method transforms input data into high-level features for better classification, particularly useful in small datasets. While this approach reduces the likelihood of overfitting and enhances generalizability, it shares a drawback with PCA in that the interpretability of features may be compromised.

1.2.4 Machine Learning: Model Interpretability

In general, the challenge of interpretability in ML often leads to models being labeled as “black boxes”, particularly because delineating the significance of input features in complex models can be a daunting task, even for field experts [64]. In a multiclass classification task, a highly regarded approach for assessing feature importance is the permutation feature importance algorithm. This method evaluates a feature’s significance by observing the variation in the model’s performance metrics – such as the loss function and prediction accuracy – after randomly shuffling the values of that feature across the dataset [65, 66]. A feature is deemed crucial if its permutation results in a noticeable degradation in model performance, indicating that the model heavily relies on that feature for making accurate predictions. When applied to both training and testing datasets, the permutation feature importance algorithm provides dual insights: it identifies the features critical for the model’s learning phase from the training set and those vital for the model’s generalization capabilities from the testing set. This dual application facilitates a comprehensive understanding of feature relevance, highlighting the model's dependency on specific inputs for both learning and prediction tasks.

ANNs offer a unique advantage in model interpretability through the analysis of weights in hidden layers, which can shed light on the decision-making processes of the classifier. The Connection Weights Algorithm (CWA), pioneered by Olden and Jackson [67], stands out as a method for quantifying the influence of each input feature on each class [64]. By calculating the Connection Weights Products (CWPs) – the sum of the products of all the connection weights that relate an input to an output – the algorithm discerns the relative importance of each feature to every class. Historical applications of this method have been predominantly on simple neural networks with one hidden layer [54, 64, 67, 68]. However, recent advancements, including a study presented in Chapter 4 conducted by our group and led by the author of this thesis, have extended its application to multilayered neural networks, validating its effectiveness alongside the permutation feature importance method [69]. The CWA not only quantifies the impact of input features on outputs in terms of magnitude but also direction, where positive or negative CWPs indicate whether a feature value above or below the average is associated with a specific class. This insight is particularly valuable in educational and training contexts, such as surgical training, where understanding the rationale behind a model's classifications can enhance the evaluation of trainees' skills and, ultimately, contribute to improved performance in complex tasks.

1.2.5 Machine Learning in Surgical Simulators

The use of ML algorithms in surgical simulations, through leveraging the extensive datasets derived from surgical simulators, has marked a significant advancement in the classification of surgical expertise [70]. These algorithms surpass previous methods in both granularity and precision, offering nuanced insights into the components of surgical performance that delineate various levels of expertise [48, 53, 70]. The integration of these algorithms with VR surgical simulators not only enhances the specificity of performance classification but also deepens our understanding of how diverse performance metrics influence overall skill assessment [56]. Utilizing metrics generated by surgical simulations, ML models can discern patterns linking specific performance indicators to levels of surgical proficiency, such as distinguishing between trainees and experts. Surgical performance metrics are usually categorized based on the aspect of the surgical performance that is being measured. For instance, kinematic metrics related to the motion of the surgical tool, including speed, accelerations, and rotations are considered part of the motion metrics. Forces, torques, and contacts with critical anatomical structures (such as nerve contacts) are categorized as safety metrics. Tool path lengths, anatomical structures volume removals, and time to completion of tasks are considered as efficiency metrics. Previous research has applied these metrics as input for various ML architectures, including K-nearest neighbors, Naïve Bayes, SVM, and conventional neural networks, achieving predictive accuracies ranging from 65% to 97.6% [48, 69-74]. While these studies primarily focus on skill classification and feedback provision, there is often a gap in exploring the underlying reasons for these classifications or in quantifying the impact of different performance metrics [74].

However, recent investigations have begun to bridge this gap by employing one-layer and, more innovatively, two-layer ANNs, complemented by the Connection Weights Algorithm (CWA) [54, 64, 67-69]. This combination elucidates the relative importance of specific features in classifying surgical performance, enabling surgical educators to tailor training programs more effectively. By pinpointing areas requiring improvement, such personalized training can optimize the development of surgical skills, embodying the concept of “Technical Abilities Customized Training” (TACT) [75]. This approach not only aims to enhance the proficiency of all trainees but also facilitates early identification and support for those who may struggle with surgical tasks [69].

The progression towards AI-based intelligent tutor systems stands as a compelling augmentation to the established proficiency of ML algorithms in the accurate classification and analysis of surgical performance. Our research group has showcased the potential of such systems in enhancing resident training through the provision of real-time performance feedback [49, 76]. These innovative systems aim to emulate the mentorship provided by expert surgeons by delivering immediate, task-specific evaluations and highlighting potential risks. The development of these systems can be approached via two distinct methodologies, as evidenced by the work of Mirchi, et al. [49] and Yilmaz, et al. [76]: one method involves utilizing an offline pre-trained ML model for assessment and feedback purposes, while the alternative strategy leverages an adaptive algorithm that evolves with continuous input from new data, thus offering ongoing feedback to trainees. A notable challenge that arises with the implementation of these systems is the risk of "negative training", where trainees might inadvertently be steered towards incorrect skill levels [77]. To mitigate this concern, it's imperative to conduct thorough validations of the surgical skill levels as benchmarked by the ML algorithms. One approach to achieving this involves the application of realistic physics-based forces, mirroring the physiological forces encountered during actual surgical procedures. Addressing this issue necessitates an advanced understanding of the surgical technique analyzed with a deeper exploration into the biomechanics of physiological tissues and their behavior during the surgical interactions. Subsequently, it calls for an accurate representation of these physiological tissue reactions within surgical simulations, with a particular emphasis on physics-based simulations. This focus enhances the realism and efficacy of surgical training. It ensures that the skill levels benchmarked by ML algorithms accurately reflect the complex dynamics of surgical procedures, thereby minimizing the likelihood of negative training and fostering the development of truly proficient surgical practitioners.

1.3 Physics-Based Surgical Simulators in Minimally Invasive Surgeries

1.3.1 Minimally Invasive OLLIF Surgery

Spinal fusion surgery primarily addresses spine instability, deformity, or pain. It is often performed in the lower region of the spine, known as the lumbar region, in which case the procedure is termed lumbar interbody fusion (LIF) [78]. During this procedure, a surgeon removes the intervertebral disc (IVD) and fuses two or more vertebrae – the bones forming the spinal column – permanently. This fusion eliminates the relative motion between the connected vertebrae, treating the underlying ailment. Historically, LIF surgery was invasive, with surgeons accessing the spine from the patient's back using a posterior approach, or from the front via an anterior approach [1]. These methods required stripping muscles and soft tissue to reach the spine, often leading to surgical complications [1]. In contrast, recent advancements have seen the adoption of minimally invasive (MI) spinal fusion techniques. Surgeons may now use a lateral approach, specifically the Oblique Lateral Lumbar Interbody Fusion (OLLIF), to access the operative area from the side of the patient. This technique involves making small, approximately 15 mm incisions and entering the spine through the Kambin's triangle [1]. Kambin's triangle, as depicted in Figure 1-3, is a triangular region in the lumbar spine, formed and bounded by an exiting spinal nerve as the hypotenuse, the superior endplate of the inferior vertebral body as the base, and either a traversing nerve root or the superior articular process (SAP) of the inferior vertebra as the height [1]. This area also encompasses a portion of the facet joint that connects the SAP of the inferior vertebra to the inferior articular process (IAP) of the superior vertebra. Kambin's triangle serves as an electrophysiological silent window, offering surgeons a safe zone for surgery with a reduced risk of nerve damage.

The general steps of a MI OLLIF procedure begin with the surgeon establishing the location of Kambin's triangle, followed by the stabilization of the surgical area. This is achieved by inserting a surgical port through the incision, thereby completing the gaining access phase of the procedure [1]. Once access to the IVD is secured, the surgeon undertakes a facetectomy, which involves removing a portion of the facet joint – a set of synovial joints between two adjacent vertebrae that facilitate spinal movement. Following the facetectomy, the procedure continues with a discectomy, entailing the removal of the nucleus pulposus and annulus fibrosus, the core components of the IVD. The final steps involve the insertion of a cage and the placement of bone graft material to simulate the bone healing process and promote bone formation [2].

Quillo-Olvera, et al. [79] and Morgenstern, et al. [80] provide a detailed description of the OLLIF surgical steps. Gaining access to the target IVD necessitates puncturing through muscle and tissue layers with minimal manipulation of instruments and limited exploratory movements to reduce tissue and nerve disruptions. The operation within Kambin's triangle eliminates the need for direct visualization, relying instead on tactile feedback for precise navigation to the surgical site. Following gaining access to the IVD, the goal is to make adequate space for discectomy and the safe placement of the graft and cage. This requires adequate facetectomy and endplate preparation. Facetectomy typically involves removing portions of the SAP and IAP, along with their connecting joint, to decompress and protect the exiting and traversing nerves while also exposing the IVD for further steps. Subsequent endplate preparation involves removing cartilage and other inhibitory soft tissue to reveal the vascular bone, crucial for bone growth, without compromising the vertebral structural integrity. Excessive endplate preparation risks breaching the endplate, potentially leading to complications like subsidence, where the fusion device may sink into the softer bone beneath the endplate.

Following the description of the procedural steps for an MI OLLIF, Quillo-Olvera, et al. [79], drawing from these insights, documents specific recommendations for surgeons. These guidelines are derived directly from documented surgical cases, emphasizing the critical aspects of the surgery to minimize complications and optimize outcomes. The first recommendation urges surgeons to minimize manipulation and interaction with neural elements, aiming to reduce postoperative sensory disturbances such as paresthesia, dysesthesia, or direct injury to traversing or exiting nerves. Secondly, it is essential to create sufficient space around the IVD for the safe placement of the interbody cage, ensuring that this is done without compromising the traversing or exiting neural elements. The third guideline focuses on the careful preparation of the endplates, advising against over-preparation, which risks damaging the endplates and increasing the likelihood of implant subsidence or nonunion. Conversely, insufficient preparation may complicate the placement of the interbody implant or impede the fusion process. These strategic recommendations underscore the precision and care required for surgical excellence and could serve as benchmarks for assessing surgical performance during training.

In all of these steps, the surgeon mostly relies on the somatosensory feel of the task especially during the gaining access phase; surgeons have a limited view of the operating region in the following phases, which can only be examined through a 2D planar view of the laparoscopic camera [81]. Hence, developing a system that mimics both the visual and somatosensory reality of the procedure is essential for proper training. A deep understanding of the biomechanics of the spine is thus required to have a better appreciation of the impact of accurate physics-based force feedback on surgical training.

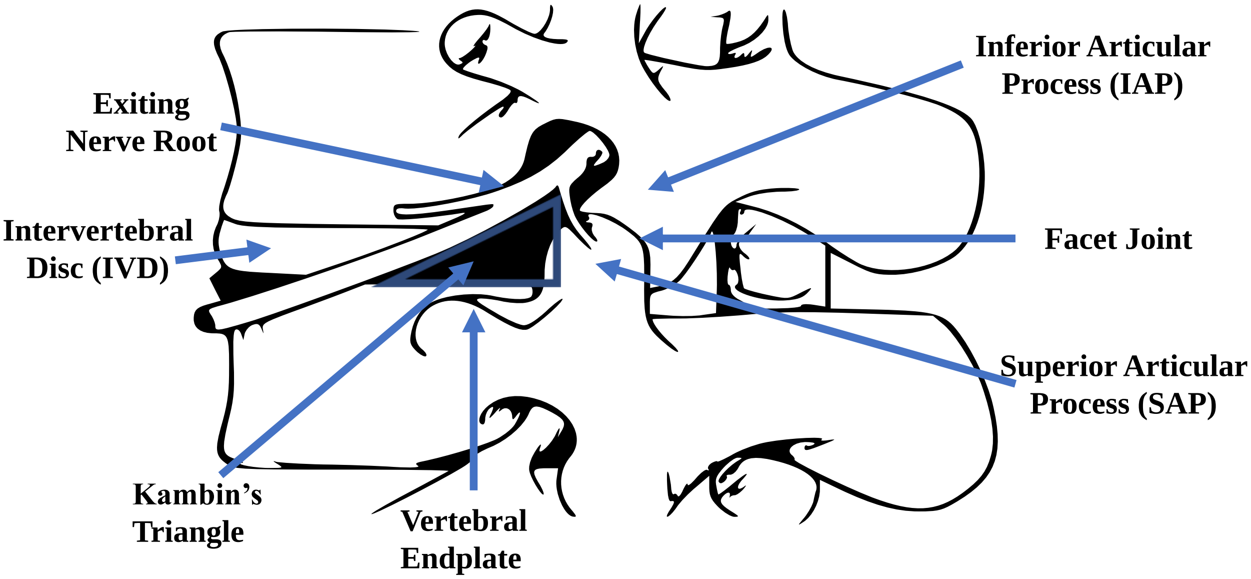


Figure 1-3 The Kambin's triangle is formed by the exiting nerve root, the superior articular process, and the superior endplate of the inferior vertebral body. It offers an electrophysiological silent window that provides surgeons a safe surgical corridor with a reduced risk of nerve damage (adapted from Abbasi and Abbasi [1]).

1.3.2 Spine Biomechanics and Modeling

The human spine can be deconstructed into three main systems based on their respective roles in supporting the body's numerous movements while protecting the spinal cord [82]. The systems can be categorized into passive, active, and neural systems [82]. Passive components facilitate the spinal functions by providing the appropriate structural support. These include the vertebrae, IVDs, and connective tissues including ligaments and fascia. Active components use direct contractile forces to promote healthy physiological movements of the spine. Active elements include spinal muscles and tendons that connect the spine to different parts of the musculoskeletal system. The spinal neural system controls the passive and active systems by relaying neural signals to maintain the required movements and functions of the spine.

In general, the macroscopic material attributes of the human physiological organs, including components of the spine, are manifestations of the microstructural design and the micro-mechanisms that the constituents undergo at the microscopic level [83]. As such, recognizing the material properties of the microconstituents of the different components of the human spine helps in understanding the overall bulk material properties [84]. This methodology can be applied when studying any of the spinal components listed above.

Macroscopically, the vertebra is a strong, stiff, tough, and lightweight mechanical support composed of the strong and stiff cortical bone and the energy absorbent cancellous bone [83]. In general, bone is an anisotropic material, meaning that its mechanical properties vary depending on the direction of the load applied. Microscopically, both the cortical and cancellous bones are composed of differing concentrations of hydroxyapatite (HA) crystal minerals, organic collagenous constituents, and water. The HA crystal minerals give bone its high stiffness, the organic collagenous constituents provide its ductility and elasticity, while the fluid filled pores contribute to its viscoelasticity; it is found that these constituents are organized and oriented in certain directions, giving rise to the anisotropic behavior seen at a macroscopic level [83]. Further investigation reveals that cortical bone has a high concentration of mineral constituents with low porosity, which is manifested macroscopically in its high stiffness. While cancellous bone has a high organic phase with high porosity, which is manifested in its viscous features – its ability to store and dissipate energy [85]. From a mechanical modeling perspective, both cortical and cancellous bones behave in a viscoelastic manner – materials that have time dependent responses [86]. Numerous models have been developed to study the mechanics of the different hierarchical structure of bone. These models include: analytical models based on the strength of materials theories, analytical models based on the micromechanics theories, and computational models based on the finite element method (FEM) [83]. Certain modelling techniques may be more appropriate than others, depending on the scale of the modeled structure and the required accuracy of the model. Also, some employ multiscale analyses to try and better encapsulate this complex behavior. The strength of materials techniques are mostly variants of the Voight and Reuss bounds, which assumes the different hierarchical levels of bone to mimic fiber-reinforced materials [86]. The models based on micromechanics theories can be seen as a more complex extension of the ones based on the materials theories, allowing for a multi-phase analysis that describe all the constituents at the nanoscale, namely collagen, HA crystals, and water [86]. The analytical models based on both these theories provide only approximate solutions for simplified geometries, not allowing the user to specify anisotropic or nonlinear behavior [86]. Hence, they are very limited in terms of accurately modeling the geometry of bone, as well as capturing the complex features of nonlinearity and anisotropy exhibited by it [86]. The use of FEMs in modeling bone has several advantages over the analytical methods; FE models allow complex geometries to be depicted using medical image data such as micro-CT. Also, FE software allows the specifications of both isotropic and anisotropic material properties, as well as, linear and non-linear properties [86]. Therefore, most studies model bone using the FEM [87]. Nevertheless, due to the computational complexity in modeling viscoelastic behavior at a macroscopic scale, most FEM studies assume bone as being linearly elastic at small strains [87].

The IVD is made up of three subcomponents: an inner soft, deformable tissue known as the nucleus pulposus (NP) (composed of 4% Collagen, 14% Proteoglycans, and 77% water), surrounded by fibrous concentric layers of the anulus fibrosus (AF) (composed of 15% Collagen, 5% Proteoglycans, and 70% water), and bounded superiorly and inferiorly by the thin layers of the cartilaginous end plates (CEPs) [88]. The IVD exhibits a non-isotopic non-homogeneous viscoelastic behavior at the macroscopic level when subjected to external loading [88]. The NP has a high concentration of water allowing it to have fluid-like behavior. The AF has differing concentrations of constituents at different locations. Moving from the outer to the inner portion of the AF, the concentrations of water, Proteoglycans, and type II Collagen increases while type I Collagen decreases, allowing it to withstand higher compressive forces internally and higher tensile forces externally [88]. Modelling the IVD using the FEM typically differentiates the NP from the AF. Available models usually range from assuming the NP as an ideal incompressible fluid interacting with a linearly elastic AF, to biphasic models simulating the fluid-structure interactions between an incompressible porous solid saturated with an incompressible fluid [89]. As with the limitations of modeling bone, complex non-linear viscoelastic behavior results in high computational cost, requiring the adaptation of simplified models for real time interactions [89].

Spinal muscles attach to the skeleton through connective tissues –known as tendons and fascia – and are designed to provide a pulling force at extremities. The complexity of modelling skeletal muscles lies in the multitude and redundancy of control points between the muscles and skeletal elements, not only in focus on the spine but for all controlled joints. A complexity that is further exacerbated by the feedback interaction between muscles and the central nervous system (CNS). Therefore, it is extremely beneficial to reduce the model to its most important features. As a result, researchers have relied on simplified skeletal muscle models to better understand the role of spinal muscles. Lumped phenomenological models that represent the mechanical behaviour of muscles have been extensively studied and developed due to their simplicity [90].

Phenomenological models, such as the Hills Muscle Model, are composed of distinct elements that give rise to the contractile and elastic behaviours of muscles [90].

In general, incorporating realistic material properties to generate physically-accurate visuals and force profiles of spinal tissues remains integral to achieving high simulator fidelity [91]. However, biological tissues require high computational cost for very accurate modeling [89, 92, 93]. As such, most biomechanical modelling used in surgical simulators rely on simplified assumptions of linearity and homogeneity to reduce computational complexity [93]. In most cases simplifying the model yields an optimal compromise between fidelity and efficiency. Yet, the challenge intensifies in VR simulations, which require real-time representation of physiologically accurate models [93]. For immersive VR experiences, the visual feedback update rate must be at least 30 Hz to ensure the continuous perception of motion, while haptic feedback necessitates a minimum of 1000 Hz to provide stable and smooth tactile sensations [94]. Addressing these demands, recent efforts, including those by our research group, have focused on optimizing realistic physiological modeling with real-time physics-based haptic fidelity. These efforts use haptic rendering on a separate thread from the visual model and apply empirical formulas based on cadaver studies for accurate haptic feedback. This method is computationally less demanding than integrating tissue models into a FEM for haptic response [91, 94, 95].

1.3.3 Importance of Physics-Based Force Feedback in Minimally Invasive Surgical Training

The employment of haptic-feedback in surgical simulators has substantially impacted surgical practice learning curves as demonstrated by numerous studies [88, 89, 92, 93]. Haptic feedback is generally described as providing the user with both kinesthetic (forces and torques sensed by muscles, tendons, and joints) and tactile (vibrations sensed by mechanoreceptors on the skin) feedback resulting from the interactions between the virtual tool and components in the virtual scene [92]. The importance of force feedback, achieved by haptics, is further highlighted in MIS, such as the OLLIF, where surgeons extensively rely on somatosensory force differentials between the various soft and hard tissue to make decisions [92]. Nevertheless, not all simulators employing haptics provide realistic force outputs [93]. In fact, some of state-of-the-art simulators leverage advanced voxel-based gaming engines, incorporating haptic and auditory feedback based on geometric models to enhance the experiential realism of the simulation [93]. Simulators utilizing discrete or heuristic methods rather than constitutive modeling based on continuum mechanics, introduce a risk of providing forces that may not accurately reflect those encountered in actual surgical procedures. This discrepancy could lead participants to apply forces in the simulator that are unrealistic, potentially compromising the training's effectiveness and fidelity [93]. Recent studies have highlighted the importance of using accurate physics-based haptics rather than geometric models to ensure the accuracy and reliability of the generated force feedback [96]. Integrating a tissue model that offers both visual and haptic feedback presents a notable challenge. The quest for realism and real-time response in surgical simulations introduces difficulties for each component, often leading to a situation where enhancing one aspect comes at the expense of the other [93]. To that end, recent studies have identified utilizing data from cadaveric experiments to implement realistic physics-based feedback as a more computationally efficient approach to simulate accurate haptic and visual interactions [32, 91, 97].

Even with the current evidence supporting the integration of physics-based haptic feedback in MI surgical training, there are no studies that objectively measure the influence of physically accurate force profiles on simulation training outcomes [96]. Currently, there are no objective measures for haptic fidelity that have been established in the literature as most rely on subjective assessments. Subjective assessments, while effective when establishing the foundational validation of simulators, are not sufficient for evaluating the impact of haptic accuracy on surgical outcomes. The use of simulators that lack realistic haptic rendering in surgical training can lead to negative transfer effects in the operating room. This scenario poses a risk where learners might apply inappropriate forces, learning habits that could be challenging to correct later [96]. Naturally, this error is further exacerbated in MIS whereby the applied forces are crucial for guidance [92]. Therefore, it is imperative to establish an objective method to measure the impact of haptics on surgical training. The rise of ML in surgical simulation training allowed for a new opportunity to use a data-driven approach. This approach can objectively measure the influence of haptic and force feedback accuracy on surgical training and performance.

Chapter 2. Research Objectives and Hypotheses

The global objective of the thesis is to establish the validity of a physics-based VR/AR spinal surgical simulator in training and assessing surgical trainees. The main objective was attained by the sequential validation of different aspects of the simulator. Starting with subjectively assessing the resemblance of the simulated scenario to reality, a side-by-side comparison with a cadaveric surgery is conducted to further support face and content validity of the developed simulator. Afterwards, data related to the psychomotor interactions of participants generated by the simulator were transformed into novel surgical metrics that are utilized to assess construct validity. The surgical metrics were subsequently used in a machine learning algorithm to give further insight into aspects of surgical performance that defines expertise. Once sufficient validity was established, the focus shifted to assessing the impact and importance of using physics-based haptic feedback on surgical training. Apart from subjective methods that use expert opinions in capturing the impact of physics-based simulations, a novel method used by this thesis was to make use of the trained machine learning algorithms to obtain an objective measure. By varying the force-feedback generated by the haptic device, new surgical participants were recruited to perform the virtual procedure and then subsequently classify their new performance metrics using the previously developed machine learning algorithms. The change in the accuracy of the machine learning model was a measure of the impact of both the use of physics-based force feedback and the machine learning model used. Therefore, with regards to achieving the global objective, the following tasks and sub-objectives were constructed to stream towards establishing the validity of the developed simulator.

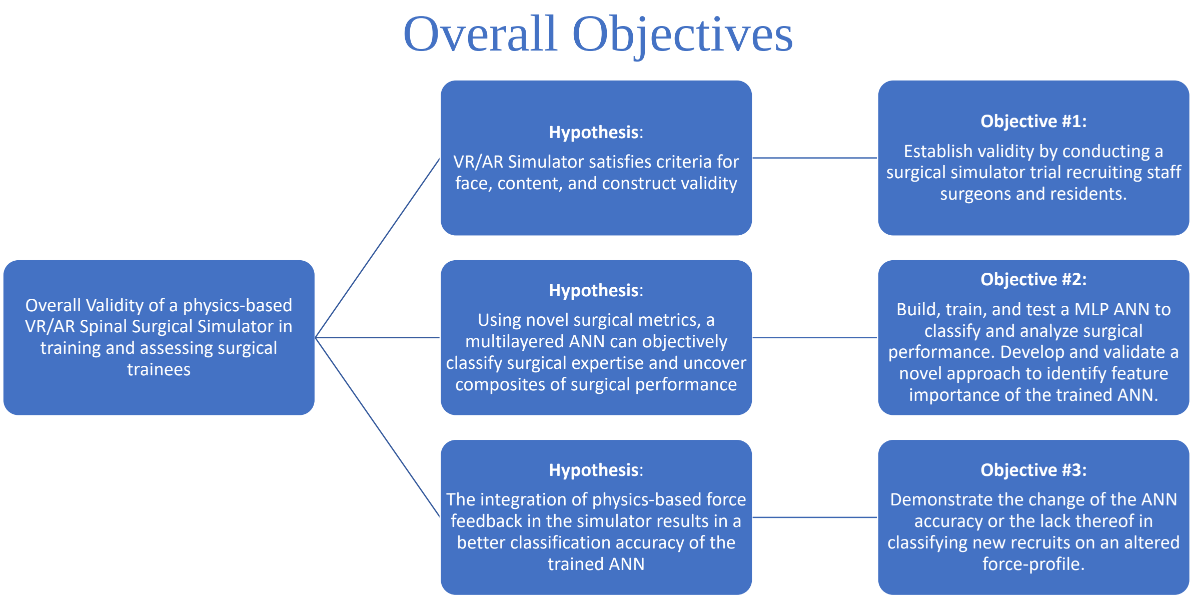

Overall Objectives
Overall Validity of a physics-based VR/AR Spinal Surgical Simulator in training and assessing surgical trainees
Hypothesis:
VR/AR Simulator satisfies criteria for face, content, and construct validity
Hypothesis:
Using novel surgical metrics, a multilayered ANN can objectively classify surgical expertise and uncover composites of surgical performance
Hypothesis:
The integration of physics-based force feedback in the simulator results in a better classification accuracy of the trained ANN
Objective #1:
Establish validity by conducting a surgical simulator trial recruiting staff surgeons and residents.
Objective #2:
Build, train, and test a MLP ANN to classify and analyze surgical performance. Develop and validate a novel approach to identify feature importance of the trained ANN.
Objective #3:
Demonstrate the change of the ANN accuracy or the lack thereof in classifying new recruits on an altered force-profile.

Figure 2-1 Overall hypotheses and objectives of the current dissertation.

Task 1:

Hypothesis 1: The novel mixed reality spinal surgical simulator satisfies the criteria for face, content, and construct validity. The simulator satisfies face and content validity criteria by having a median score greater than 3 on a 5-point Likert Scale as assessed by surgical experts with the use of a questionnaire. The simulator satisfies the statistical criteria (P<0.05) for construct validity defined by the KruskalWallis, WelchAnova, and Anova test statistics.

Objective 1: Prepare and conduct a study recruiting surgical staff neurosurgeons, orthopedic surgeons, fellows, and residents in orthopedic and neurosurgery. Study preparation includes the creation of the study protocol, participant consent forms, the face and content validity questionnaire, and obtaining the approval of the McGill Faculty of Medicine Research Ethics Board. During the trial, collect participant questionnaire responses as well as psychomotor data using the haptic device of the simulator system. Following study termination, statistically analyze the results of the questionnaire. Furthermore, leverage the data collected during the trial to generate novel metrics of surgical performance based on expert opinion, publications that focused on the Oblique Lateral Lumbar Interbody Fusion (OLLIF) surgery, and novel metrics derived from the data. Conduct appropriate statistical analyses for construct validity on the generated metrics to demonstrate statistical significance of surgical performance.

Task 2:

Hypothesis 2: Using novel surgical metrics, a multilayered ANN can objectively classify different levels of surgical expertise with a minimum accuracy of 80% and uncover composites of surgical performance that uniquely define expertise.

Side Study: Develop and test a novel multilayered ANN approach on prior spine simulator data.

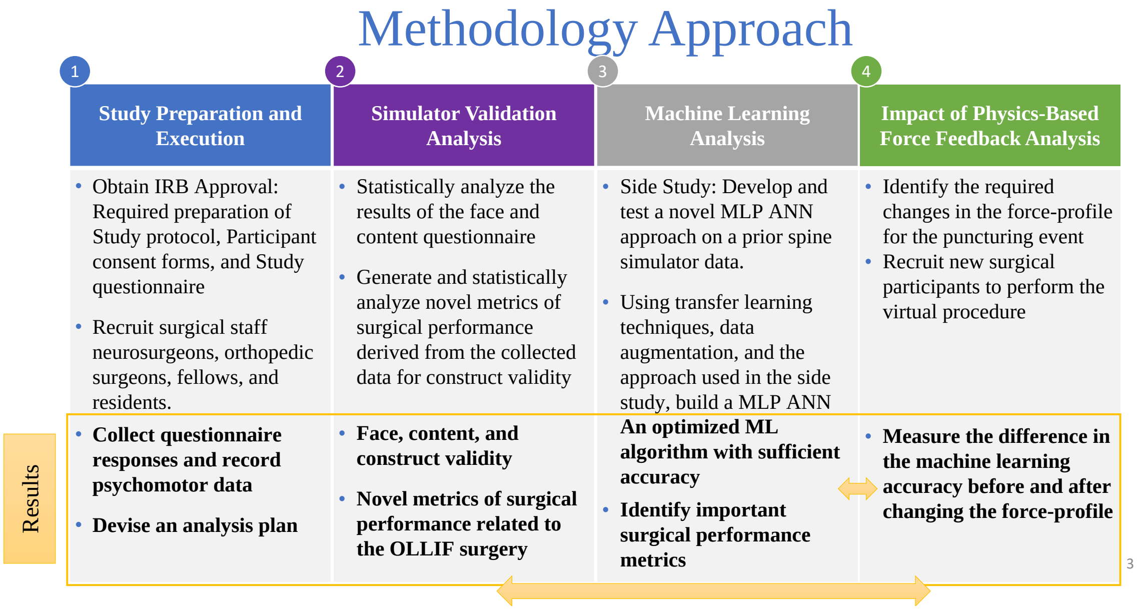
Utilizing a Multilayer Perceptron Artificial Neural Network to Assess a Virtual Reality Surgical Procedure (Published) – Journal of Computers in Biology and Medicine.

Objective 2: Using the novel metrics generated in Objective 1, build, train, and test a multilayer perceptron (MLP) artificial neural network in classifying and analyzing surgical performance. Leveraging data augmentation techniques and transfer learning using the model developed in the side study, identify feature importance of the trained ANN and subsequently validate the novel approach by comparing results to the permutation feature importance method.

Task 3:

Hypothesis 3: The integration of physics-based force feedback in the novel mixed reality spinal surgical simulator results in a better classification accuracy of the trained ANN model by a minimum of 16.67% as compared to the performance of the ANN model on scores generated using non-realistic force profiles.

Objective 3: Identify the required change in the force-profile for the puncturing event of the virtual surgery using biomechanical principles, and subsequently recruit new surgical participants to perform the virtual procedure. Using the new generated performance scores, measure the difference in the trained machine learning models' accuracies before and after changing the force-profile and analyze the performances using the novel approach developed in Objective 2.



Methodology Approach

	1 Study Preparation and Execution	2 Simulator Validation Analysis	3 Machine Learning Analysis	4 Impact of Physics-Based Force Feedback Analysis
	• Obtain IRB Approval: Required preparation of Study protocol, Participant consent forms, and Study questionnaire • Recruit surgical staff neurosurgeons, orthopedic surgeons, fellows, and residents.	• Statistically analyze the results of the face and content questionnaire • Generate and statistically analyze novel metrics of surgical performance derived from the collected data for construct validity	• Side Study: Develop and test a novel MLP ANN approach on a prior spine simulator data. • Using transfer learning techniques, data augmentation, and the approach used in the side study, build a MLP ANN	• Identify the required changes in the force-profile for the puncturing event • Recruit new surgical participants to perform the virtual procedure
Results	• **Collect questionnaire responses and record psychomotor data** • **Devise an analysis plan**	• **Face, content, and construct validity** • **Novel metrics of surgical performance related to the OLLIF surgery**	**An optimized ML algorithm with sufficient accuracy** • **Identify important surgical performance metrics**	• **Measure the difference in the machine learning accuracy before and after changing the force-profile**

Figure 2-2 Summary of the Methodology Approach used to attain the Thesis Objectives

The methodology of the current thesis revolved around the preparation, execution, and analysis of a study recruiting surgical staff neurosurgeons, orthopedic surgeons, fellows, and residents in orthopedic and neurosurgery to perform the virtual surgical scenario on the newly developed VR/AR surgical simulator as described in Figure 2-2. The study preparation phase of the project involved obtaining the approval of the McGill Faculty of Medicine Research Ethics Board under ethics number A03-M15-20A, which required the development of the study protocol, participant consent form, and the face and content validity questionnaire. The trial required recruiting surgeons of varying expertise to firstly complete the surgical simulation followed by answering the developed questionnaire for face and content validity. A small subgroup of expert participants was recruited in a side-by-side cadaver study, where the experts perform the surgery on a cadaver followed directly by completing the virtual procedure. During a simulation run, psychomotor data relating to the participants' use of the surgical tools were collected. The collected data included variables such as position, time, and angles of the simulated surgical tools, as well as applied forces and torques, removed volumes, and surgical tool contacts of specific anatomical structures. The recorded data were extracted and processed to generate surgical performance metrics that were used as a set of criteria to assess the performance of the participants in the virtual procedure. For example, position and time were combined to generate velocity metrics, forces and contact detection were used to determine the forces used when removing anatomical structures, and position and contact detection were used to determine the path length used while interacting with anatomical structures.

Following data collection, participants were prospectively divided into three groups based on the surgical training level: Post-Resident group (included post-residents and consultants in neurosurgery and orthopedic surgery), Senior-Resident group (included PGY[3] 4-6 neurosurgical residents and PGY 4-5 orthopedics residents), and Junior-Resident group (included PGY 1-3 neurosurgery residents, PGY 1-3 orthopedics residents). Face validity and content validity part of objective 1 as presented in Chapter 3 were established by analyzing the questionnaire scores given

[3] PGY: Postgraduate year denoting the progress of the postgraduate medical resident in the residency program.

by the recruited surgeons in the Post-Resident group. Construct validity (objective 1 and Chapter 3) was established by statistically analyzing the differences in the scores of the surgical performance metrics generated from the collected data during the study among the three groups. Furthermore, the surgical performance metrics were subsequently used in a machine learning analyses to further shed light on aspects of surgical performance that defines expertise (objective 2 and Chapter 4). Objective 3 required further recruitment of participants followed by varying the pre-set force-profiles generated by the haptic device of the developed simulator. Comparing and analyzing the surgical performance metrics prior and post changing the force-profiles shed light on the importance and impact of the used physics-based forces as presented in Chapter 5.

Chapter 3. Validation Studies of the Surgical Simulator

3.1 Background of First Article

This study aimed to establish the face, content, and construct validity of the newly developed VR/AR surgical simulator. As outlined in Chapter 2, this was attained by conducting a surgical simulator trial, recruiting spine surgeons with varying level of expertise, and obtaining ethics approval under ethics number A03-M15-20A. As a result, a detailed study protocol was developed to fulfill this objective. The protocol clearly outlined the trial's objectives and hypotheses, methods and procedures (including recruitment, data gathering, data storage, and study questionnaires), statistical analyses with associated power studies, and ethical considerations, such as confidentiality and consent. An initial power study indicated a target number of participants necessary to achieve sufficient statistical power. Subsequently, another power study was conducted using preliminary data collected on metrics that were borderline statistically significant to ensure the adequacy of the sample size for detecting meaningful differences. Special consideration was given to the development of the face and content validity questionnaires to accurately capture aspects of the simulation essential for establishing validity. Specifically, questions were designed to elicit meaningful feedback from expert surgeons regarding both the visual realism, as characterized by face validity, and skill realism, as outlined by content validity, ensuring the questions clearly distinguished between the two aspects of validity. Moreover, these aspects were evaluated in both the VR and AR dimensions of the simulation. In the VR context, assessments were made on the graphical appearances, movements, and haptic feedback of the virtual tools as the user interacted with virtual surfaces and structures. Conversely, in the AR domain, evaluations centered on the overall realism of the surgical setup including fluoroscopy, neuro-monitoring and navigation tools, as well as the appearances and maneuverability of the physical tools, including the tactile feedback experienced from the benchtop spine model, which incorporates intrinsic spine components. The questionnaires also took into account the side-by-side cadaver comparison study, a novelty of this research. In this comparison, expert surgeons from DePuy Synthes (part of Johnson & Johnson Medical Inc.) sequentially performed a similar surgery on a cadaver and then engaged with the simulation. In general, each questionnaire item prompted participants to draw upon their surgical knowledge and experience. For those specifically involved in the side-by-side cadaver study, they were further guided to compare aspects of the simulator directly with the cadaveric procedure they had just executed, ensuring that feedback was rooted in an immediate, tactile comparison. The use of the median of responses as a criterion to assess the validity of face and content questionnaires was chosen for its robustness to outliers and skewed data. The median provides a clearer consensus among responses, indicating that half of the responses were above a certain value. This approach helps to maintain centrality, unlike the mean, which can be shifted by extreme values.

During a trial run, both psychomotor interactions and questionnaire responses were recorded. As described in Chapter 2, participants were categorized into three distinct groups during the post-trial analysis: the Post-Resident group (comprising fellows and consultants in neurosurgery and orthopedic surgery), the Senior-Resident group (consisting of PGY 4-6 neurosurgical residents and PGY 4-5 orthopedics residents), and the Junior-Resident group (including PGY 1-3 neurosurgery residents and PGY 1-3 orthopedics residents). To assess face and content validity, questionnaire responses from expert post-residents, including those involved in the cadaveric side-by-side trial, were exclusively considered. The other two groups were utilized to measure agreement between the questionnaire responses. This approach helped identify specific aspects of the simulator where surgeons' perspectives evolve with increasing surgical experience, pinpoint areas that are well-developed or need enhancement, and highlight those aspects that warrant a comprehensive change.

Psychomotor data underwent a pre-processing phase before being combined to derive innovative surgical performance metrics. These metrics were derived based on consultations with expert surgeons, surgical literature on minimally invasive spine surgeries – specifically focusing on MI-LIF – and elements unique to the current simulation. Initially, a comprehensive list of 276 surgical performance metrics was generated. Drawing on previous publications by our research group and its collaborators, these metrics were classified into three main categories: safety, motion, and efficiency. The safety category encompasses metrics assessing actions crucial to patient safety. For instance, it measures forces exerted on anatomical structures and interactions – like the forces applied to, contacts made with, and distances maintained from – critical structures such as nerves and cauda. Motion metrics evaluate the stability and consistency of the surgical tools' movements during the procedure. These are characterized by features like tool velocities, tool accelerations, and changes in tool acceleration as determined by the jerk – a measure of tool tremors. Lastly, efficiency metrics gauge the surgeon's proficiency in achieving the surgical objective optimally. Features in this category include the duration taken for completion, the shortest and safest tool path length to the surgical site, and volume metrics like the removal of specific structures. Volume removals do not only assess unnecessary tissue removal in clearing the pathway to the surgical zone but also evaluate the volume removed from distinct structures, like the nucleus and annulus, essential for the successful completion of the procedure. Chapter 4 subsequently used the developed surgical performance metrics in a machine learning analysis to shed light on aspects of surgical expertise.

To that end, the feedback from the questionnaires along with the recorded data from this trial were used to assess the face, content, and construct validity in the following study entitled “Face, Content, and Construct Validity of a Novel VR/AR Surgical Simulator of a Minimally Invasive Spine Operation”. The attainment of Objective 1 and Hypothesis 1 are presented in the manuscript for which the contribution of the first author is 85%. The manuscript was published online in the Medical & Biological Engineering & Computing Journal on February 26, 2024 (https://doi.org/10.1007/s11517-024-03053-8). Furthermore, subsets of this work were presented at (1) the Orthopedic Research Society (ORS) and Philadelphia Spine Research Society (PSRS) 6th International Spine Research Symposium held in Pennsylvania, USA in November 2022, and (2) the 11th Interdisciplinary World Congress on Low Back & Pelvic Girdle Pain held in Melbourne, Australia in November 2023, under the name “Validating a Novel VR/AR Spinal Surgical Training Device with Focus on Physics-Based Force Feedback’.

3.2 Article 1: Face, Content, and Construct Validity of a Novel VR/AR Surgical Simulator of a Minimally Invasive Spine Operation

Sami Alkadri, Rolando Del Maestro, Mark Driscoll

Author names:

Sami Alkadri B.Eng., Ph.D Student [1,3], Rolando F. Del Maestro MD, PhD [3], Mark Driscoll, PEng., Ph.D., Associate Professor [1,2]

Institutional affiliations:

(1) Musculoskeletal Biomechanics Research Lab, Department of Mechanical Engineering, McGill University, Macdonald Engineering Building, 815 Sherbrooke St W, Montreal, H3A 2K7, QC, Canada.

(2) Orthopaedic Research Lab, Montreal General Hospital, 1650 Cedar Ave (LS1.409), Montreal, Quebec, Canada, H3G 1A4

(3) Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology & Neurosurgery, Montreal Neurological Institute, McGill University, 2200 Leo Pariseau, Suite 2210, Montreal, Quebec Canada, H2X 4B3.

Corresponding author:

Mark Driscoll

Mailing Address: Macdonald Engineering Building, RM 153, 817 Sherbrooke St W, Montreal, Quebec H3A 0C3, Canada

Phone: 514-398-6299

Fax: 514-398-7365

Email Address: mark.driscoll@mcgill.ca

3.2.1 Abstract

Background

Mixed-reality surgical simulators are seen more objective than conventional training. The simulators' utility in training must be established through validation studies.

Objective

Establish face-, content-, and construct-validity of a novel mixed-reality surgical simulator developed by McGill University, CAE-Healthcare, and Depuy Synthes.

Design

This study, approved by a Research Ethics Board, examined a simulated L4-L5 oblique lateral lumbar interbody fusion (OLLIF) scenario. A 5-point Likert scale questionnaire was used. Chi-square test verified validity consensus. Construct validity investigated 276 surgical performance metrics across three groups, using ANOVA, Welch-ANOVA, or Kruskal-Wallis tests. A post-hoc Dunn’s test with a Bonferroni correction was used for further analysis on significant metrics.

Setting

Musculoskeletal Biomechanics Research Lab, McGill University, Montreal, Canada. DePuy Synthes, Johnson & Johnson Family of Companies, research lab.

Participants

34 participants were recruited: spine surgeons, fellows, neurosurgical and orthopedic residents. Only seven surgeons out of the 34 were recruited in a side-by-side cadaver trial, where participants completed an OLLIF surgery first on a cadaver, then immediately on the simulator. Participants were separated a priori into three groups: post-, senior-, and junior-residents.

Results

Post-residents rated validity, median>3, for 13/20 face-validity and 9/25 content-validity statements. Seven face-validity and 12 content-validity statements were rated neutral. Chi-square

test indicated agreeability between group responses. Construct validity found eight metrics with significant differences (P<0.05) between the three groups.

Conclusions

Validity was established. Most face-validity statements were positively rated, with few neutrally rated pertaining to the simulation’s graphics. Although fewer content-validity statements were validated, most were rated neutral (only four negatively rated). The findings underscored the importance of using realistic physics-based forces in surgical simulations. Construct-validity demonstrated the simulator's capacity to differentiate surgical expertise.

Keywords

VR/AR Surgical Simulation. Face, Content, & Construct Validity. Physics-based haptic feedback.

3.2.2 Introduction

Virtual reality (VR) surgical simulators have been rapidly adopted as a more objective method of training and evaluating surgical technical skills, especially when compared to conventional training methods [1, 2]. VR training modules provide safe and controlled training platforms that allow residents to further develop their surgical skills [3]. Furthermore, the ability to generate automated scoring systems further supports the notion of integrating VR simulator systems in the training and the objective assessment of surgical residents in performing procedures. VR simulators collect enormous sets of data pertaining to the psychomotor interactions of the user during the completion of the simulated tasks. Such data are often transformed into performance metrics that play an important role in training and assessing surgical trainees. Recent developments have coupled the VR systems with haptic technology, which allowed trainees to develop their “feel” of the procedure before performing in-vivo surgeries. This haptic technology allows real-time force-feedback which enhances the authenticity of the training programs [3]. In fact, our group strives to demonstrate the potential benefits of incorporating accurate physics-based haptic technology on learning outcomes through detailed quantification of surgical forces [4].

Despite the advancements of VR simulators in the surgical field, spinal surgeries lagged behind other disciplines [3]. In particular, a clear gap was present in VR simulators for spinal minimally invasive surgeries; until recently, spinal simulation training was still in its infancy with very little research in the past two decades to create a spinal surgical simulator [3]. Moreover, the high demand of spinal surgeries led to the continuous improvements of both the surgical techniques and the skills of the surgeons. Numerous efforts were directed to establish novel minimally invasive spine surgical procedures that enhance patient safety and recovery [5]. Coupling the high demand for novel minimally invasive spine surgeries (MISS) with the range of difficulty associated with spine surgery has led to the development of novel spinal VR simulators with haptic feedback [6, 7]. These simulator platforms can deconstruct complex surgical procedures such as the Oblique Lateral Lumbar Interbody Fusion (OLLIF) into discreet steps allowing trainees to concentrate on specific technical skills in need of enhancement rather than those already acquired [7-9]. The OLLIF surgery requires learners to master a broad spectrum of surgical techniques and each of these components can be assessed and trained utilizing virtual reality simulators [7, 10]. One such system is the VR/AR training platform developed by our group to train orthopedic and neurosurgeons on a novel minimally invasive OLLIF surgery.

The promising preliminary results exhibited by VR surgical training systems further encouraged its adaptation to surgical curriculums [11]. However, proper fundamental validation studies of the simulator systems are required. More specifically, the utility of such simulators in effectively training and assessing surgical trainees must be established through foundational subjective and objective validation steps, namely: face, content, and construct validity.

Face and content validity are established using a questionnaire. Face validity is the extent to which the developed simulation environment mimics the real surgery, whereas content validity is the extent to which the developed system is representative of the skills required to successfully complete the real surgery [12]. Construct validity refers to the ability of the simulator to distinguish between different levels of surgical expertise [1, 13, 14]. It is an objective validation step that relies on the enormous sets of data generated from the interactions of the user during the simulated task. Such data are often transformed into surgical performance metrics that play an important role in not only establishing construct validity, but also in training and assessing surgical trainees. The use of statistical analyses is the gold standard for establishing construct validity [1, 13, 14]. Statistically significant differences in the scores among experts and trainees on the generated surgical performance metrics highlight the ability of the simulator to adequately differentiate between levels of surgical expertise.

While recent literature reflects a growing interest in more advanced forms of validation, such as concurrent and predictive validity, there is a discernible gap in studies demonstrating concrete foundational face, content, and construct validations [15-17]. Concurrent and predictive validity, evaluate how closely the outcomes of a newly developed simulator align with those of an established gold standard and assess whether skills acquired on the simulator yield better results in real surgical settings, respectively. The current research aims to address this gap by focusing on and establishing the foundational validation steps. These initial validations are crucial as they establish the basic authenticity and educational relevance of the simulator, which is a necessary precursor to more complex forms of validation like concurrent and predictive validity [18].

Hence, the scope of our work is deliberately concentrated on face, content, and construct validity of a novel OLLIF surgical approach that has not been explored previously. Therefore, the generated surgical metrics used as part of the construct validity step are considered unique and novel as they describe aspects of this new surgical approach. Furthermore, the study sheds light on the impact of using accurate physics-based force feedback on surgical simulation training, an aspect that to the best of the authors' knowledge is not previously studied. Lastly, the novelty explored in this study also includes a unique face and content validation approach by making use of a side-by-side cadaver study where participants directly complete the surgical scenario on a cadaver followed by completing the same surgical operation on the simulator.

3.2.3 Material and Methods

3.2.3.1 The VR/AR Simulator & The Simulated Scenario

This study utilized a novel VR/AR surgical training system developed by McGill University in affiliation with CAE Healthcare and Depuy Synthes part of Johnson & Johnson. The surgical simulator under consideration is a physics-based simulator of a minimally invasive spine single level fusion. The geometry of the surgical scenes in the simulator are reconstructed from patient specific data. The simulation runs on a high-performance gaming laptop (i7-8750H) with Windows 10 operating system. Similar to the surgical reality, the rendered images are displayed on two flat panel monitors to match the interface in the operating room: a built-in monitor and an external touch screen monitor. The monitor on the left in Figure 3-1 provides general surgical guides including a recorded animation displaying how to operate instruments during a step. The other monitor is an interactive touch screen displaying the laparoscopic views of the surgical area with which the surgeons interact. Haptic feedback is provided from a combination of a six-degrees of freedom ENTACT W3D device and a benchtop model that includes 3D printed vertebrae components, also exemplified in Figure 3-1. This is conveyed to the surgeons hand via analogue surgical tools interchangeably hooked up the haptic system.

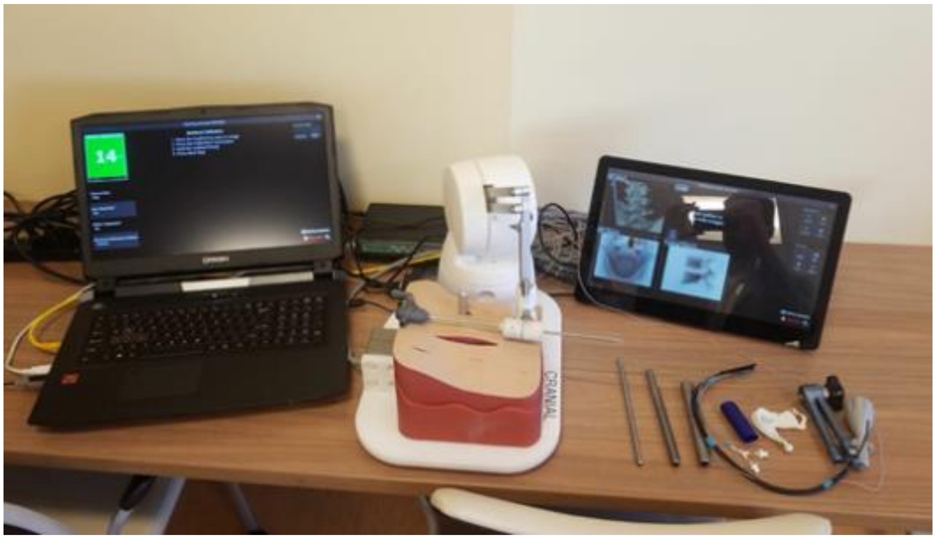


Figure 3-1 The summarized simulator layout. Left is the laptop runs 120Hz display, which indicates the instruction of the surgery process. The haptic device and benchtop model are in the middle. And right is the external display runs 60Hz which indicates the four cameras that demonstrate the surgical area. The surgeon operates the haptic device based on the visual feedback from both monitors.

The simulation focusses on three phases of the spinal surgery: gaining access through the back muscles, removing the intervertebral disc, and inserting graft and a spinal cage. The detailed steps along with the surgical tools used at each phase are demonstrated in Figure 3-2. The first phase of the simulated surgery includes the use of a multiprobe tool to gain access to the surgical area. Phase 2 requires the surgeon to firstly use a Burr tool for drilling and performing a facetectomy, followed by using the Concord tool’s suction mechanism to remove the remaining parts of the disc. Lastly, the surgeon is required to insert graft and a cage using the graft and cage insertion tools. The virtual volumetric model contains artificial muscle layers and an intervertebral disc, each providing realistic force feedback through interaction with the haptic device. The force feedback replicates the resistance provided by the instruments when penetrating through the muscles during an actual surgery using tailored mechanical properties. Prior to the start of the simulation, participants were made aware of all steps and instruments needed to complete the procedure via verbal and written instructions. No time limit was imposed on completing the simulated scenario.

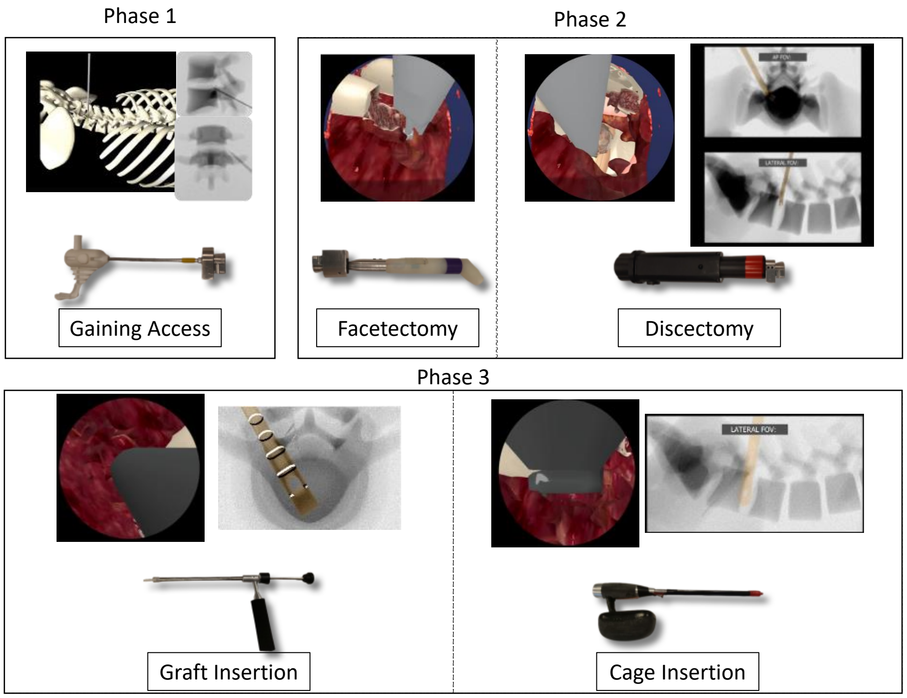


Figure 3-2 The three phases of the simulated surgery: Phase 1 includes gaining access to the disc using a Multitool; Phase 2 includes facetectomy using a Burr Tool followed by a discectomy using a Concord Tool; Phase 3 includes inserting graft followed by inserting a cage using the respective tools.

3.2.3.2 Participants

Thirty-five participants were recruited to perform the virtual reality OLLIF scenario. Seven expert orthopedic surgeons out of the thirty-four participants were recruited in a side-by-side cadaver trial, where participants completed a minimally invasive spinal fusion surgery on a cadaver, then immediately repeated the identical procedure on the surgical trainer/simulator. The remaining participants completed the trial without performing a cadaver surgery. All 35 participants were included in the face and content validity analyses. Due to errors during the simulation runs, only 24 individuals were included in the construct validity analysis: 10 post-residents, 5 senior residents, and 9 junior residents. Table 3-1 and Table 3-2 present the demographics and the difference in experiences and knowledge of the 34 participants, respectively. The participants were divided into three groups: A post-resident group (3 neurosurgeons, 12 spine surgeons, 2 spine fellows, and 2 neurosurgical fellows), a Senior-Resident group (4 PGY 4-6 neurosurgery and 3 PGY 4-5 orthopaedics residents), and a Junior-Resident group (4 PGY 1-3 neurosurgery and 5 PGY 1-3 orthopaedics residents). This study was approved by an appropriate Research Ethics Board. All participants signed an approved written consent form prior to completing the simulation of the virtual spine surgery which took on average 1-hour to complete.

Table 3-1 Demographics of the post-resident, senior-resident, and junior-resident groups.

	Junior Residents	**Senior Residents**	**Post-Residents**
No. of individuals	9	7	19
Sex			
Male	8	6	18
Female	1	1	1

Surgical Specialty Level of Training	Neurosurgery	Orthopaedic Surgery
PGY 1-3	4	5
PGY 4-6	4	3
Fellows	3	2
Consultants	2	12

Table 3-2 Differences in previous experience, knowledge, and comfort level of the groups.

	Junior Residents	**Senior Residents**	**Post-Residents**
No. of individuals in each group who:			
Have previous experience using a surgical simulator	2 (22%)	5 (71%)	17 (89%)
Assisted on a TLIF	7 (77%)	7 (100%)	17 (89%)
Performed a TLIF solo	0 (0%)	0 (0%)	14 (73%)
Medina self-rating on 5-point Likert scale:			
Textbook Knowledge of a TLIF	3.0 (3.0 – 4.0)	3.0 (3.0 – 4.0)	3.5 (1.0 – 5.0)
Surgical Knowledge of a TLIF	3.0 (2.0 – 4.0)	3.0 (3.0 – 4.0)	3.5 (1.0 – 5.0)
Comfort level performing a TLIF with a consultant in the room	3.0 (1.0 – 4.0)	4.0 (2.0 – 5.0)	4.5 (2.0 – 5.0)
Comfort level performing a TLIF solo	1.0 (1.0 – 2.0)	2.0 (1.0 – 4.0)	3.0 (1.0 – 5.0)

3.2.3.3 Face and Content Validity

All participants completed a questionnaire pertaining to face and content validity of the developed simulator using a 5-point Likert scale, where 1 indicated "strongly disagree" and 5 indicated "strongly agree. There is no consensus on an acceptable median for sufficient face and content validity in the literature. In the current study, sufficient validity is assumed to be achieved if a median > 3.0 on a 5-point Likert scale is obtained for the Post-resident group. Usually, face and content validity rely solely on the evaluations of the training system by expert surgeons [1, 12]. However, this study utilized responses made by non-experts (Junior and Senior-Resident groups) to rate the consensus among experts and trainees on certain aspects of the simulator pertaining to both face and content validity [1, 12]. A Chi-Square test was utilized to establish statistical significance of validity consensus. Comparing the consensus between the experts and trainees may be used to analyze the change in perspective with surgical experience [1]. This also allows for detailed analyses of validity that pinpoints aspects of the simulator that are adequately developed, requires further improvements, or require a complete change [1].

The questionnaire was designed to gather detailed feedback from expert surgeons on two primary aspects: visual realism (face validity) and skill realism (content validity), evaluated within both the VR and AR dimensions of the simulation. Surgical and industry experts were consulted to ensure the questions were pertinent, clear, and effectively targeted the intended aspects of validity. For face validity, the questionnaire differentiated between the VR and AR components of the simulator, assessing graphical appearances of virtual anatomical structures and tools in VR, and the overall realism of the surgical setup in AR, including fluoroscopy, neuro-monitoring, and navigation tools. Content validity was similarly bifurcated, with VR questions examining the movements and haptic feedback of virtual tools, and AR questions focusing on the maneuverability and tactile feedback of the physical tools. Additionally, the questionnaires incorporated elements from the side-by-side cadaver comparison study, an innovative aspect of our research. In this study, 7 expert surgeons from DePuy Synthes performed a Transforaminal Lumbar Interbody Fusion (TLIF) on a cadaver, followed by a simulation procedure. The subgroup completed the entire experiment within 1-hour to ensure that the participants contrasted their experience on the virtual procedure to that on the cadaveric surgery in a side-by-side comparison. This direct comparison enabled the questionnaire to prompt participants, especially those involved in the cadaver study, to draw on their surgical experience and make direct comparisons between the simulator and the cadaveric procedure, ensuring a grounded and immediate tactile feedback assessment.

3.2.3.4 Construct Validity

Construct validity was assessed using a priori metrics established independently for each Module. During a simulation procedure, the surgical simulator recorded a series of data relating to the participants' use of the surgical tools. The collected data included variables such as position, time, and angles of the simulated surgical tools, as well as applied forces, removed volumes, and surgical tool contacts of specific anatomical structures. In total 73 variables were collected throughout a simulation run. Subsequently, the recorded data were extracted and processed to generate surgical performance metrics that were used as a set of criteria to assess the performance of the participants in the virtual procedure. For example, position and time were combined to generate velocity metrics, forces and contact detection were used to determine the forces used when removing anatomical structures, and position and contact detection were used to determine the path length used while interacting with anatomical structures. A total of 276 metrics were initially generated based on expert opinion, publications that focused on spinal fusion surgical performance, and novel metrics derived from the data. Subsequently, all derived metrics data were normalized using z-score normalization to reduce impact of outliers. Metrics were divided into three categories: motion, safety, and efficiency.

For all the generated surgical performance metrics, normality was tested using the Shapiro-Wilk test. For normally distributed data, variance homogeneity was further tested using the Levene's test. To statistically measure the differences between the surgical groups, one of three statistical tests was used depending on the normality and variance homogeneity of the data. The standard Anova test was used if the data distribution was normal with equal variances across the groups. Welch Anova was used if normality was met but with heterogeneous variances. Lastly, Kruskal-Wallis parametric test was used for non-normally distributed data. A post-hoc Dunn’s test with a Bonferroni correction was utilized to investigate differences between groups on significant metrics.

3.2.4 Results

3.2.4.1 Participants

Table 3-2 highlights the main differences between the groups based on previous experience, knowledge and comfort levels performing and/or assisting in a TLIF (most similar procedure to simulated OLLIF). The senior-resident group (PGY 4 and higher) assisted in more TLIF surgeries and have a higher level of comfort assisting a TILF solo than the junior-resident group (PGY 1-3). Both the senior- and the junior-resident groups have no experience and a low comfort in performing a TILF solo. Despite being the highest group having performed and assisted in a TLIF, the post-resident group ratings demonstrated that some recruited surgeons were non-spinal specialty and do not have textbook or surgical expertise in the TLIF surgery (median 3.5; range 1.0 – 5.0). In fact, 11% of the post-resident group have not performed or assisted in a TLIF previously.

3.2.4.2 Face and Content Validity

The face and content validity questionnaire consisted of 45 statements, 20 statements assessed face validity and 25 statements assessed content validity. For face validity, post-resident group rated 13 statements positively (median > 3) and seven statements neutrally (median = 3) with no negatively rated statements (median < 3). For content validity, post-resident group rated nine statements positively (median > 3), 12 statements neutrally (median = 3), and four statements negatively (median < 3). The four negatively rated statements were all pertaining to interactions of the users with the Burr tool. The median responses for each of the face and content validity statements are shown in Table 3-3 and Table 3-4, along with the corresponding p-values for a chi-square test to assess the agreement in the response between junior, senior, and post-resident participants. All p-values were greater than 0.05, indicating no significant differences among group responses.

Table 3-3 Face validity median responses of the post-resident group with the Chi-square p-values assessing inter-group agreeability.

Validity Statements	Post-Residents Median Responses	Chi-Square P-Value
The Physical Multitool accurately resembles the real surgical tool.	4	0.356
The Virtual Multitool accurately resembles the real surgical tool.	4	0.638

I am able to accurately set up the benchtop model to resemble a real surgery through the use of the Fox Arm and port.	4	0.279
The orientation and angulation of the port in the physical world matches what is seen in the virtual world.	4	0.567
The Physical Burr accurately resembles the real surgical tool.	4	0.807
The Virtual Burr accurately resembles the real surgical tool.	3	0.177
The Physical Tissue Retractor accurately resembles the real surgical tool.	4	0.331
The Virtual Tissue Retractor accurately resembles the real surgical tool.	3	0.547
The Physical Concorde Clear accurately resembles the real surgical tool.	4	0.627
The Virtual Concorde Clear accurately resembles the real surgical tool.	4	0.487
The Physical Graft Delivery Device accurately resembles the real surgical tool.	4	0.341
The Virtual Graft Delivery Device accurately resembles the real surgical tool.	4	0.637
The Physical Cage Insertion Device accurately resembles the real surgical tool.	3.5	0.1
The Virtual Cage Insertion Device accurately resembles the real surgical tool.	4	0.511
The animation representing the cage insertion is similar to a real surgery.	3	0.802
The visual guides shown during the simulation are similar to the ones used during a real surgery.	4	0.586
The simulator system setup – including the positioning of the screen, the haptic device, and the benchtop model – is similar to a real surgical setup.	3	0.515
The visual graphics shown in the Port Cam view are similar to reality.	3	0.324
The internal impressions of the tissue model shown in the Port Cam view are similar to reality.	3	0.554
The external impressions of the tissue model shown in the Port Cam view are similar to reality.	3	0.67

Table 3-4 Content validity median responses of the post-resident group with the Chi-square p-values assessing inter-group agreeability

Validity Statements	**Post-Residents Median Responses**	**P-Value Chi-Square**
I am able to maneuver the Multitool similar to a real surgery when puncturing on the model	4	0.527

The forces experienced using the Multitool during the gaining access step are similar to those experienced during a real surgery	3	0.689
The force difference between the soft tissue layers is appropriate.	3	0.341
I can clearly distinguish between the soft and hard tissue.	4	0.054
I am able to remove bone and soft tissue as needed to gain IVD access	4	0.536
I can clear an adequate access area.	4	0.769
I am able to maneuver the Burr tool similar to a real surgery.	2	0.051
The amount of bone removed using the Burr tool during each pass of the facetectomy step is similar to a real surgery.	2.5	0.722
The bone forces experienced using the Burr Tool during the facetectomy step are similar to those experienced during a real surgery:	2	0.158
The soft tissue forces experienced using the Burr Tool during the facetectomy step are similar to those experienced during a real surgery:	2	0.42
I am able to use the Tissue Retractor Tool to protect the nerve similarly to a real surgery	3	0.546
The method of selecting annulotomy size is reasonable.	3	0.541
I am able to remove the amount of soft tissue that I wanted.	3	0.115
I am able to maneuver the Concorde Clear tool similar to comparable Curettes in a real surgery.	4	0.527
The forces experienced using the Concorde Clear tool during the discectomy step are similar to those experienced using comparable Curettes during a real surgery	3	0.313
The torques experienced using the Concorde Clear tool during the discectomy step are similar to those experienced using comparable Curettes during a real surgery	3	0.319
I am able to remove IVD similar to a real surgery.	3	0.494
I am able to scrape and prepare the endplates similar to a real surgery.	3	0.274
I am able to tell how far into the IVD I have penetrated	3	0.421
The amount of disc removed as presented by the simulator metrics matches my expectations.	3.5	0.302
When impacting on the Graft Delivery Device the changes at each mallet impact resemble a real surgical procedure.	4	0.71
When impacting on the Cage Insertion Device, the changes at each mallet impact resemble a real surgical procedure.	3	0.533

The movement of the Physical tools resemble a real surgical procedure as the graft is inserted in the IVD	4	0.456
The movement of the Physical tools resemble a real surgical procedure as the cage is inserted in the IVD	4	0.253
The overall tasks and the associated skills required to complete the simulation run are similar to those required to complete a real surgery	3	0.179

3.2.4.3 Construct Validity

Construct validity results showed significant differences between the three groups for eight metrics (Table 3-5). Box plots and pairwise comparisons of significant metrics are presented in Figure 3-3. The significant metrics spanned all three metric categories of motion, efficiency, and safety. Furthermore, the metrics differentiated the performance of the three groups while performing the most critical steps of the procedure.

Table 3-5 Construct validity results.

Surgical Step	Significant Metrics	3 Group Split (Junior vs Senior vs Post)		
		Data Distribution	Variance Homogeneity	Test Statistic
Gaining Access	**Total Multi-Tool Tip Path Length**	Normal	Homogenous Variance	ANOVA: P=0.032
Facetectomy, Discectomy, & Annulotmy	**Number of Sign Changes of the Acceleration of the Burr Tool in the Z-Direction**	Normal	Homogenous Variance	ANOVA: P=0.022
	Average Jerk of the Concorde Tool in the Y-Direction	Non-Normal	-	KruskalWallis: P=0.04
	Volume Removed of the L4 Endplate above the Annulus Fibrosus	Normal	Homogenous Variance	ANOVA: P=0.041
	Volume Removed of the L5 Endplate under the Annulus Fibrosus	Normal	Homogenous Variance	ANOVA: P=0.042
	Maximum Force Applied on the IAP Using the Burr Tool	Non-Normal	-	KruskalWallis: P=0.036
	Average Distance to the Nerve while operating the Concorde Tool	Normal	Homogenous Variance	ANOVA: P=0.03
	Average Distance to the Cauda while operating the Concorde Tool	Normal	Homogenous Variance	ANOVA: P=0.032

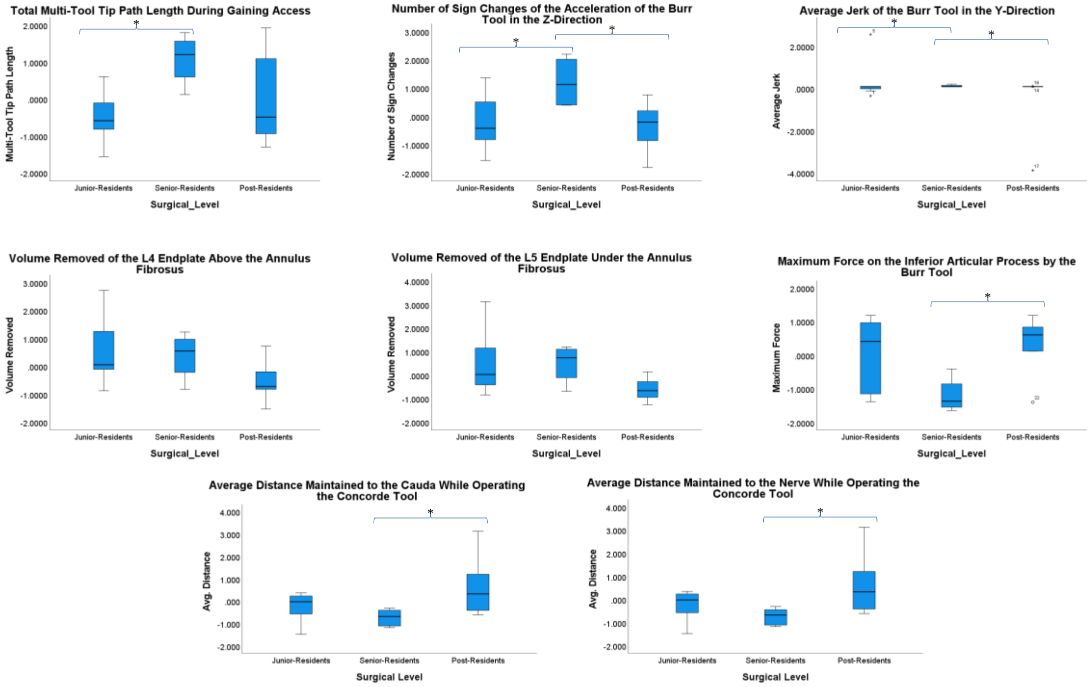


Figure 3-3 Box plots & post-hoc Dunn's test with a Bonferroni correction of the 8 statistically significant metrics.

3.2.5 Discussion

3.2.5.1 Overall Validity

The newly developed VR/AR surgical simulator has been shown to attain face, content, and construct validity, making it a promising formative educational tool of a novel OLLIF surgical approach that has not been explored previously. Therefore, the generated surgical metrics used as part of the construct validity step are considered unique and novel as they describe aspects of this new surgical approach. The novelty explored in this study also includes a unique face and content validation approach by making use of a side-by-side cadaver study where participants directly complete the surgical scenario on a cadaver followed by completing the same surgical operation on the simulator. The study also gives an insight into the importance of using accurate physics-based force profiles in spinal surgical training.

3.2.5.2 Face and Content Validity

The results of the subjective validity assessment of the new surgical simulator show a high level of face validity with 13 out of the 20 statements reaching a median score greater than 3 on a 5-point Likert scale. The high number of positively rated statements and the lack of any negative feedback in the face validity questionnaire indicate that the system was perceived as having a good overall level of realism. Only seven statements were neutrally rated and did not reach validity (Table 3-3). Among the virtual tools displayed during the procedure, only the virtual Burr tool and the virtual tissue retractor did not reach validity, with the rest of the tools in both the physical and virtual versions having sufficient validation in the face validity questionnaire. The rating of the appearance of the virtual Burr tool might have been impacted by the negatively rated user experience of the tool. In fact, the only negatively rated statements in the content validity questionnaire were related to the interactions of the participants with the Burr tool, which is further discussed in more detail later in this section. The virtual tissue retractor tool was the only tool with incomplete responses among participants; the use of the tissue retractor tool was optional during the simulation as in the case of the real surgery and some participants chose not to utilize the tool, which may have contributed to the tool not reaching validity. Nevertheless, the physical versions of both the Burr and the tissue retractor tools reached face validity, indicating that the graphics were not as effective in mimicking reality. In fact, six out of the seven neutrally rated statements were related to the graphics and animation, indicating that there may be room for improvement in this aspect of the simulation. However, refining the visual graphics and animations of a simulation negatively impacts the computational time per frame, which in turn impacts the ability of the simulator to maintain a realistic interactive experience [19]. When the frame rate per second becomes less than 20 Hz, discontinuous and lagging graphic feedback affects the user experience, which is related to the rate at which the brain processes visual data [19]. The current simulation was optimized to maintain the minimum required frame rate per second that ensures a realistic interactive experience while maximizing the realism of the graphics and animations [20]. The lack of any negative feedback in the face validity questionnaire supports the optimization decision and the fact that a good balance was found between realistic graphics and a realistic interactive experience.

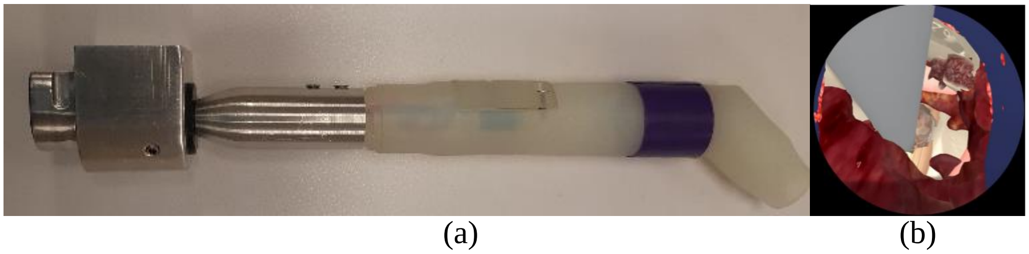


Figure 3-4 (a) The physical Burr tool; (b) Camera view of the virtual Burr tool with a shield during the simulation.

Despite the relatively lower number of validated statements in the content validity questionnaire, the majority of the statements that did not reach validity were rated neutral and only four statements were negatively rated, which were all pertaining to the Burr tool (Table 3-4). A recurring comment during the course of the trial was made regarding the use of a shield with the Burr tool as demonstrated in Figure 3-4. The reduced depth perception of the camera view in the simulation coupled with the shield resulted in difficulties while handling the tool, which is demonstrated by the low median rating of the statement assessing the maneuverability of the Burr tool (Table 3-4). In general, one must be accurate in investigating the subjective rated aspects of a simulator system, as what can be perceived as a negative aspect of a simulator might be essential to capture reality. Careful investigation is required to determine if an added difficulty is representative of the skills required to complete the real surgery or if its an unnecessary addition that needs refinement. While the overall graphics require further refinement, in the case of the current simulation, the reduced depth perception is essential to capture the true difficulties faced in the actual MI surgery. Therefore, the feedback obtained on the level of difficulty in handling the Burr tool further supports the notion of face and content validity. Paradoxically, the importance of using realistic physics-based force profiles in surgical simulation is highlighted by the negatively rated statements regarding the forces experienced while operating the Burr tool. The Burr tool is the only tool in the simulation that is programmed with forces that are not based on cadaveric experiments. User interactions with the Multi-tool and the Concorde Tool that incorporated realistic forces based on cadaveric experiments were rated either positively or neutrally. This finding further supports the use of accurate physics-based force profiles in surgical simulations.

The chi-square test was further used to assess the agreeability between group responses. For each statement, the null hypothesis was that the three groups had no differences in the ratings. All p-values presented in Table 3-3 and Table 3-4 had values greater than 0.05, failing to reject the null hypothesis and indicating that no statistically significant differences exist. These results support the notion that the groups were in agreement when assessing the aspects of the simulation.

3.2.5.3 Construct Validity

The eight statistically significant metrics were derived from the surgical tools utilized in the most critical steps of the procedure (gaining access, facetectomy, and discectomy steps) and spanned all three metric categories. Developed through a collaboration of expert surgeons' insights and existing surgical literature, these metrics crucially demonstrate the simulator's ability to differentiate between various levels of surgical expertise, which is fundamental for construct validity. This differentiation suggests that the simulator can effectively measure the specific skills it intends to. Furthermore, these metrics have the potential to be teachable objectives for junior surgeons. They provide quantifiable targets in critical aspects of surgical performance, offering a pathway for skill development towards the benchmarks of more experienced surgeons. Thus, this construct validity analysis not only validates the simulator's assessment capabilities but also hints at its potential as a comprehensive training tool, which could significantly contribute to the advancement of surgical education.

During the gaining access step, the efficiency of the surgeons in reaching the surgical area represented by the Multitool pathlength was significantly different among the groups. The results also indicated significant differences in handling the Burr tool in the facetectomy step and the Concorde tool in the discectomy step, highlighted by the Burr Tool acceleration sign changes and the Concorde tool average jerk, respectively. Six safety metrics were identified during the facetectomy and discectomy steps. Metrics fall under the safety category if their effect result in either direct or indirect risk of injury or danger to the patient. Indirect safety metrics include unnecessary removals of anatomical structures such as the unnecessary removals of the L4 and L5 endplates identified in Table 3-5. Direct safety metrics include metrics that capture the preservation of important anatomy during the procedure, such as forces applied on anatomical structures and the proximity maintained to critical structures such as the nerve or the cauda. The maximum forces applied on the Inferior Articular Process (IAP), and the average distances maintained to the nerve and cauda were identified as significant metrics in Table 3-5.

In general, a discontinuous learning pattern is characterized with non-sequential progression of skills while progressing from the junior-resident to the post-resident surgical level, passing through an inconsistent senior-resident level. Consider Figure 3-3, a clear discontinuous learning pattern can be seen in the motion and efficiency metrics. More specifically, both the post-residents and junior-residents were efficient with stable motions having seemingly smaller pathlengths and less directional changes in their motion as compared to the senior-residents. Despite the similarity in the performances of the junior and post-residents in the motion and efficiency metrics, they are attributed to different reasons. The expert post-resident group utilize precise and deliberate movements while the junior-residents are more reluctant and cautious. In the remaining metrics, it is not directly evident that a significant difference exists between the performances of the junior and senior-residents. The figure suggests that post-residents seem to remove less L4 and L5 endplates while being more wary of operating in critical proximity to the nerve and cauda when compared to the junior-residents and senior-residents.

The analysis done for construct validity is not just a validation of the simulator's effectiveness in distinguishing between different levels of expertise. It also lays the groundwork for its use as a comprehensive training tool, offering measurable and attainable goals for surgical skill improvement used for both training and assessment.

3.2.6 Conclusion

This study has established the face, content, and construct validity for the MI OLLIF simulated surgery on the newly developed VR/AR simulator. The unique side-by-side cadaver study and the use of accurate physics-based force profiles contributed in establishing the realism and educational value of the simulator. While some aspects, such as the graphics and animation, could be improved, the system has been optimized to balance realistic graphics with a realistic interactive experience.

The face and content validity of the simulator were largely favorable, with only a few negative ratings. The majority of issues encountered were related to the virtual Burr tool including the unrealistic force feedback for that particular tool as well as tool-handling difficulties. Upon further analysis, this feedback was shown not only to support the face and content validity of the simulator, but also to highlight the importance of using realistic physics-based force profiles in surgical simulations as used for other surgical tools in the current simulation. The construct validity of the simulator is supported by the significant differences in performance metrics across different levels of surgical expertise. The analysis validates the simulator's ability to differentiate various expertise levels and establishes it as a comprehensive training tool, providing measurable goals for enhancing surgical skills in both training and assessment contexts. A discontinuous learning pattern was observed in the motion and efficiency metrics, with post-residents and junior-residents displaying seemingly smaller pathlengths and fewer directional changes in their motion compared to senior-residents. In other metrics, post-residents demonstrated more precise and cautious behavior in terms of preserving important anatomy and maintaining safe distances from critical structures.

Overall, the VR/AR surgical simulator represents a promising formative educational tool for the OLLIF surgical approach. With further refinements and optimization, it has the potential to become an invaluable resource for training the next generation of surgeons in this innovative technique.

3.2.7 Compliance with Ethical Standards

Ethical Approval

The institutional IRB approval was received for the study protocol and consent forms. This research involved recruiting surgeons to perform the virtual surgery on the simulator. Proper consent was obtained.

Competing Interests Statement

No competing interests to declare.

Funding

NSERC Collaborative Research Development (CRD) Grant; Franco Di Giovanni Foundation; Brain Tumour Foundation of Canada Brain Tumour Research Grant; a Medical Education Research Grant from the Royal College of Physicians and Surgeons of Canada; and the Montreal Neurological Institute and Hospital.

3.2.8 References

[1] M. Goldenberg and J. Y. Lee, "Surgical Education, Simulation, and Simulators-Updating the Concept of Validity," (in eng), *Curr Urol Rep*, vol. 19, no. 7, p. 52, May 17 2018. https://doi.org/10.1007/s11934-018-0799-7

[2] M. Pfandler, M. Lazarovici, P. Stefan, P. Wucherer, and M. Weigl, "Virtual reality-based simulators for spine surgery: A systematic review," *The Spine Journal*, vol. 17, 05/01 2017. https://doi.org/10.1016/j.spinee.2017.05.016

[3] N. Vaughan, "A review of virtual reality based training simulators for orthopaedic surgery," (in eng), *Medical Engineering and Physics*, vol. 38, no. 2, p. 59, 2016.

[4] K. El-Monajjed and M. Driscoll, "Analysis of Surgical Forces Required to Gain Access using a Probe for Minimally Invasive Spine Surgery via Cadaveric-based Experiments towards use in Training Simulators," *IEEE Transactions on Biomedical Engineering*, pp. 1-1, 2020. https://doi.org/10.1109/TBME.2020.2996980

[5] "McGill University to partner with industry in developing virtual-reality training platform for spinal surgery," 11 JUN 2018. Available: https://www.mcgill.ca/newsroom/channels/news/mcgill-university-partner-industry-developing-virtual-reality-training-platform-spinal-surgery-287588

[6] S. Alkadri, "Kinematic Study and Layout Design of a Haptic Device Mounted on a Spine Bench Model for Surgical Training," Undergraduate Honours Program - Mechanical Engineering, Mechanical Engineering, McGill University, 2018.

[7] N. Ledwos, N. Mirchi, V. Bissonnette, A. Winkler-Schwartz, R. Yilmaz, and R. F. J. O. N. Del Maestro, "Virtual reality anterior cervical discectomy and fusion simulation on the novel sim-ortho platform: validation studies," *Operative Neurosurgery*, 2020.

[8] T. Cotter, R. Mongrain, and M. J. J. o. M. D. Driscoll, "Design synthesis of a robotic uniaxial torque device for orthopedic haptic simulation," vol. 16, no. 3, p. 031008, 2022.

[9] S. Patel, J. Ouellet, M. J. M. Driscoll, B. Engineering, and Computing, "Examining impact forces during posterior spinal fusion to implement in a novel physics-driven virtual reality surgical simulator," pp. 1-7, 2023.

[10] N. Mirchi *et al.*, "Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance," *Operative Neurosurgery*, vol. 19, no. 1, pp. 65-75, 2019. https://doi.org/10.1093/ons/opz359

[11] Y. Munz, "Laparoscopic virtual reality and box trainers: is one superior to the other?," (in eng), *Surgical Endoscopy*, vol. 18, no. 3, p. 485, 2004.

[12] F. J. Carter *et al.*, "Consensus guidelines for validation of virtual reality surgical simulators," *Surgical Endoscopy And Other Interventional Techniques*, vol. 19, no. 12, pp. 1523-1532, 2005/12/01 2005. 10.1007/s00464-005-0384-2

[13] C. Huang, H. Cheng, Y. Burreau, H. M. Ladak, and S. K. Agrawal, "Automated Metrics in a Virtual-Reality Myringotomy Simulator: Development and Construct Validity," (in eng), *Otol Neurotol*, vol. 39, no. 7, 2018. https://doi.org/10.1097/mao.0000000000001867

[14] R. M. Kwasnicki, R. Aggarwal, T. M. Lewis, S. Purkayastha, A. Darzi, and P. A. Paraskeva, "A Comparison of Skill Acquisition and Transfer in Single Incision and Multi-port Laparoscopic Surgery," *Journal of Surgical Education*, vol. 70, no. 2, pp. 172-179, 2013/03/01/ 2013. https://doi.org/10.1016/j.jsurg.2012.10.001

[15] B. Stew, S. S.-T. Kao, N. Dharmawardana, and E. H. Ooi, "A systematic review of validated sinus surgery simulators," vol. 43, no. 3, pp. 812-822, 2018. https://doi.org/10.1111/coa.13052

[16] S. Chawla, S. Devi, P. Calvachi, W. B. Gormley, and R. Rueda-Esteban, "Evaluation of simulation models in neurosurgical training according to face, content, and construct validity: a systematic review," *Acta Neurochirurgica*, vol. 164, no. 4, pp. 947-966, 2022/04/01 2022. 10.1007/s00701-021-05003-x

[17] S. S. Van Nortwick, T. S. Lendvay, A. R. Jensen, A. S. Wright, K. D. Horvath, and S. Kim, "Methodologies for establishing validity in surgical simulation studies," *Surgery*, vol. 147, no. 5, pp. 622-630, 2010/05/01/ 2010. https://doi.org/10.1016/j.surg.2009.10.068

[18] O. Søvik *et al.*, "Virtual reality simulation training in stroke thrombectomy centers with limited patient volume—Simulator performance and patient outcome," p. 15910199231198275, 2023.

[19] J. Y. Chen, J. E. J. I. T. o. S. Thropp, Man,, C.-P. A. Systems, and Humans, "Review of low frame rate effects on human performance," vol. 37, no. 6, pp. 1063-1076, 2007.

[20] S. K. Card, *The psychology of human-computer interaction*. Crc Press, 2018.

3.3 Additional Studies Related to Article 1

Article 1 assessed the impact of visuals and 'feel' on immersion within the developed VR system. A side study was conducted to further investigate the relationship between graphics, computational complexity, and their collective effect on VR realism and user immersion. The goal was to develop an objective metric to identify the minimal threshold that upholds both graphical realism and immersive experience. This study aligns with the overall goals set forth in Objective 1 and Hypothesis 1, which seek to identify and establish the foundational validation for the current simulator platform. Specifically, the side study examines the aspects of VR/AR simulations that influence immersion and at the same time assesses how these aspects affect both face and content validity. The contribution of the first author is 80%, which involved utilizing a subset of the preliminary data collected for Article 1, along with data previously gathered by Tianqi Wang. The first author not only merged this data but also carried out new analyses. Furthermore, the author was responsible for drafting the manuscript that was ultimately submitted to the European Spine Journal. This subset of data formed the basis of the side study, which was a preliminary subset of the broader dataset applied in the main Article 1.

3.4 Article 2 – Assessment of the Fidelity of a Mixed Reality Surgical Spine Simulator using direct comparison of Cadaver and Simulator trials

Authors:

Sami Alkadri, B. Eng., Ph.D Student

Tianqi Wang, M. Eng.

Mark Driscoll, P.Eng., Ph.D., Assistant Professor *

Institutional Affiliation:

Department of Mechanical Engineering

McGill University

Macdonald Engineering Building

815 Sherbrooke St W, Montreal, Quebec H3A 2K7, Canada

Corresponding Author:

Mark Driscoll

Mailing Address: Macdonald Engineering Building, RM 153, 817 Sherbrooke St W, Montreal, Quebec H3A 2K7, Canada

Phone: 514-398-6299

Fax: 514-398-7365

Email Address: mark.driscoll@mcgill.ca

3.4.1 Abstract

Background

Virtual reality surgical simulators offer a cost-effective and ethical training environment. High visual fidelity enhances trainee satisfaction, confidence, and skill transfer. Quantifying visual fidelity can improve finite element physics-driven simulations in balancing high visual realism and computing complexity.

Purpose

(1) Establish visual fidelity through a subjective face validity questionnaire. (2) Define a quantitative approach linking design parameters and subjective visual feedback.

Methods

Under ethical approval, 16 senior surgeons performed a minimally invasive spinal fusion on cadavers and then directly on a simulator containing visual, auditory, and haptic feedback. Surgeons evaluated the simulation's visual fidelity via a questionnaire. The impact of quantifiable model parameters on visual fidelity was established through image comparisons and computational speed evaluations of five varying-complexity models.

Results

Expert surgeons rated visual fidelity of the simulation with a median $\geqslant$ 3 in the questionnaire. A negative linear correlation was found between frame rate per second and number of nodes (r^2=0.925). The number of nodes impacted subjective perception (P-value<<0.05).

Conclusions

Visual fidelity was established by experts' ratings. Increasing the nodes enhanced the visual, physical, and haptic feedback, but decreased the frame rate. Balancing nodes and frame rate optimizes user satisfaction and real-time interaction. The study enabled a quantitative approach to visual fidelity.

Keywords

Surgical Simulator; Face Validity; Visual Fidelity; Visual Graphics; Computational Complexity; Comparative Study.

COMPLIANCE WITH ETHICAL STANDARDS

Ethical Approval

The institutional IRB approval was received for the study protocol and consent forms. This research involved recruiting surgeons to perform the virtual surgery on the simulator. Proper consent was obtained.

Competing Interests Statement

No competing interests to declare.

Funding

NSERC Collaborative Research Development (CRD) Grant.

3.4.2 Introduction

Medical simulation, an innovative method in surgical education, aims to replace traditional training systems currently in use. More specifically, virtual reality surgical simulators allow surgical trainees to develop and refine their skills while eliminating ethical and cost constraints associated with cadavers or animals [1]. In the field of medical simulation, high visual fidelity is becoming an essential target. In simulation, fidelity is described as *"the degree to which a model or simulation reproduces that state and behaviour of a real-world object or perception of a real-world object, feature, condition or standard in a measurable or perceivable manner"*[4]. High-fidelity simulators can lead to student satisfaction, self-confidence, and the ability to transfer the acquired skills to a surgical setting [2, 3]. Most studies refer to fidelity only qualitatively using subjective validity questionnaires while failing to quantify high and low fidelity [5]. Few studies discuss which parameters researchers should consider while developing the visual component of virtual reality simulators. Often, these studies emphasize the importance of realism to trigger the same psychological and physical immersion as would be experienced in surgery, without explicitly discussing the simulation design parameters leveraged to attain the required level of realism. Although the criteria for sufficient fidelity in simulation for educational purposes are based on subjective validation methods, researchers are increasingly pursuing standards of high fidelity, imposing a quantification to the method of achieving fidelity [6]. This is especially advantageous for simulation platforms that simulate soft tissue models using the finite element method, as such models contain a plethora of measurable design variables that can influence fidelity. Therefore, one approach of quantifying fidelity is to establish a relationship between the subjective validity and the measurable simulation parameters related to the finite element models used.

To accurately represent soft tissues in finite element medical simulations, both surface and volumetric models are utilized with varying proportions depending on whether computer efficiency or physical accuracy is most required. On the one hand, surface models are more advantageous when modelling geometrically-identical tissues due to the minimal number of nodes and the low computational costs associated with such models; however, these models often lead to invalid physical deformations [7]. On the other hand, volumetric models are more relevant for delivering both precise data matrices and geometrically accurate representations especially when considering geometric deformations that occur in tissues undergoing cutting and penetration processes in the simulation. Nevertheless, the added computational costs induced by the volumetric models often impose a constraint when designing a high visual fidelity simulation. Fortunately, this design constraint is being addressed with the rapid development of the computational industry, which is demonstrating increased computational power reducing the calculation time and optimizing the graphics for simulation display. In general, both surface and volumetric models are used concurrently to achieve better visual fidelity. As such, model refinement methods to augment the visual feedback provided by both model types are required.

The most general model augmentation methods include: (1) mapping textures on the surface of the objects and (2) adding complexity to the objects. The augmentation methods of both the surface and the volumetric models are associated with similar drawbacks as each respective model type. The latter method adds nodes to the volumetric model for increased force feedback accuracy and realistic graphical representation of the topological changes under usual surgical manipulations, such as deformations, cutting, and penetration [8]. Even though the extra polygons of the model improve the geometrical and physical details, the complexity increases the computational time per frame, and thus decreases the frame rate per second (FPS). When the FPS becomes less than 20 Hz, discontinuous and lagging graphic feedback affects the user experience, which is related to the rate at which the brain processes visual data [9]. In addition to graphical accuracy, the lagging in the frame processing rate also affects the visual and topographical responses of tissues undergoing loading input by the surgeon. Therefore, while the generated volumetric model provides physical accuracy, over-refining such models may hinder the interactions in real time simulation. The ability to provide real-time interaction is important for the educational value of the simulation platform. Real-time simulations allow surgeons to practice the virtual surgical procedure efficiently and effectively, with instant and accurate operational performance evaluations [10]. Therefore, to avoid the downsides of over-augmenting the volumetrics models, texturing techniques – the first augmentation method listed above – are often utilized in conjunction with volumetric model augmentation to further improve visual fidelity. To optimize the simulation fidelity and performance, one must use the right combinations of both volumetric and surface models that generate high fidelity simulators with real time interactions.

The current study utilizes a surgical simulator platform for the training of a novel spinal fusion surgery. The platform utilizes a concrete benchtop with visual display and interactive haptic output to mimic the reality of the surgical environment [2, 11, 12]. The medical simulator focusses on three phases of the spinal surgery: gaining access through the back muscles, removing the intervertebral disc, and inserting a spinal cage. The simulated surgical technique is associated with high risk complications, such as injuring the nerve root, that can gravely impact patient outcomes [13]. To potentially reduce the risk of such iatrogenic injuries, the developed simulator should ensure realism. This includes, amongst other important elements, the development of a platform that conveys accurate anatomic morphology and interactive response. Therefore, the simulator should not only achieve fidelity qualitatively, but also quantitatively by establishing a link between model parameters that effect fidelity and the real-time interactions.

The objectives of the current study are to (1) firstly, establish visual fidelity qualitatively using a subjective face validity questionnaire, and (2) subsequently define and present a quantitative approach that highlights the relationship between the frame rate (FPS), the model augmentation methods, and the subjective visual feedback.

3.4.3 Methods

3.4.3.1 Participants

A total of 16 senior neurosurgeons and orthopedic surgeons participated in two distinct trials. Nine orthopedic surgeons out of the 16 participants were recruited in the first trial, while the rest participated in the second one. Participants of both trials completed a minimally invasive spinal fusion surgery on a cadaver, then immediately repeated the identical procedure on the surgical trainer/simulator. Participants signed informed consent forms approved by the Ethics Board.

3.4.3.2 Simulator Platform

The surgical simulator under consideration is a physics-based simulator of a minimally invasive spine single level fusion. The geometry of the surgical scenes in the simulator are reconstructed from patient specific data. The simulation runs on a high-performance gaming laptop (i7-8750H) with Windows 10 operating system. Similar to reality, the rendered images are displayed on two flat panel monitors to match the interface in the operating room: a built-in monitor and an external touch screen monitor, with 120Hz and 60Hz refresh rate, respectively. The monitor on the left in Figure 3-5 provides general surgical guides including a recorded animation displaying how to operate instruments during a step. The other monitor is an interactive touch screen displaying the laparoscopic views of the surgical area with which the surgeons interact. Haptic feedback is provided from a combination of a six-degrees of freedom ENTACT W3D device and a benchtop model, also exemplified in Figure 3-5. Due to the visual processing constraint, the mesh model containing the element size that produce the 30Hz was used in the training platform during the simulation procedure to provide complete visual and haptic feedback for the participants. The simulation focusses on three phases of the spinal surgery: gaining access through the back muscles, removing the intervertebral disc, and inserting a spinal cage. The main focus of the current study is the access gaining step, which requires the user to access the disc through virtual volumetric models. The virtual volumetric model contains artificial muscle layers and an intervertebral disc, each providing realistic force feedback through interaction with the haptic device. The force feedback replicates the resistance provided by the instruments when penetrating through the muscles during an actual surgery using tailored mechanical properties.

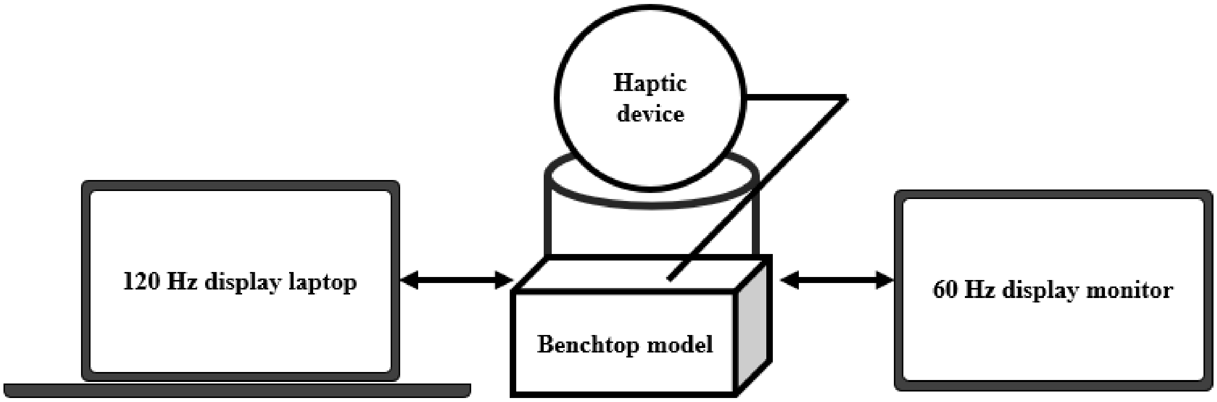


Figure 3-5 The summarized simulator layout. Left is the laptop runs 120Hz display, which indicates the instruction of the surgery process. The haptic device and benchtop model are in the middle. And right is the external display runs 60Hz which indicates the four cameras that demonstrate the surgical area. The surgeon operates the haptic device based on the visual feedback from both monitors

3.4.3.3 Visual Fidelity & Face Validity

The surgeons completed questionnaires using a 5-point Likert scale comparing the developed simulator in terms of visual satisfaction contrasted with the cadaver surgery (Table 3-6). Surgeons recruited in the second trial answered additional questions pertaining to the overall visual face validity of the simulator. For the purpose of this study, only questions pertaining to the visual fidelity were included, the face validity of the entire simulator is assessed in separate studies by our group. All questions were explained to each surgeon and the entire experiment was completed within 1 hour to ensure that the participants contrasted their experience on the virtual procedure to that on the cadaveric surgery in a side-by-side comparison. Validity for the visual face fidelity was deemed sufficient if a median ≥ 3.0 on a 5-point Likert scale was achieved. Furthermore, a one sample sign non-parametric test was utilized to establish statistical significance of validity consensus.

3.4.3.4 Quantifying Fidelity

To determine the relationship between the FPS, the augmentation methods (defined by the number of nodes and the texturing treatments), and visual feedback, two sets of six different simulation runs were conducted by incrementally varying the tetrahedral mesh from 3mm to 8mm, while simultaneously measuring the FPS. The second set of simulation runs measured the affect of texturing treatments on the FPS by using solid colour to replace the diffuse and normal maps, as shown in Figure 3-6. The texturing treatments were based on videos and images collected during cadaver experiments and were implemented using the industrial graphical painting software *Autodesk Maya* (2019, San Rafael, United States) and *Substance Painting* (2019.3.1, San Jose, United States). Lastly, user feedback was obtained using images of the different computational complexities that were recorded previously from the simulator (Questionnaire B - Figure 3-8).

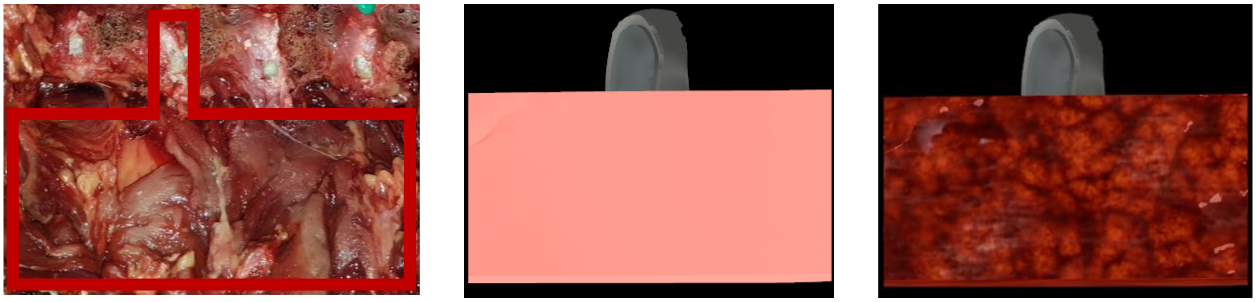


Figure 3-6 Impact of Texturing on FPS - Compares a reference real cadaver image (left) with simulation outputs: one with a basic solid color (middle) to assess minimal texturing impact, and another with full texture treatment (right), illustrating the effect of detailed texturing on FPS and visual realism.

3.4.3.5 Statistical Analysis

The one sample sign non-parametric test was used to establish significance in the questionnaire responses. Normality was assessed using the Shapiro-Wilk test, which demonstrated non-normally distributed data (P <0.05). To complete the required statistical analysis for

Questionnaire B (Figure 3-8), the score of each image was ranked from 5 to 1 based on the number of nodes used, where 5 and 1 were allocated for the models with the most and the least number of nodes, respectively. A score of 3 was allocated when the participants found no difference between the models. The null hypothesis of Questionnaire A was that the median responses of face validity were less than and equal to 3, whereas the null hypothesis of Questionnaire B was that the number of nodes has no impact on visual feedback. A p-value of less than 0.05 was deemed to provide statistical significance.

3.4.4 Results

3.4.4.1 Visual Fidelity & Face Validity

The median scores and ranges for face validity are outlined in Table 3-6 (Questionnaire A). The expert surgeons recruited in both trials rated the visual graphics of the simulation with a median of 4 as compared to the cadaver surgery. The seven participants in the second trial rated the visual guides used in the simulation with a median of 4, and the simulator system setup with a median of 3 as compared to the cadaver surgery, respectively. Participants of both trials rated both the internal and external impressions of the tissue model used in the simulation with a median of 3 as compared to the cadaver surgery, respectively. Furthermore, the median responses of three out of the four statements were found to be significantly greater than a value of 3 using the one sample sign non-parametric test.

Table 3-6 Face Validity Questionnaire A

Face Validity Questionnaire A			
Question No. – Validity Statement	**Median**	**Range**	**P value[Ŧ]**
1. The visual graphics shown in the Port Cam view are similar to the cadaver surgery[+]	4	2 – 5	0[*]
2. The visual guides shown during the simulation are similar to the ones used during the cadaver surgery	4	3 – 5	0[*]
3. The simulator system setup – including the positioning of the screen, the haptic device, and the benchtop model – is similar to the cadaver surgical setup	3	3 – 4	0.0059[*]
4. The internal and external impressions of the tissue model shown in the Port Cam view are similar to reality[+]			
a. Internal Impression	3	2 – 4	0.3438
b. External Impression	3	2 – 4	0.0078[*]

Ŧ P-value for the one sample sign non-parametric test

+ Questions were part of both trials (total of 16 participants)

*** Significant P-value for the one sample sign non-parametric test (P<0.05)**

3.4.4.2 Quantifying Fidelity

1. The relationship between number of nodes and FPS. The frame rate corresponding to each number of nodes were collected during the simulated operation (Table 3-7). A negative and linear correlation was found between the FPS and the number of nodes: $FPS = -0.0004NON + 77.952$ with $r^2 = 0.925$, where NON means number of nodes (Figure 3-7).

Table 3-7 - Model performance in the simulation platform.

Element Size (mm)	Number of Nodes	FPS(Average)	FPS (Solid Colour)
3	139788	21	30
4	111744	28	40
4.5	102216	31	41
5	97896	37	42
6	90426	42	53
8	83622	43	58

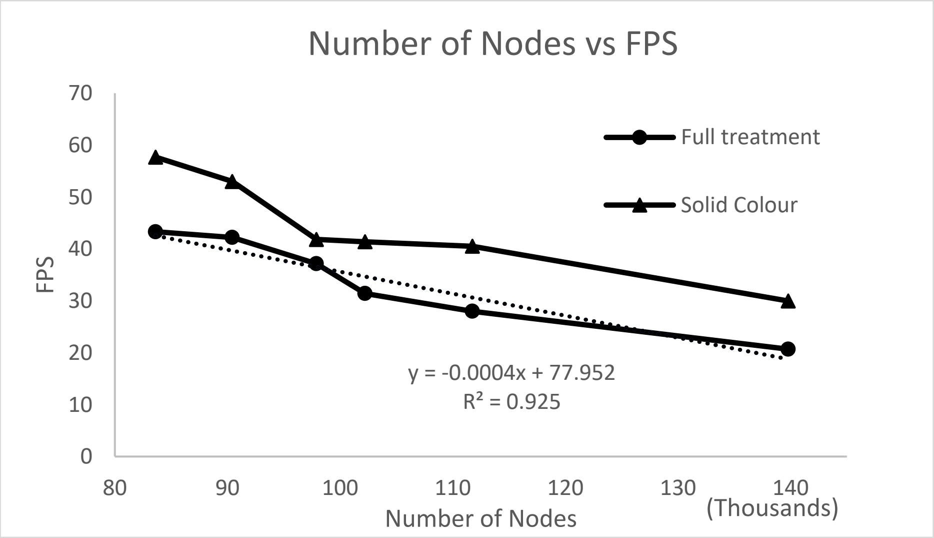


Figure 3-7 - Comparison of average frame rate between different surface treatment with different number of nodes

2. The effects of surface treatment on computational complexity. After replacing the texture produced using diffuse and normal mapping, with a solid colour on the models, an improved and linear frame rate was observed for all sized meshes (Figure 3-7).

3. Evaluation of the visual feedback by senior surgeons For Questionnaire B, five of the surgeon participants reported that either the model with the first or the second most number of nodes best suited their experience in the operating room compared to the cadaver surgery. The remaining four participants indicated no difference among the five element size models. Using the one sample sign non-parametric test, a significant value was obtained, indicating that the mean of the responses was significantly larger than 3 (P-value=0).

Table 3-8 Face Validity Questionnaire B

Face Validity Questionnaire B												
Validity Statement	**Participant No.**									**Mean**	**Range**	**P value**^{Ŧ}
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**			
Which image is best suitable for your experience?	3	3	4	3	4.5	4	5	5	3	3.83	3-5	0*

3.4.5 Discussion

Visual Fidelity and Face Validity

The visual components of the surgical simulator under consideration exhibit visual fidelity and face validity, which are important in establishing student satisfaction, self-confidence, and the ability to transfer the acquired skills to a surgical setting. All statements in Questionnaire A (Table 3-7) attained median values ≥ 3, indicating that visual fidelity was achieved qualitatively. To further support the notion of validity, the one sample sign non-parametric test was conducted. All but one aspect of the visual fidelity attained median values significantly greater than 3; the expert surgeons did not rate the internal impression of the tissue model used in the simulation significantly greater than 3. The internal and external impressions are major visual components of the minimally invasive endoscopic surgical viewpoint, as such they were chosen to examine the visual feedback associated with the actual surgical experience. The internal muscle texture represents the deep muscle tissue under cutting and penetration, while the external texture reflects the superficial area of the muscle that produces the specular effect of biological fluid during surgery. Due to the nature of this simulation platform, the surgeon would spend significantly more time looking at and working with the internal texture. Thus, the internal texture including the cross-section area have a greater impact on visual feedback than the external effects, which might justify the lower rating associated with the internal impression as compared to the external impression of the tissue model.

Quantifying Fidelity

The approach used in this study for fidelity quantification required establishing the relationships that connect each of the FPS, the augmentation methods defined by the number of nodes and the texturing techniques, and the corresponding subjective user feedback. To determine such relationships, two sets of six simulation runs were conducted on the simulator, focusing mainly on the access gaining step, which requires the user to access the disc through virtual volumetric models. More specifically, at each run the volumetric model was augmented by varying the number of nodes of the tissue model. Although the element size is the fundamental parameter to determine the model shape, it is impossible to define a uniform element size for each face. Therefore, the number of nodes was chosen as the augmentation parameter, as it is a better representation for the computational complexity of the model. To account for the time the surgeon spends on the simulator's module, every muscle model is run through the simulation for five minutes to collect the average frame rate. During a simulation run, the frame rate remained stable with little change, allowing for a constant frame rate to be directly linked to a certain number of nodes. The muscle models used in this study employed a texturing technique that uses diffuse and normal mapping methods to maximize the realism for participants and maintain an immersive experience while using the medical simulator. To test the affect of the texturing treatment on the computational time (FPS), a second set of simulation runs were conducted by replacing the texturing treatments with solid colors.

Based on the present study, with respect to fidelity, the number of nodes provided important information. It visually affects the level of realism in displaying the cross-sectional area after surgical tool penetration. Increasing the number of nodes allows the use of smaller tetrahedral mesh, which leads to smoother cross-sectional area with no sharp edges, especially during surface cutting [19]. Based on the results of the surgeons' satisfaction for the images provided in the questionnaire B, a significant portion of the experts preferred the models with the increased number of nodes, supporting the notion that larger number of nodes produces a more realistic muscle cross section. Similarly, in a real-time game engine, adding higher mesh resolution enhances the apparent geometric detail of fracture [16]. Furthermore, increasing the number of nodes results in more accurate physical behavior from the finite element analysis standpoint [15]. On the same note, refining the model mesh could potentially increase the accuracy of haptic fidelity, which is associated with the information collected during model contacts [20]. At the same time, increasing the number of nodes in a model decreases the response time of the computer. For instance, the models with the largest number of nodes (>103,000 nodes) were below the 30 Hz range, which is the minimum frame rate required to prevent losing real-time interaction (Figure 3-7) [14]. Even the current most advanced computers would still fall short in producing a realistic representation of an over-refined volumetrics surgical area while maintaining the interactive experience. The purpose of achieving high degrees of realism and visual fidelity is to maximize the transference of surgical skills, which would not be feasible without real-time feedback [17]. As such, optimizing the number of nodes is required to maximize user satisfaction while maintaining the minimum required frame rate for real time interaction. Using the linear relationship found in this study, an estimate of the generated frame rate can be determined for any given number of nodes, and thus allows for a quantitative approach to reach visual fidelity. Similar studies, such as Ullrich's work, demonstrated the same linear relationship between number of nodes and computational time per frame [21]. Thus, the methodology of this experiment provides a quantifiable value for fidelity, which may be utilized by researchers to validate the visual framework of future surgical simulators.

The texturing treatments also have an impact on the FPS, resulting in an overall decrease in the frame rate with an unusual average of 5 Hz frame rate drop after the number of nodes was increased beyond 97,896 nodes (Figure 3-7). However, the affect of this surface augmentation method is not as strong as volumetrically augmenting the model, allowing for an increased user satisfaction lower computational cost. Nevertheless, further investigation is still required to fully understand how textures influence computational complexity.

3.4.6 Limitations

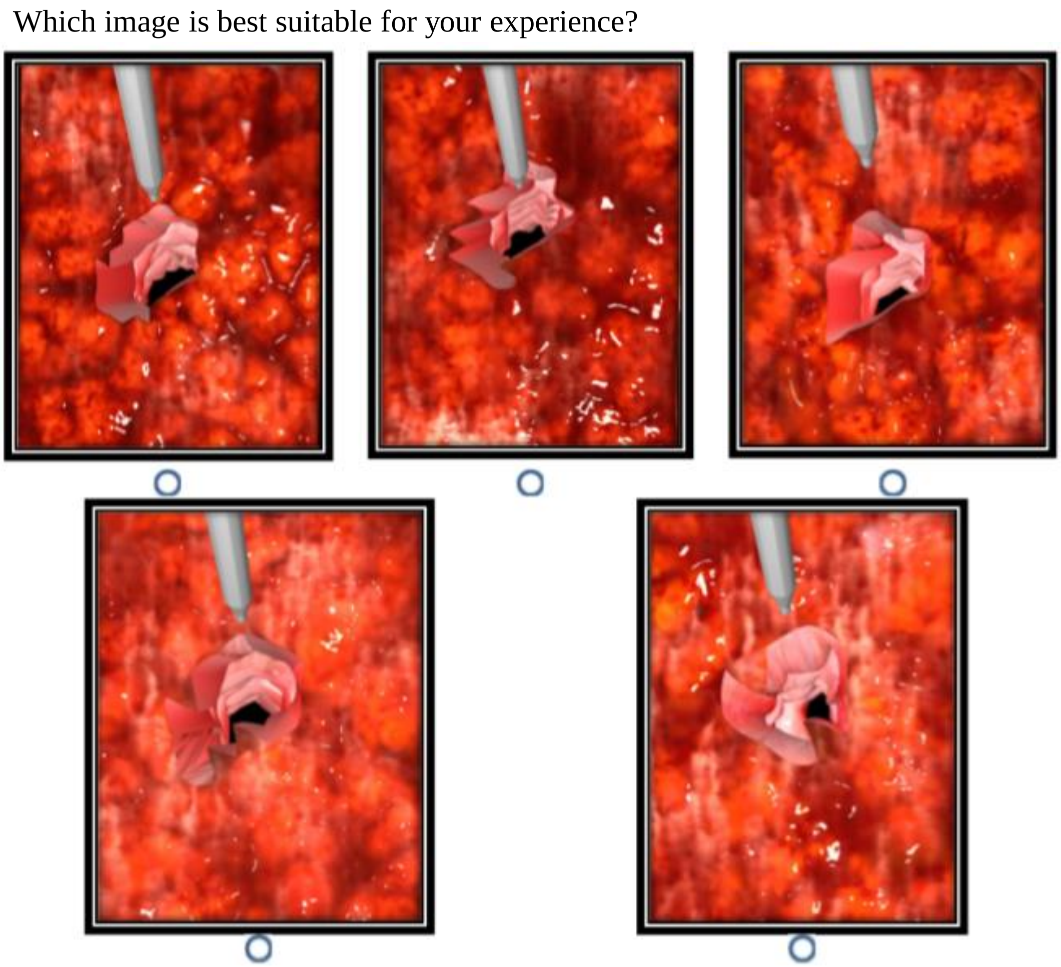
The sample size of the questionnaire is seen as a limitation of the current study as a relatively small sample of surgeons were recruited in both trials as compared to the ideal sample size of 30 participants. Nevertheless, participants of both trials were senior surgeons with broad experience in an operating room, thus providing invaluable insights. Given the unique set-up of this experiment, using an immediate comparison to a cadaver surgery provides further confidence in the reported results. A second limitation is associated with time constraints that resulted in the participants not being able to use the simulator with the different models and frame rates, and instead relied on the images of the different models to acquire the visual feedback from the experts. While similar studies examined the effect of frame rates, it would still be valuable to test the impact of varying the frame rate with senior surgeons in the context of this study and its parameters.

Another limitation is associated with the lack of an established standard for determining fidelity. Describing fidelity in the context of a surgical simulator is still based on physical and psychological subjective perception [2]. It is often vague and differently defined based on the trainer's needs. For example, the simulation platform in this study is considered to have fidelity based on the subjective feedback from the face validity Questionnaire. However, the level fidelity is not readily given from such a validation method. Therefore, it is feasible to consider adding the capability of frame rate production in conjunction with the number of nodes to quantify the fidelity of a simulator. On the one hand, both the frame rate and the number of nodes are critical parameters, which heavily influence the cost of a training platform and enhance the immersion to connect the virtual and real-life experience. On the other hand, they are quantifiable parameters providing a method to determine whether a platform meets the needs of a surgical method such as the head-mounted display or the continuously accurate haptic output [22].

3.4.7 Conclusion

In summary, the relationships between model complexity and simulation frame rate time were examined in this study. Sixteen senior surgeons provided an opinion based on experience compared with the procedure on a cadaver within a short period of time. This unique methodology produced a final compromise when graphical requirement exceeded the computational budget. A new approach to determine simulation fidelity is suggested in this study. Future work will focus on implementing a dynamic platform that allows various number of nodes to be modelled in a desired region to maximize visual performance with a given computational power while investing in optimization.

3.4.8 Appendix A



Which image is best suitable for your experience?

○ ○ ○

○ ○

Figure 3-8 User Feedback on Simulation Fidelity - This figure displays the interface of Questionnaire B, which was used to collect user feedback on the visual fidelity of the simulation across different computational complexities. It showcases various images from the simulator that participants evaluated, highlighting their perceptions of realism and the effectiveness of different mesh complexities in achieving a lifelike surgical simulation experience.

3.4.9 References

[1] A. Frisoli *et al.*, "Simulation of real-time deformable soft tissues for computer assisted surgery," vol. 1, no. 1, pp. 107-113, 2004.

[2] S. Barry Issenberg, W. C. Mcgaghie, E. R. Petrusa, D. Lee Gordon, and R. J. J. M. t. Scalese, "Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review," vol. 27, no. 1, pp. 10-28, 2005.

[3] Z. Samawi, T. Miller, and M. S. J. N. E. P. Haras, "Using high-fidelity simulation and concept mapping to cultivate self-confidence in nursing students," vol. 35, no. 6, pp. 408-409, 2014.

[4] M. Roza, J. Voogd, H. Jense, and P. Van Gool, "Fidelity requirements specification: A process oriented view," in *Fall Simulation Interoperability Workshop*, 1999: Citeseer.

[5] R. T. Hays and M. J. Singer, *Simulation fidelity in training system design: Bridging the gap between reality and training*. Springer Science & Business Media, 2012.

[6] N. J. Maran and R. J. J. M. e. Glavin, "Low ‐ to high ‐ fidelity simulation – a continuum of medical education?," vol. 37, pp. 22-28, 2003.

[7] H. J. P. o. t. I. Delingette, "Toward realistic soft-tissue modeling in medical simulation," vol. 86, no. 3, pp. 512-523, 1998.

[8] H. Zhang, S. Payandeh, and J. J. D.-S. Dill, "Simulation of progressive cutting on surface mesh model," 2002.

[9] J. Y. Chen, J. E. J. I. T. o. S. Thropp, Man,, C.-P. A. Systems, and Humans, "Review of low frame rate effects on human performance," vol. 37, no. 6, pp. 1063-1076, 2007.

[10] S. K. Card, *The psychology of human-computer interaction*. Crc Press, 2018.

[11] M. A. Seropian, K. Brown, J. S. Gavilanes, and B. J. J. o. n. e. Driggers, "Simulation: Not just a manikin," vol. 43, no. 4, pp. 164-169, 2004.

[12] R. P. Cant and S. J. J. J. o. a. n. Cooper, "Simulation ‐ based learning in nurse education: systematic review," vol. 66, no. 1, pp. 3-15, 2010.

[13] M. G. Lykissas *et al.*, "Nerve injury after lateral lumbar interbody fusion: a review of 919 treated levels with identification of risk factors," vol. 14, no. 5, pp. 749-758, 2014.

[14] A. Liu, G. Tharp, L. French, S. Lai, L. J. I. T. o. r. Stark, and Automation, "Some of what one needs to know about using head-mounted displays to improve teleoperator performance," vol. 9, no. 5, pp. 638-648, 1993.

[15] C. Basdogan *et al.*, "Haptics in minimally invasive surgical simulation and training," vol. 24, no. 2, pp. 56-64, 2004.

[16] E. G. Parker and J. F. O'Brien, "Real-time deformation and fracture in a game environment," in *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2009, pp. 165-175.

[17] T. R. J. I. J. o. N. E. S. Kirkman, "High fidelity simulation effectiveness in nursing students' transfer of learning," vol. 10, no. 1, pp. 171-176, 2013.

[18] G. Norman, K. Dore, and L. J. M. e. Grierson, "The minimal relationship between simulation fidelity and transfer of learning," vol. 46, no. 7, pp. 636-647, 2012.

[19] D. Bielser, V. A. Maiwald, and M. H. Gross, "Interactive cuts through 3 ‐ dimensional soft tissue," in *Computer Graphics Forum*, 1999, vol. 18, no. 3, pp. 31-38: Wiley Online Library.

[20] D. Bielser and M. H. Gross, "Interactive simulation of surgical cuts," in *Proceedings the Eighth Pacific Conference on Computer Graphics and Applications*, 2000, pp. 116-442: IEEE.

[21] S. Ullrich, D. Rausch, and T. W. Kuhlen, "Bimanual Haptic Simulator for Medical Training: System Architecture and Performance Measurements," in *EGVE/EuroVR*, 2011, pp. 39-46.

[22] C. J. Luciano, P. P. Banerjee, and S. H. Rizzi, "GPU-based elastic-object deformation for enhancement of existing haptic applications," in *2007 IEEE International Conference on Automation Science and Engineering*, 2007, pp. 146-151: IEEE.

3.5 Conclusion

Article 1 delved into the impact of graphical and operational realism on immersion within the VR/AR surgical simulation. It assessed both the visual (face validity) and the skill (content validity) realism for the virtual and the augmented reality versions of each component of the simulation. It demonstrated the foundational validation of the newly developed simulator. It established that the simulator is realistic, satisfying face validity, and it effectively evaluates the intended competencies, ensuring content validity. Furthermore, the study underscored the simulator's utility in distinguishing surgical skill levels as defined by construct validity, highlighting its potential adoption into surgical assessment and training programs. Article 2 ventured into examining the influence and constraints imposed by current hardware on the immersive experience within VR simulations, as previously introduced in the literature review (Section 1.1.4). This exploration goes beyond the mere assessment of realism to consider the technological underpinnings essential for maintaining user engagement. Specifically, the study delved into the optimization strategies necessary to achieve a balance between achieving lifelike simulation experiences (face validity) and sustaining optimal graphical rendering speeds, measured in frames per second (FPS), to preserve the depth of immersion. This exploration highlighted how the technological infrastructure of VR hardware can either enhance or limit the educational potential of surgical simulations.

One of the main advantages of simulation-based training is the ability to open the avenue for elevating resident training from competency-based to expertise-based levels, a goal highlighted in the literature review Section 1.1.1. Attaining these aspirations is feasible through the use of advanced VR/AR systems that are equipped with sophisticated haptic feedback. This technology enables the creation of unique and innovative surgical performance metrics, similar to those developed for construct validity as detailed in Article 1. These advanced VR/AR systems enable a more nuanced real-time feedback mechanism – initially by benchmarking against expert performance to pinpoint the precise elements of surgical expertise, and subsequently by analyzing trainees' performances to direct improvements where needed. Achieving expertise-based training in surgical skills requires a profound analysis of participants' technical performance as defined by the surgical performance metrics. As such, integrating machine learning algorithms with VR/AR surgical simulations is very advantageous, enabling a more granular deconstruction of surgical performance through the detection of subtle and intricate patterns. The efficacy of this approach in analyzing the novel surgical performance metrics derived as part of the construct validity in Article 1 is explored in depth in the next chapter (Chapter 4).

Chapter 4. Machine Learning Study on the OLLIF Virtual Surgical Performance

4.1 Background of Third & Fourth Articles

Leveraging the innovative surgical performance metrics established in Objective 1, Articles Three and Four explored the application of ANNs for the classification and analysis of surgical performance. Article Three specifically examined the incision task within an anterior cervical discectomy and fusion (ACDF) scenario, utilizing the Sim-Ortho simulator—a virtual reality surgical simulation platform developed by OSSimTech. This article not only contributed to the broader thesis project as a pivotal side study but also introduced and validated a novel adaptation of the Connection Weight Algorithm for assessing feature importance. This exploration was crucial for achieving Objective 2 of the thesis, offering an original methodology for evaluating surgical performance metrics and developing a two-layered ANN model. The trained model in Article Three was subsequently leveraged through transfer learning techniques to achieve Objective 2.

Article Four built upon the tools and insights from Article Three to fulfill Objective 2, presenting the outcomes of these works across both manuscripts. The first author's contribution to these articles is 85%. Article Three, which highlighted the initial phase of this research, was successfully published in the Computers in Biology and Medicine Journal on August 18, 2021 (https://doi.org/10.1016/j.compbiomed.2021.104770). Following this, Article Four, that addressed the completion of Objective 2, was submitted to the same journal for publication in January 2024.

4.2 Article 3: Utilizing a Multilayer Perceptron Artificial Neural Network to Assess a Virtual Reality Surgical Procedure

Sami Alkadri, Nicole Ledwos, Nykan Mirchi, Aiden Reich, Recai Yilmaz, Mark Driscoll, Rolando Del Maestro

Author names:

Sami Alkadri B.Eng., Masters Student [(1)], Nicole Ledwos MSc[(2)], Nykan Mirchi MSc[(2)], Aiden Reich MSc [(2)], Recai Yilmaz MD[(2)], Mark A. Driscoll, PEng., Ph.D., Assistant Professor [(1)], Rolando F. Del Maestro MD, PhD [(2)]

Institutional affiliations:

(1) Musculoskeletal Biomechanics Research Lab, Department of Mechanical Engineering, McGill University, Macdonald Engineering Building, 815 Sherbrooke St W, Montreal, H3A 2K7, QC, Canada.

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

Corresponding author:

Mark Driscoll

Mailing Address: Macdonald Engineering Building, RM 153, 817 Sherbrooke St W, Montreal, Quebec H3A 2K7, Canada

Phone: 514-398-6299

Fax: 514-398-7365

Email Address: mark.driscoll@mcgill.ca

ABSTRACT

Background

Virtual reality surgical simulators are a safe and efficient technology for the assessment and training of surgical skills. Simulators allow trainees to improve specific surgical techniques in risk-free environments. Recently, machine learning has been coupled to simulators to classify performance. However, most studies fail to extract meaningful observations behind the classifications and the impact of specific surgical metrics on the performance. One benefit from integrating machine learning algorithms, such as Artificial Neural Networks, to simulators is the ability to extract novel insights into the composites of the surgical performance that differentiate levels of expertise.

Objective

This study aims to demonstrate the benefits of artificial neural network algorithms in assessing and analyzing virtual surgical performances. This study applies the algorithm on a virtual reality simulated annulus incision task during an anterior cervical discectomy and fusion scenario.

Design

An artificial neural network algorithm was developed and integrated. Participants performed the simulated surgical procedure on the Sim-Ortho simulator. Data extracted from the annulus incision task were extracted to generate 157 surgical performance metrics that spanned three categories (motion, safety, and efficiency).

Setting

Musculoskeletal Biomechanics Research Lab; Neurosurgical Simulation and Artificial Intelligence Learning Centre, McGill University, Montreal, Canada.

Participants

Twenty-three participants were recruited and divided into 3 groups: 11 post-residents, 5 senior and 7 junior residents.

Results

An artificial neural network model was trained on nine selected surgical metrics, spanning all three categories and achieved 80% testing accuracy.

Conclusions

This study outlines the benefits of integrating artificial neural networks to virtual reality surgical simulators in understanding composites of expertise performance.

Keywords

Multilayered artificial neural network, feature importance, virtual reality, surgical simulation, surgical education, performance metric, surgical expertise, anterior cervical discectomy and fusion

Funding

NSERC Collaborative Research Development (CRD) Grant, the AO Foundation, Davos, Switzerland, the Di Giovanni Foundation, the Montreal Neurological Institute and Hospital, and the Department of Orthopaedic Surgery at McGill University.

Conflict of interest statement

No competing interests to declare.

4.2.1 Introduction

Virtual reality surgical simulators have been rapidly adopted as a more objective method of training and evaluating surgical technical skills [1, 2]. The incorporation of haptic technology has resulted in increased positive learning outcomes [3]. The range of difficulty associated with spinal surgery has led to the development of novel spinal virtual reality (VR) simulators with haptic feedback [4, 5]. These simulator platforms can deconstruct complex common surgical procedures such as the anterior cervical discectomy and fusion (ACDF) into discreet steps allowing trainees to concentrate on specific technical skills in need of enhancement rather than those already acquired [5]. The ACDF requires learners to master a broad spectrum of surgical techniques and each of these components can be assessed and trained utilizing virtual reality simulators [5, 6].

Virtual reality simulators collect enormous sets of data pertaining to the psychomotor interactions of the user during the completion of the simulated tasks. Such data are often transformed into performance metrics that play an important role in assessing and training surgical trainees. Several studies have established the value of performance metrics in classifying individuals into the correct level of expertise and training individuals to improve their level of performance [6-11].

Artificial intelligence (AI) algorithms employing the vast data sets available from surgical simulators have been able to classify surgical expertise with greater granularity and precision than has been previously demonstrated in surgery [12]. These algorithms have also provided insights into the composites of surgical performance that differentiate levels of expertise [6, 10, 12]. Artificial intelligence can be described as the ability of computational algorithms to make "smart" decisions [13]. Machine learning, a subset of AI, is a term used to describe the ability of algorithms to make classifications or decisions by identifying and learning from hidden patterns within datasets, without the need for explicit instructions [14]. Machine learning includes both simple linear algorithms and more complex non-linear ones [14]. Deeper subsets of machine learning, such as artificial neural networks (ANNs), can correctly learn complex non-linear patterns within the given dataset. ANNs consist of a series of layers containing nodes or neurons. The layers are interconnected via the nodes that pass information through connections with different weights [14]. The algorithm adaptively learns the weights associated with connections between nodes in different layers to generate a better representation of the true model. When combined to virtual reality surgical simulators, the algorithm not only has the potential to increase the granularity of classification of surgical performance, but can also provide deeper insights into the impact of the different performance metrics on the classifications [14]. Most studies utilizing artificial intelligence with surgical simulators only exploit the ability of the algorithms to classify participants, while failing to account for the underlying reasons for the classifications or to quantify the relative importance of the performance metrics used in developing the model [14]. Nevertheless, recent studies applied one-layered ANN combined with the Connection Weights Algorithm to highlight the relative feature importance in classifying surgical performance [13, 15-17]. The Connection Weights Algorithm, originally developed by Olden and Jackson [17], was used to understand and quantify the relative impact of each metric on the classification task in one-layered ANN. To the best of the authors' knowledge, no prior studies implemented this algorithm on multilayered ANN.

Thus, the objective of the study was to assess the ability of a multilayered ANN algorithm to: 1) classify surgical performance on an ACDF virtual reality simulated scenario and, 2) identify the relative importance of specific performance metrics in the surgical expertise classification in this virtual reality spinal procedure. In addition to establishing the effectiveness of an ANN algorithm in distinguishing surgical performance, the novelty explored in this study seek to validate a new adaptation of the Connection Weights Algorithm on a multilayered ANN to assess feature importance.

4.2.2 Material and Methods

4.2.2.1 The Virtual Reality Simulator & The Simulated Scenario

This study utilized the Sim-Ortho VR simulator developed by OSSimTech^{TM} (Montreal, Canada) and the AO Foundation (Davos, Switzerland). The scenario simulated is the ACDF surgical procedure. The VR simulator exploits the use of 3D glasses and graphics from a gaming system to provide 3D visuals of the procedure [5, 6]. This platform immerses individuals in an active and dynamic learning process providing instrument haptic and auditory feedback.

The ACDF simulated scenario utilized in this study has been extensively employed by our group to assess surgical expertise. The simulation includes 3 animated steps (neck incision, placement of retractors, and fusion) and 4 deconstructed interactive steps (C4-C5 vertebral disc annulus incision, discectomy, osteophyte removal, and posterior longitudinal ligament removal) [5, 6, 18]. Each of the interactive simulated steps have been shown to have face, content and construct validity [5]. Prior to the start of the simulation, participants were made aware of all steps and instruments needed to complete the procedure via verbal and written instructions. No time limit was imposed on completing the simulated scenario. The current study focuses on the first interactive step which consists of performing a 2cm transverse box incision exposing the disc annulus using a virtual No.15 scalpel. The second interactive step, discectomy, has been assessed by Mirchi, et al. [6] and the third interactive step, osteophyte removal by Reich, et al. [18] have been previously reported.

4.2.2.2 Participants

This study utilized participant data previously collected in a prior ACDF simulated scenario validation study [5, 6]. Twenty-seven participants were initially recruited to perform the virtual reality ACDF scenario. Since the simulator is optimized for right-handed individuals, data from left-handed participants were excluded. In the previous studies, data from post-residents with non-spine focused clinical practices were excluded. However, since the first interactive step, C4-C5 vertebral disc annulus scalpel incision was not dependent on the more complex remaining interactive steps it was considered appropriate to include data from the post-resident participants. Table 4-1 presents the demographics of the 23 participants. The participants were divided into three groups: A Post-Resident group (3 neurosurgeons, 2 spine surgeons, 5 spine fellows, and 1 neurosurgical fellow), a Senior-Resident group (3 PGY 4-6 neurosurgery and 2 PGY 4-5 orthopaedics residents), and a Junior-Resident group (3 PGY 1-3 neurosurgery and 4 PGY 1-3 orthopaedics residents). Table 4-2 highlights the main differences between the groups based on previous experience, knowledge and comfort levels performing and/or assisting in an ACDF. The senior-resident group (PGY 4 and higher) assisted in more ACDF surgeries and have a higher level of comfort assisting and performing an ACDF solo than the junior-resident group (PGY 1-3). The post-resident group ratings demonstrated expert textbook and surgical ACDF knowledge (median 5.0; range 4.0 – 5.0). This study was approved by an appropriate Research Ethics Board. All participants signed an approved written consent form prior to completing the simulation.

Table 4-1 Demographics of the post-resident, senior-resident, and junior-resident groups.

	Junior Residents	**Senior Residents**	**Post-Residents**
No. of individuals	7	5	11
Age (years) ± SD	27.4 ± 1.4	30.6 ± 2.3	44.2 ± 13.2
Sex			
Male	5	4	11
Female	2	1	0

Surgical Specialty Level of Training	Neurosurgery	Orthopaedic Surgery
PGY 1-3	3	4
PGY 4-6	3	2

Fellows	1	5
Consultants	3	2

Table 4-2 Differences in previous experience, knowledge, and comfort level of the groups.

	Junior Residents	**Senior Residents**	**Post-Residents**
No. of individuals in each group who:			
Have previous experience using a surgical simulator	5 (71%)	4 (80%)	9 (82%)
Assisted on an ACDF in the last month	1 (14%)	3 (60%)	N/A
Performed an ACDF solo in the last month	1 (14%)	1 (20%)	8 (72%)
Medina self-rating on 5-point Likert scale:			
Textbook Knowledge of an ACDF	3.0 (1.0 – 4.0)	3.0 (2.0 – 4.0)	5.0 (4.0 – 5.0)
Surgical Knowledge of an ACDF	3.0 (1.0 – 3.0)	3.0 (3.0 – 4.0)	5.0 (4.0 – 5.0)
Comfort level performing an ACDF with a consultant in the room	3.0 (1.0 – 4.0)	3.0 (2.0 – 5.0)	N/A
Comfort level performing an ACDF solo	1.0 (1.0 – 3.0)	3.0 (2.0 – 4.0)	5.0 (3.0 – 5.0)

4.2.2.3 AI Analysis

A systematic approach was used in integrating an ANN in classifying the virtual surgical performance. As illustrated in Figure 4-1, the methodology was divided into two main steps: Data collection & Preprocessing and Machine Learning Model Development.

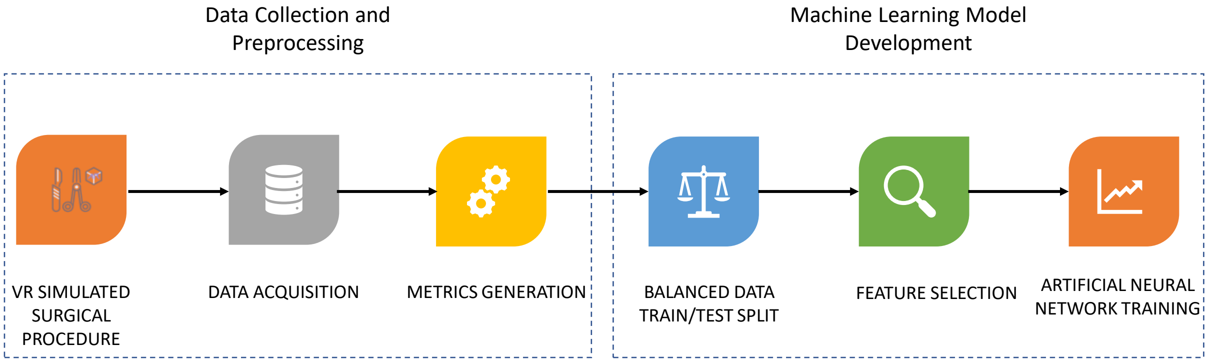

Data Collection and Preprocessing
Machine Learning Model Development
VR SIMULATED SURGICAL PROCEDURE
DATA ACQUISITION
METRICS GENERATION
BALANCED DATA TRAIN/TEST SPLIT
FEATURE SELECTION
ARTIFICIAL NEURAL NETWORK TRAINING

Figure 4-1 The study methodology consisted of two main steps: Data Collection & Preprocessing and Machine Learning Model Development

4.2.2.3.1 Data Collection and Preprocessing

During a simulation procedure, the surgical simulator recorded a series of data relating to the participants' use of the surgical tools. The collected data included variables such as position, time, and angles of the simulated surgical tools, as well as applied forces, removed volumes, and surgical tool contacts of specific anatomical structures. In total 66 variables were collected throughout a simulation run. Subsequently, the recorded data were extracted and processed to generate surgical performance metrics that were used as a set of criteria to assess the performance of the participants in the virtual procedure. For example, position and time were combined to generate velocity metrics, forces and contact detection were used to determine the forces used when removing anatomical structures, and position and contact detection were used to determine the path length used while interacting with anatomical structures. A total of 157 metrics were initially generated based on expert opinion, publications that focused on surgical incision performance, and novel metrics derived from the data [19, 20]. Subsequently, all derived metrics data were normalized using z-score normalization. The generated metrics were assigned into one of three main categories: motion, safety, and efficiency. Data extraction, metrics generation and z-score normalization were done in Python (Version 3.7, OR USA).

4.2.2.3.2 Machine Learning Model Development

Building any machine learning model requires a series of steps to ensure the development of an optimal and a generalizable model. As described by Figure 4-1, three main steps were taken during the machine learning model development. At the very start, the data analyzed was split into training, validation, and testing sets. Since the dataset in this study contained underrepresented classes, a stratified split was used to ensure similar representation of all classes in all sets (Table 4-3). To prevent leakage of information from the testing set into the model development, all subsequent steps – feature selection and model training – were only performed on the training and validation sets, which comprised approximately 78% of the total dataset. Following the split, a z-score normalization was applied on the features. The normalization transformed the mean of each feature to a value of zero and mapped the rest of the values to be centered about the mean, assigning positive and negative z-scores for feature values above and below the mean, respectively.

Table 4-3 Stratified split of the dataset into training, validation, and testing sets.

Classes	Original Dataset	Training Dataset	Validation Dataset	Testing Dataset
Junior	7	4	1	2
Senior	5	3	1	1
Post	11	7	2	2
Total	23	14	4	5

Feeding a large number of unimportant features into any machine learning algorithm would introduce noise and inefficiencies [15]. Hence, following the data split and before training the machine learning model, a sequential forward selection (SFS) algorithm was used to remove irrelevant metrics that may not be useful in distinguishing surgical performance. The SFS algorithm employs its own built-in machine learning model to determine the optimal subset of features. Starting from an empty feature subset, the SFS algorithm iteratively builds optimal feature subsets based on the performance of the built-in machine learning model on the feature subsets. More specifically, at each iteration the SFS algorithm checks the relative performance of the new subset of features as compared to the previous iteration. The algorithm continues until all the features are added, and subsequently returns the optimal subset with the best performance. This study employed a 4-fold cross validation Neural Network model as part of the SFS algorithms for feature selection. The feature selection step reduced the features into nine final metrics as shown in Table 4-4.

Table 4-4 Nine final metrics resulted from the SFS algorithm used in this study. The metrics spanned all three categories.

Metric Category	Metric Description	Metric Abbreviation
Motion	Maximum velocity in the Z direction	$v_{z_{max}}$
	Mean velocity in the Y direction while contacting the Nucleus	$v_{y_{N_{mean}}}$
Safety	Maximum force exerted on the Spinal Cord Nerves	$F_{max_{SCN}}$
	Maximum force exerted on the Right Vertebral Artery	$F_{max_{RVA}}$
	Volume removed of the Spinal Cord Nerves	$VolumeRemoved_{SCN}$
Efficiency	Contact time with the C4 Vertebra	$ContactTime_{C4}$
	Contact time with the Left Posterior Longitudinal Ligament	$ContactTime_{Left_{PLL}}$
	Contact time with the Right Posterior Longitudinal Ligament	$ContactTime_{Right_{PLL}}$
	Contact Length with the C4 Vertebra	$ContactLength_{C4}$

4.2.2.3.3 Building and Training the ANN

Following the feature selection step, a multilayer perceptron (MLP) artificial neural network was built and trained. A PyTorch framework was used to build and train the MLP model. The framework used was similar to a general framework as described by Paszke, et al. [21] and demonstrated by Chintala [22]. The cross-entropy loss was used along with the stochastic gradient descent optimization with momentum algorithm (SGD with momentum) for model training. The ReLu activation function was used with the default Lecun weights initialization technique as defined by the PyTorch built-in functions. To prevent overfitting the model on the training set, early stopping was implemented using the loss obtained on the validation set as a stopping criterion. More specifically, training was stopped once the validation loss increased. The training algorithm built in this study saves a copy of the model parameters when the validation loss is improved. It also saves a history of the training and validation accuracies and loss function value during training.

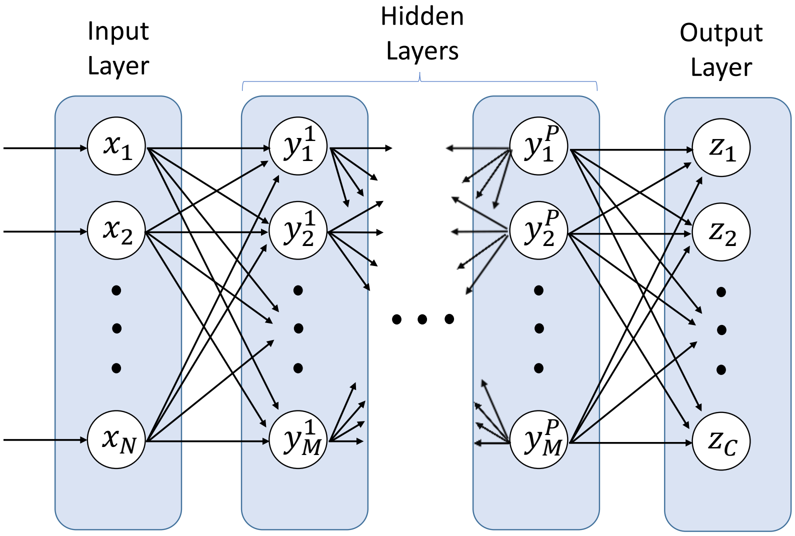


Figure 4-2 A general MLP diagram showing the input layer, the hidden layers and the interconnected hidden units, and the output layer.

An MLP architecture consists of multiple interconnected hidden neurons within multiple layers as presented in Figure 4-2. MLP optimization requires the tuning of several hyperparameters related to both model architecture and training. Model architecture hyperparameters include: the number of hidden layers and the number of hidden units. Model training hyperparameters for the MLP used in the current study (MLP with SGD) include: the learning rate and the momentum of the SGD algorithm. Table 4-5 presents a non-exhaustive list of potential values of each hyperparameter. These values were chosen based on best practices seen in literature when using the SGD learning with momentum algorithm in a multilayer perceptron neural network [23]. A semi-systematic grid search was conducted to explore the models that can be generated using the many different combinations of the presented hyperparameters. The purpose of the grid search was to find the best performing models out of the combinations. Similar to the early stopping, the performance of the models on the validation set was used as a search criterion.

Table 4-5 Hyperparameters potential values.

No. of Hidden Layers	1	2	3		
No. of Hidden Units	6	10	20	40	100
Learning Rate	0.0001	0.0005	0.001	0.005	0.01
Momentum	0.6	0.7	0.8	0.9	1

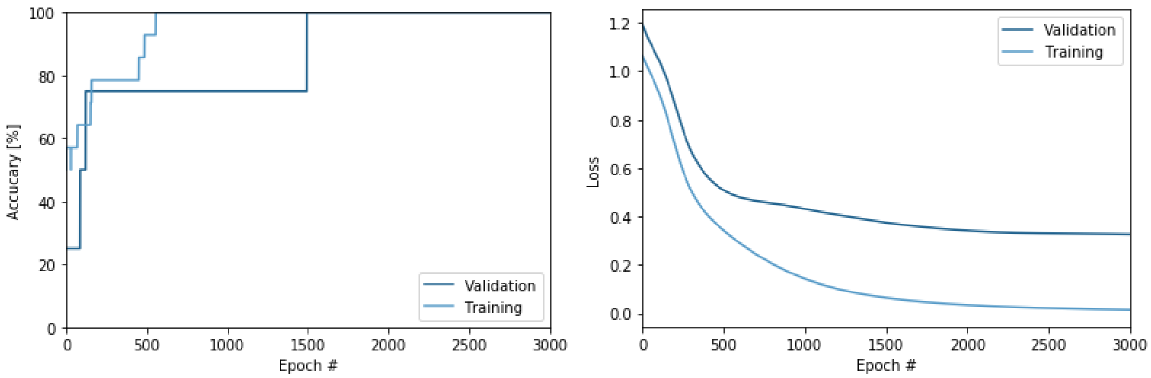
Table 4-6 presents the best performing models found based on the search criteria in the one-layered, two-layered, and three-layered ANNs. As seen in Table 4-6, the two-layered network resulted in a better model performance on the validation set. Table 4-7 shows the chosen model with the best hyperparameters. Figure 4-3 presents the training of the optimal model. After each training epoch, the model was tested on the validation set, generating the validation accuracy and loss. Early stopping was frequently used in training the models, the optimal model stopped training after 3000 epochs as the validation loss started to slightly increase (Figure 4-3).

Table 4-6 The best performing models in each of the one-layered, two-layered, and three-layered ANNs.

Hidden Inputs Per Layer	Hidden Layers	SGD Learning Rate	SGD Momentum	Validation Accuracy	Validation Loss
20	1	0.001	0.8	75%	0.56
40	2	0.001	0.7	100%	0.33
20	3	0.0001	0.8	75%	0.4

Table 4-7 Best performing model found within the grid search.

Hidden Inputs Per Layer	Hidden Layers	SGD Learning Rate	SGD Momentum
40	2	0.001	0.7



(a) (b)

Figure 4-3 The performance of the chosen optimal model at each training epoch: (a) the accuracy of the model on the training and validation sets at each training epoch; (b) the value of the loss function on the training and validation sets at each training epoch.

The Connection Weights Algorithm, originally developed by Olden and Jackson [17], was used to understand and quantify the relative impact of each metric on the classification task. The algorithm was developed for one-hidden layer networks and assigns a distinct weight for each feature-class pair by summing the products of all the connection weights that relate an input to an output, as demonstrated by Figure 4-4 and Equation (18).

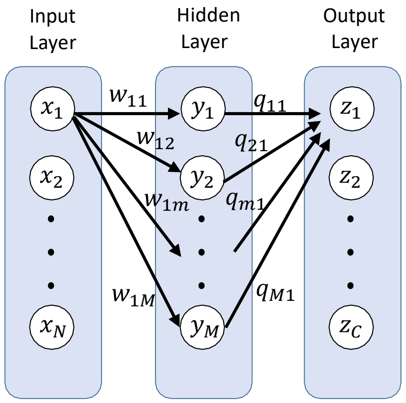


Figure 4-4 Schematic of a one hidden layer network demonstrating the weights that connect the first input node to the first output node.

$$CWP_{x,z} = \sum_{m=1}^{M} w_{xm} q_{mz} \qquad \text{Equation (18)}$$

In this work, the Algorithm was adapted to a multilayer neural network to calculate the Connection Weights Product (CWP) as recently suggested by multiple studies [24, 25]. More specifically, this study adapted the algorithm to a two hidden layer network as demonstrated by Figure 4-5 and Equation (19):

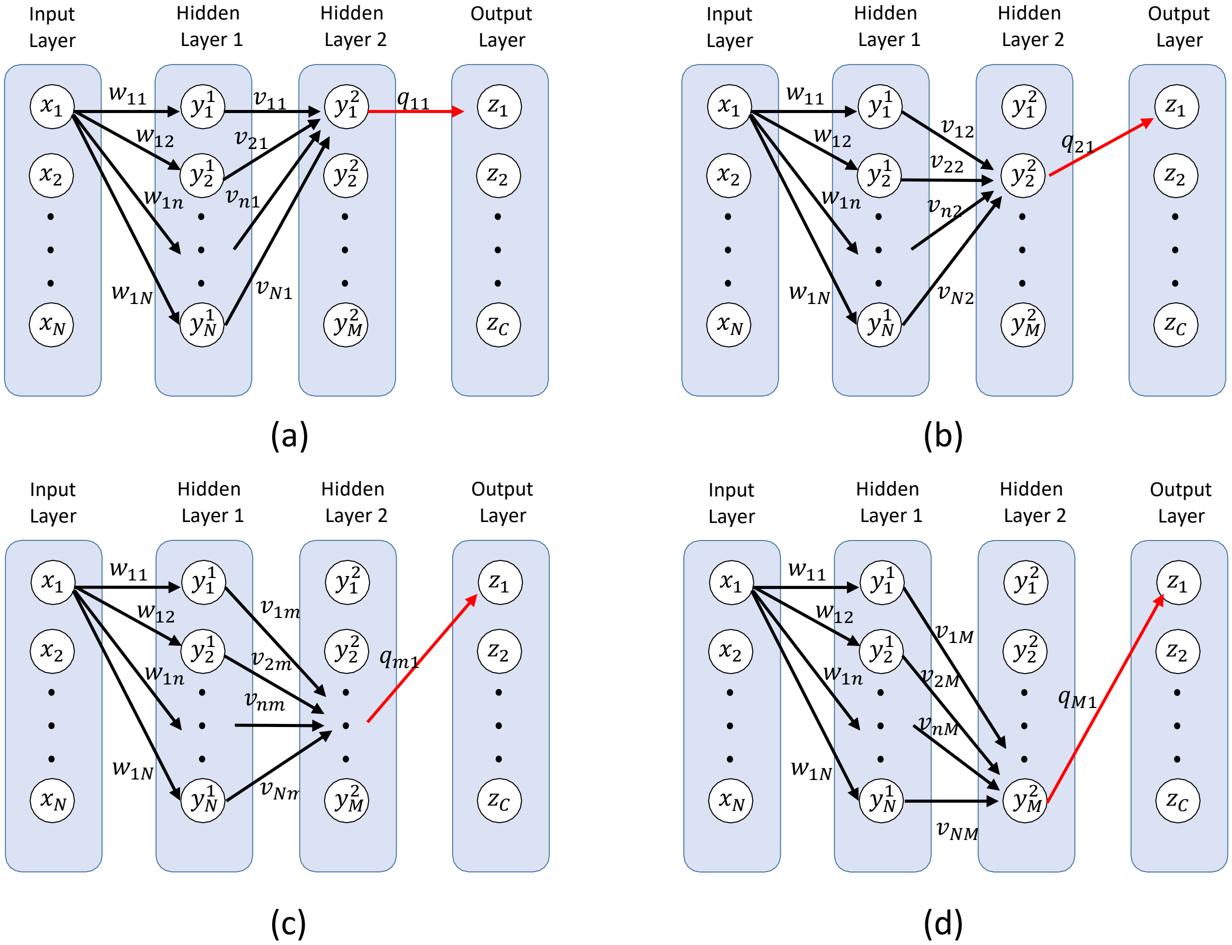


Figure 4-5 Schematic of a two hidden layer network demonstrating the weights that connect the first input node to the first output node. To simplify the illustration, the connection weights are broken into multiple schematics (a-d) by varying the last hidden layer m from 1 to M.

$$CWP_{x,z} = \sum_{m=1}^{M} \sum_{n=1}^{N} w_{xn} v_{nm} q_{mz} \qquad \text{Equation (19)}$$

Where $CWP_{x,z}$ is the connection weight product of an input metric x to a class output z, w_{xn} is the weight connecting an input metric x to a first hidden layer neuron n, v_{nm} is the weight connecting a first hidden layer neuron n to a second hidden layer neuron m, and q_{mz} is the weight connecting a second hidden neuron m to an output z. As demonstrated in Figure 4-5 and Equation

2, the new adaptation of the algorithm can be seen as computing and subsequently adding the original algorithm M times. As with the original algorithm, the CWP can attain both positive and negative values, outlining the relative contribution of each input feature to each output in both magnitude and sign. The sign of the CWP indicates whether a high or a low feature value results in a higher probability of a certain class. CWPs can be further leveraged to obtain the relative importance of the features to each class by determining the ratio of the magnitude of a feature CWP to the sum of the magnitudes of all the features CWPs for that certain class.

To further support the new adaptation of the Connection Weights Algorithm on a multilayer neural network performed in this study, feature importance was also evaluated using the permutation feature importance method and subsequently compared to the results of the Connection Weights Algorithm. The permutation feature importance algorithm captures the importance of a feature by measuring the change in the model score after permuting that feature's values [26, 27]. The loss function along with the prediction accuracy were used in this study as a measure of the model's performance. A feature is important if the model behaves poorly following the permutation of that feature's values, whereas an unimportant feature would not cause the performance of the model to deteriorate significantly. This study used both the training and testing sets when implementing the permutation feature importance. In a sense, the permutation feature importance is similar to a sensitivity study used in a typical finite element analysis.

4.2.3 Results

4.2.3.1 Surgical Performance Metrics

Surgical performance metrics generated for the incision component were divided into three categories: motion, safety, and efficiency. Initially, 157 surgical performance metrics were generated for each participant. Following the SFS (sequential forward selection) algorithm, only nine important metrics remained, as demonstrated in Table 4-4. Similar to the data from the discectomy but unlike the osteophyte removal study, the nine most significant metrics spanned all three categories [6, 18]. These nine surgical performance metrics were used as inputs to the developed ANN. More specifically, the trained model had the following architecture:

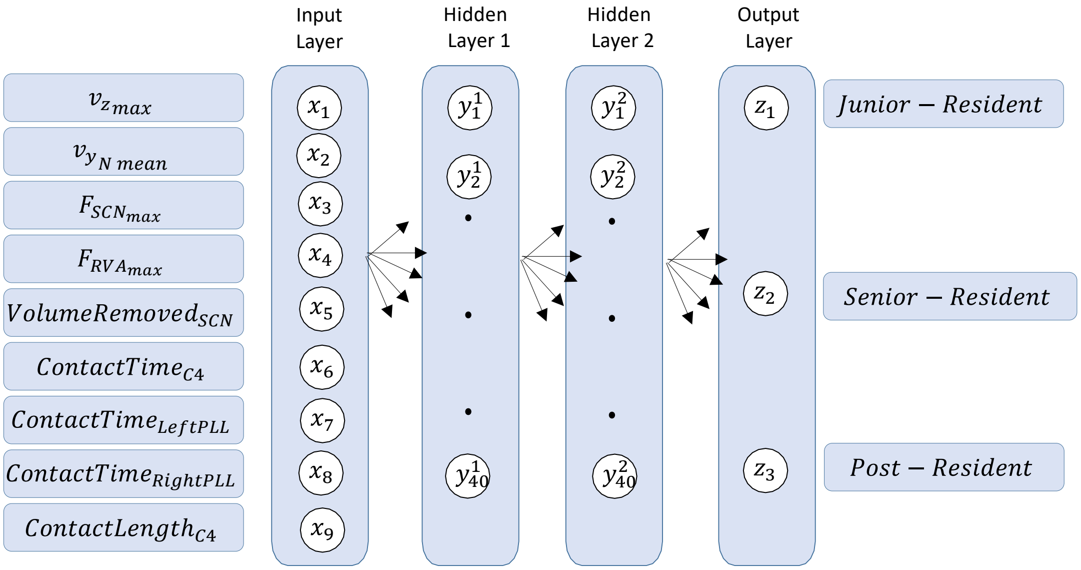


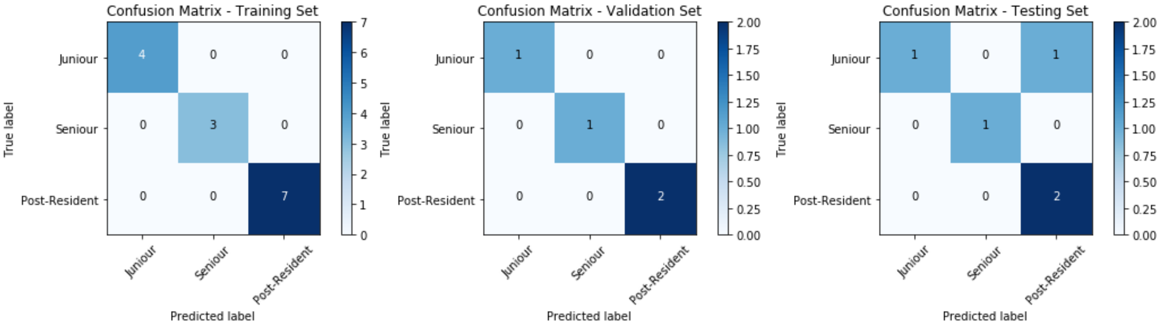
Figure 4-6 Model architecture of the final developed ANN model demonstrating the input surgical metrics, the number of hidden units and layers, as well as the output variables.

4.2.3.2 Accuracy in Classification of Surgical Performance

The final model was trained for 3000 epochs. The classification accuracies of the trained model are highlighted in Table 4-8 and confusion matrices (Figure 4-7 (a) to (c)). A confusion matrix is a table that allows the visual analysis of the performance of an ANN. Three confusion matrices were generated – on the training (14 participants), validation (4 participants), and testing sets (5 participants) – achieving accuracies of 100%, 100%, and 80% respectively.

Table 4-8 Accuracy performance of the trained model on the training set, validation set, and testing set.

No. of Training Epochs	Training Accuracy (%)	Validation Accuracy (%)	Testing Accuracy (%)
3000	**100**	**100**	**80**



(a) (b) (c)

Figure 4-7 Confusion matrices highlighting the performance of the trained model on the: (a) training set, (b) validation set, and (c) testing set.

4.2.3.3 Surgical Performance Metrics Importance

Each input feature within an ANN has a certain impact on the response output of the algorithm. This study adapted the Connection Weights Algorithm to a multilayered ANN and subsequently compared the results to the permutation feature importance method. Table 4-9, Table 4-10, and Table 4-11 present the nine surgical performance metrics along with their CWPs and the corresponding relative importance for the post-resident, senior-resident and junior-resident groups. It is to be noted that the order of feature importance, presented by the relative importance column in the tables, varies for each class of surgical level. Table 4-12 and Table 4-13 present the permutation feature importance applied to the training and testing sets, respectively. Figure 4-8 presents the learning patterns that are exhibited in each input feature. The figure presents the CWPs of each feature for the three surgical levels.

Table 4-9 Ranked Surgical Performance Metrics with Corresponding Weights and Relative Importance for Post-Residents.

Rank	Category	Metric	Connection Weights Product	Relative Importance (%)
1	Efficiency	$ContactLength_{C4}$	8.8201	23.91%
2	Efficiency	$ContactTime_{C4}$	6.9817	18.93%
3	Motion	$v_{z_{max}}$	-6.1178	16.59%
4	Motion	$v_{y_{N_{mean}}}$	-5.8321	15.81%
5	Safety	$F_{max_{SCN}}$	-2.2951	6.22%
6	Safety	$VolumeRemoved_{SCN}$	-2.2766	6.17%
7	Efficiency	$ContactTime_{PLL_{Right}}$	-2.1945	5.95%
8	Efficiency	$ContactTime_{PLL_{Left}}$	-1.3218	3.58%
9	Safety	$F_{max_{RVA}}$	-1.0443	2.83%

Table 4-10 Ranked Surgical Performance Metrics with Corresponding Weights and Relative Importance for Senior-Residents.

Rank	Category	Metric	Connection Weights Product	Relative Importance (%)
1	Motion	$v_{y_{N_{mean}}}$	4.8357	30.75%
2	Efficiency	$ContactLength_{C4}$	3.8694	24.61%
3	Motion	$v_{z_{max}}$	3.3675	21.41%
4	Safety	$F_{max_{SCN}}$	-1.6055	10.21%
5	Efficiency	$ContactTime_{PLL_{Left}}$	-1.0675	6.79%
6	Safety	$F_{max_{RVA}}$	0.3224	2.05%
7	Efficiency	$ContactTime_{PLL_{Right}}$	0.3095	1.97%
8	Efficiency	$ContactTime_{C4}$	-0.2959	1.88%
9	Safety	$VolumeRemoved_{SCN}$	0.0525	0.33%

Table 4-11 Ranked Surgical Performance Metrics with Corresponding Weights and Relative Importance for Junior-Residents.

Rank	Category	Metric	Connection Weights Product	Relative Importance (%)
1	Efficiency	$ContactLength_{C4}$	-12.3433	36.47%
2	Efficiency	$ContactTime_{C4}$	-6.4255	18.99%
3	Safety	$F_{max_{SCN}}$	3.7846	11.18%
4	Motion	$v_{z_{max}}$	3.0317	8.96%
5	Efficiency	$ContactTime_{PLL_{Left}}$	2.2582	6.67%
6	Safety	$VolumeRemoved_{SCN}$	2.1596	6.38%
7	Efficiency	$ContactTime_{PLL_{Right}}$	1.8638	5.51%
8	Motion	$v_{y_{N_{mean}}}$	1.1712	3.46%
9	Safety	$F_{max_{RVA}}$	0.8065	2.38%

Table 4-12 Permutation Feature Importance on the training set.

Rank	Category	Metric	Difference in Loss function	Prediction Accuracy(%)
1	Efficiency	$ContactLength_{C4}$	5.08	40.07%
2	Motion	$v_{z_{max}}$	3.28	63.91%
3	Efficiency	$ContactTime_{C4}$	2.32	56.27%
4	Efficiency	$ContactTime_{PLL_{Right}}$	1.58	78.57%
5	Efficiency	$ContactTime_{PLL_{Left}}$	1.58	78.57%
6	Safety	$F_{max_{SCN}}$	1.51	71.43%
7	Safety	$F_{max_{RVA}}$	1.51	71.43%
8	Safety	$VolumeRemoved_{SCN}$	1.51	71.43%
9	Motion	$v_{y_{N_{mean}}}$	1.21	84.13%

Table 4-13 Permutation Feature Importance on the testing set

Rank	Category	Metric	Difference in Loss function	Prediction Accuracy(%)
1	Efficiency	$ContactLength_{C4}$	4.37	15.62%
2	Efficiency	$ContactTime_{PLL_{Right}}$	2.58	20%
3	Efficiency	$ContactTime_{PLL_{Left}}$	2.53	20%
4	Efficiency	$ContactTime_{C4}$	2.10	52.32%
5	Safety	$VolumeRemoved_{SCN}$	1.97	60%
6	Motion	$v_{y_{N_{mean}}}$	1.52	76.02%
7	Safety	$F_{max_{RVA}}$	1.44	63.82%
8	Motion	$v_{z_{max}}$	1.42	80%
9	Safety	$F_{max_{SCN}}$	1.27	80%

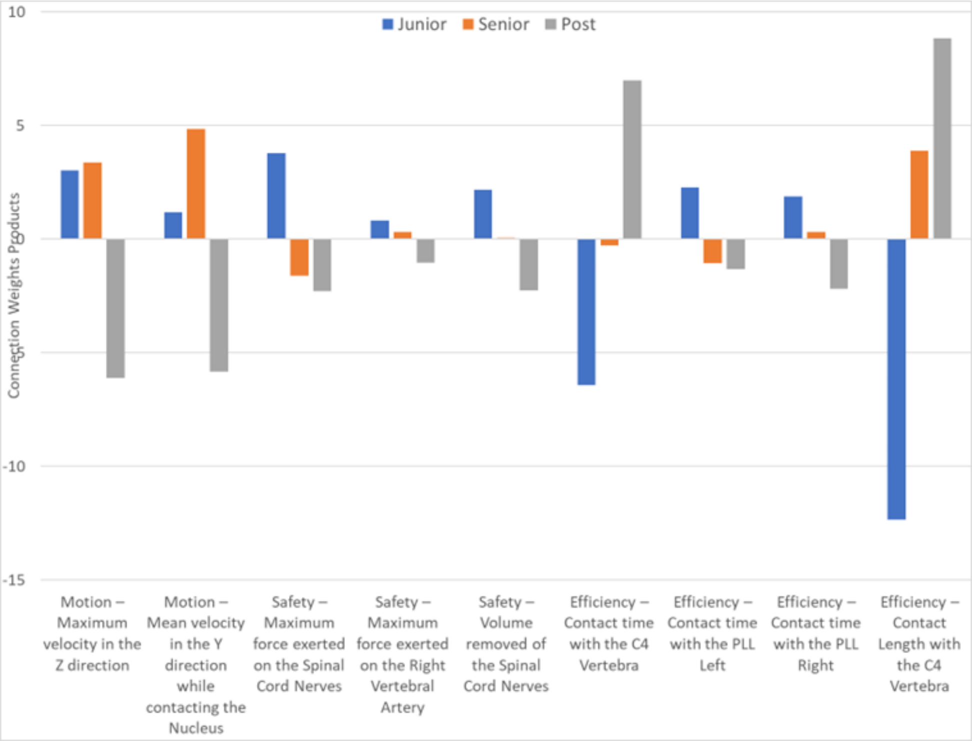


Figure 4-8 Learning patterns of the Connection Weights Products for each input feature.

4.2.4 Discussion

4.2.4.1 Performance of the ANN

The first objective of the study was to leverage an ANN algorithm in the assessment of surgical performance on an ACDF virtual reality simulated scenario. This study focused on the annulus incision step of the ACDF simulation, in which nine features were identified as the most important and subsequently utilized in the development of the neural network. The use of early stopping in model training helped in preventing overfitting. The utilized methodology was successful in developing and training a two-hidden layer neural network that performs well on all three datasets (100% training accuracy, 100% validation accuracy, and 80% testing accuracy). Due to the limited data size used in this study, the accuracy results on the testing set were within the acceptable range. Analysis of the one misclassified individual revealed that the performance associated with this junior resident not only diverged from the junior group, but also resembled the post-resident performance in the most important features that were related to both the junior and post-resident groups (Table 4-9, Table 4-11, and Table 4-14). The participant had positive scores in the contact length (z-score of 0.43) and time (z-score of 0.95) with the C4 vertebra, and a negative score (-0.34) for the maximum velocity in the z-direction. The z-scores specify the number of standard deviations the surgical performance is from the mean values of each feature. Thus, this individual used longer than average contact length and contact time with the C4 vertebra, while utilizing slower than average movements. Based on the CWPs, one interpretation is that these values might increase the likelihood of a post resident classification while they reduce the likelihood of a junior resident classification (Table 4-14). However, this interpretation might not directly hold true without additional analyses, such as the use of other feature importance algorithms as discussed in the next sections.

Table 4-14 Surgical performance metric scores of the misclassified junior resident participant. The performance of this individual diverged from the junior group and resembled the senior group performance, which is evident when comparing the scores to the CWPs of the Junior and Senior resident groups.

Category	Metric	Score	Junior: CWP (%Importance)	Senior: CWP (%Importance)
Efficiency	$ContactLength_{C4}$	0.43	-12.3433 (36.47%)	8.8201 (23.91%)
Efficiency	$ContactTime_{C4}$	0.95	-6.4255 (18.99%)	6.9817 (18.93%)
Motion	$v_{z_{max}}$	-0.34	3.0317 (8.96%)	-6.1178 (16.59%)

4.2.4.2 Insights and Surgical Performance Patterns Revealed by the ANN

The second objective of the study focused on revealing hidden insights identified by the developed neural network model in classifying the ACDF surgical performance level using a new adaptation of the Connection Weights Algorithm. The “black box” analogy has been frequently cited when using deep neural networks, as capturing the true importance of input features can become tedious [15]. In surgical training applications it is important to identify the impact and the relative importance of input features. In a multi-classification task, a useful method of revealing the importance of input features is the Connection Weights Algorithm, which quantifies the impact of each input feature (surgical performance metric) to each class (surgical level) [15]. The algorithm assigns a distinct weight for each feature-class pair by summing the products of all the connection weights that relate an input to an output. The calculated values, termed as the CWPs, can be further leveraged to identify the relative importance of the features to each surgical class. To the best of the author's knowledge, previous studies implemented this algorithm on simple one-hidden layer neural networks [13, 15-17]. As such, the current study is the first to explore the usefulness of the method on multilayered neural networks and subsequently validate the approach using the permutation feature importance method. The significance of the Connection Weights Algorithm lies in its ability to capture the relative contribution of each input feature to each output in both magnitude and sign. For instance, a positive (or a negative) CWP implies that a higher (or a lower) than average feature value is related to a certain class. The use of the CWPs combined with the feature relative importance helps surgical educators design surgical training programs to help guide individual surgical trainees to enhance specific aspects of their skill sets that may need to be improved. This type of personalized residency technical skills training program could maximize trainee bimanual psychomotor training dependent on initial and ongoing information from simulation studies. Our group has proposed a conceptual framework referred to as "Technical Abilities Customized Training" (TACT) [28]. Surgical TACT programs could focus on accelerating top performers, improving areas of weakness in average performers and early identification of trainees with poor surgical performance, while initiating multiple validated methods to enhance and to maintain the bimanual performance of all groups.

4.2.4.2.1 Insights of the ANN Classifications

The Connection Weights Algorithm provides a detailed description of the differences in the surgical performance metrics of the incision task between groups. Differences in the surgical performances are highlighted by the differing values of the CWPs and their relative importance for each input feature among the three groups. Obtaining the relative importance of the features for each of the surgical level groups identifies the most impactful metric that defines a certain surgical level. Consider Table 4-9, Table 4-10, and Table 4-11, the most impactful metrics that distinguish level of surgical performance between the junior, senior, and post-resident groups are efficiency and motion metrics – mainly the C4 vertebra contact length and time ($ContactLength_{C4}$ & $ContactTime_{C4}$) and the maximum velocity in the z direction ($v_{z_{max}}$). Junior group surgical performance differs from the senior and the post-resident groups with respect to the C4 contact length and time metrics, pinpointing the main aspects of the surgical performance that uniquely distinguishes the junior group. Even though the senior and post-resident groups behave similarly in their interactions with the C4 vertebra ($ContactLength_{C4}$), their surgical performance diverges in the motion metrics resulting in a unique performance signature for each group. This might imply that for a new participant, the values scored in these most impactful metrics would influence the likelihood of the surgical performance classifications. For instance, there is an increased likelihood of classifying an individual as a post-resident, as opposed to a junior or a senior resident, when the participant uses relatively slow movements and interacts with the C4 vertebra using relatively long paths and time. This is exemplified by the misclassified junior resident participant in the testing set discussed in the previous section. These results are consistent with the construct validity findings of Ledwos, et al., which found that post-residents utilize longer contact paths and time as compared to the junior group during the incision step [5].

4.2.4.2.2 Educational Learning Patterns Revealed by the ANN

The CWP not only allows for a better understanding of the insights behind the ANN classifications, but it also may help guide trainees in their progression towards surgical expertise. Figure 4-8 demonstrates a visualization of the CWP trends between the junior, senior, and post-resident groups for each feature. Two main learning patterns have been identified using ANN to assess the surgical performance of post-residents, senior and junior residents during the simulated ACDF procedure on the Sim Ortho Platform [6, 18]. These two patterns have been identified as continuous and discontinuous learning. More specifically, continuous learning is associated with sequential improvements of skills as the surgical training level evolves from junior to senior then finally to post-resident surgical level. Discontinuous learning pattern is characterized with non-sequential progression of skills while progressing from the junior resident to the post-resident surgical level, passing through an inconsistent senior resident level. The CWPs of all the safety and efficiency metrics exhibit a continuous learning pattern, while the motion metrics show a discontinuous one.

In all three of the safety metrics, the junior resident group utilizes higher forces on both the right vertebral artery and the spinal cord nerves as well as removes larger volumes of the spinal cord nerves as compared to the senior and post-resident groups. The post-residents use less forces and remove the smallest volumes among the three groups. Hence, a trainee might aim to use lower forces and remove smaller volumes of critical anatomical structures to improve their surgical incision performance. It is to be noted, however, that the incision step would not usually result in significant forces being translated to the right vertebral artery and spinal cord nerves. Nevertheless, the patterns identified in this analysis still underly differences in surgical performances. Efficiency metrics also display continuous learning patterns; however, the direction of the trends differ. Post-residents employ longer paths and more time when interacting with the C4 vertebra compared to senior and junior residents, while junior residents use more time when interacting with both the right and left posterior longitudinal ligaments as compared to the senior and post-resident groups. To improve surgical performance, a trainee would want to limit the interactions to the C4 vertebra while minimizing interactions with the posterior longitudinal ligaments.

The CWPs of the motion features presented in Figure 4-8 exhibit a discontinuous learning pattern that passes through an inconsistent senior surgical training level. Both the junior and post-residents are associated with slower movements as compared to the senior group, with the post-residents using substantially slower controlled movements than the other two resident groups. A dilemma exists for the discontinuous learning patterns, as it is not directly clear from the data generated by the Connection Weights Algorithm whether junior trainees should be trained to the senior resident surgical level or alternatively to the expert post-resident surgical level. Studies are needed to determine the appropriate training approach when discontinuous learning patterns are identified when utilizing VR intelligent tutoring systems.

Rao, et al. provides a detailed description of the ACDF operation [29]. In the annulus incision step, the surgeon is required to perform the incision by using the borders of the vertebra along with the vertebral joint as a guide to avoid injuries to anatomical structures [29]. This description is consistent with the expert performance extracted from the CWPs of post-residents. Their performance is characterized by patient safety related considerations: controlled movements, long paths along the C4 vertebra, low exerted forces on both the right vertebral artery and the spinal cord nerves, and minimal interactions with the posterior longitudinal ligaments. The consistency of the post-resident surgical performance to that described by Rao, et al. increases the confidence in classifying the post-residents as "experts". Our group has developed a performance model for virtual reality procedures which focuses on the expert surgeon primary concern being the safety and efficiency of procedures. It appears reasonable to speculate that for the incision step of the ACDF it may be appropriate to train junior residents to mirror expert level of performance rather than that of the senior group [9, 30].

Unveiling the patterns generated by the neural network and using the Connection Weights Algorithm illuminates some aspects of the "black box" principally focused on safety and efficiency providing new insights on these crucial characteristics of surgical performance.

4.2.4.2.3 Permutation Feature Importance

To further support the novel application of the Connection Weights Algorithm on a multiple hidden ANN, this study further analyzed the importance of the surgical performance metrics by applying the permutation feature importance algorithm. The algorithm was applied on both the training and testing sets, as each can give different insights on aspects of surgical performance and the associated classifications. Using the training set, the permutation feature importance underscores the metrics that are seen important during the learning phase of the model. It highlights the features that the model used in building the connections between surgical performance metrics and surgical classifications. Utilizing the testing set, the algorithm highlights the critical features for the model to perform well on unseen data. It highlights the features that the model relies on when making new predictions. Furthermore, applying the algorithm on both the training and testing sets allows for a comparison of metrics that overlap between the two analyses, thus underscoring the true importance of metrics in both the model's learning and prediction phases.

Using both the training and the testing sets, the most impactful metrics outlined by the permutation feature importance algorithm fall under the efficiency category (Table 4-12 and Table 4-13). More specifically, the contact length and time with the C4 vertebra are seen to be among the top metrics, with the C4 contact length being the most important metric, conforming to the results obtained using the CWPs. The results obtained from the use of the training set (Table 4-12) reached a higher conformity with the results of the CWPs, which is expectable since both utilize the information stored by the final model during training. Similar to the results of the CWPs for the three different classes, the permutation algorithm on the training set found the top three features to be the contact length and contact time with the C4 vertebra, and the maximum velocity in the z-direction. Furthermore, among the safety category, the top feature was the maximum force applied on the spinal cord nerves, similar to the results of the CWPs. While basing the analysis on the training set might be discouraged, the results shed some insights on aspects of surgical classifications that aid in the study's objectives of understanding the most impactful metrics that differentiate surgical performances.

Similarly, the results obtained from applying the algorithm on the testing set demonstrate that the contact length and time with the C4 vertebra to be among the most impactful metrics (Table 4-13). However, there are some discrepancies among the remaining feature rankings when compared to the results of the CWPs, highlighting some of the limitations in interpreting feature importance. While using the trained model to highlight important features might give insights on surgical performance, the identified features might not be directly transferrable to be impactful in the prediction of unseen data. For the current study, two of the most important features found using the permutation feature importance algorithm on the testing set coincided with both the results on the training set and the results of the CWPs. This further supports the findings and analysis of the CWPs and the associated impact of CWP values on predictions, such as the analysis made on the misclassified individual.

4.2.4.3 ACDF Surgical Simulation

The ACDF simulation is a four-part surgical scenario allowing each step to be independently validated and used for training. Each component of the ACDF simulation was previously validated by Ledwos, et al. [5]. The second and third steps of the surgical simulation, concerning the discectomy and osteophyte removal components, have been outlined [6, 18]. These studies utilized some of the same participant data to generate metrics and extract CWPs from developed ANNs, employing similar methodology. These studies only used a single layer ANN with a different optimization technique and included 2 less post-operative participants. Table 4-15 presents a comparison between the analysis conducted on the three simulation components. The discectomy component of the simulation is more complex since three different surgical instruments can be used to complete the task and sixteen metrics to distinguish surgical performance spanning four metric categories. The annulus incision step is the least complex only requiring one surgical instrument and nine metrics spanning three categories to distinguish performance. The osteophyte removal component employs an active drill but can be considered intermediate in complexity using six metrics arising from one category. The discectomy and osteophyte removal requires more expertise to safely complete these tasks, which is consistent with the increased number of safety metrics outlined (Table 4-15). The current study identified nine metrics spanning three categories with the efficiency metrics being more important in distinguishing surgical performance for the annulus incision step.

Table 4-15 Comparison Between the Annulus Incision Step, the Discectomy Step, and the Osteophyte Removal Step of the ACDF surgical Simulation.

	Annulus Incision	Discectomy	Osteophyte Removal
No. of Instruments Used	1 (No. 15 Blade)	3 (Bone Curette, Pituitary Rongeur and Disc Rongeur)	1 (Burr)
No. of Metrics Identified	9	16	6
Metrics Categories	Motion, Safety & Efficiency	Motion, Safety, Efficiency & Cognitive	Safety
Top 3 Ranked Metrics	Motion & Efficiency	Safety & Cognitive	Safety
Most Important Category of Metrics	Efficiency	Safety	Safety
Accuracy of the Model	80%	83.3%	83.3%
Lowest & Highest Magnitude of CWP	0.05 & 12.34	0.02 & 5.24	0.08 & 1.5
Hidden Learning Patterns	Continuous & Discontinuous	Continuous & Discontinuous	Continuous & Discontinuous

4.2.5 Limitations

4.2.5.1 ANN Limitations

The development of the MLP artificial neural network model in this study followed a systematic approach that is based on best practices of utilizing machine learning algorithms for surgical performance assessments (Figure 4-1) [10, 31]. The methodology used in building and training the model focused on avoiding common pitfalls related to overfitting and computational cost. A two-layer network MLP was trained with early stopping to improve the model generalizability and save computational time. Several limitations are associated with the model developed in this study. First, the generalizability of the model is restricted due to the limited available data from only one center. Training the model on larger datasets that span multiple institutions is necessary to develop a more robust model. Second, most studies utilizing Connection Weights Algorithm were based on one-hidden layer neural networks rather than the multiple hidden layer network used in this study [6, 13, 18]. This study adapted the algorithm to be applicable on multiple hidden layer networks and further studies are necessary to support this application. Nevertheless, this study re-analyzed the feature importance using the permutation method to further support the novel adaptation of the Connection Weights Algorithm. The findings of the permutation algorithm suggests that features found important using the training set are not necessarily transferrable to metrics that aid in new predictions. However, metrics that overlapped between the training and testing sets supported the findings of the Connection Weights Algorithm. In the current study, the top two impactful metrics coincided between the training, testing, and the CWPs results, therefore further supporting the current analysis.

4.2.5.2 ACDF Surgical Simulation Limitations

The ACDF simulator utilized in this study does not encompass the many complex interactions that occur in the performance of a patient ACDF procedure. Several important components of the procedure are automated preventing an assessment of important aspects of surgical exposure of the appropriate cervical disc space. The OSSimTech simulator used was developed for right-handed users limiting both its applicability to left-handed participants and the ability to quantitate bimanual performance. Previous studies in our group have demonstrated differences in right-and left- handed ergonomics and modifications in the platform are necessary to allow bimanual skills performance to be assessed and provide a more holistic understanding of the expertise necessary to safely carry out an ACDF [30, 32].

The simulator utilizes an advanced voxel-based gaming engine that generates the graphical representation of the anatomical structures and instrument interactions and leverages haptic and auditory feedback to augment the experiential realism of the simulation. Recent studies have highlighted the importance of using physics-based haptics to ensure the accuracy and reliability of the generated force feedback and the importance of extracting and implementing realistic physics-driven feedback using data from cadaveric experiments [3, 33, 34]. Forces generated using simulators with discrete or heuristic approaches, not based on constitutive modeling from the continuum mechanical method, may not accurately provide or, consequently, record the forces experienced in real patient operations which might tend participants to respond with forces not used in reality. Naturally, this error presents a further limitation when utilizing the force metrics in surgical training, as the benchmark values identified by the simulator might be different to reality and thus resulting in training junior residents to wrong skill levels. On a similar note, the simulator used in the current study has detected and identified interactions with anatomical structures that usually are not experienced during the incision step. The results indicate that applying pressure on the annulus resulted in forces being translated to the vertebral arteries, the posterior ligaments, and the spinal cord nerves. Although this might be a misrepresentation of the actual surgical step, the main outcomes of the analysis still hold. Indeed, multiple studies including the present one has found that more experienced surgeons tend to use lower and more controlled forces as compared to junior trainees [6, 18, 32]. Moreover, the expert surgeons in the current study were able to avoid unnecessary interactions with the mentioned anatomical structures by following the path of the vertebral body, indicating that expert performance would not generate forces on irrelevant anatomical structures. This result further supports the validity of the simulator in successfully differentiating between surgical levels. The development of smart operative instruments capable of measuring force application during patient procedures, as being developed in the Musculoskeletal Biomechanics Research Lab, to the forces assessed in identical scenarios utilized in virtual reality simulators will allow educators to more accurately assess the formative role of these platforms.

4.2.6 Conclusion

This study demonstrates the use of an ANN to distinguish virtual reality surgical performance for assessment and training of surgical performance. Our results outline the significant potential of extracting hidden patterns within neural networks to highlight the important composites of expert and less skilled surgical performances, and the potential integration of ANNs with virtual reality surgical simulator platforms for formative and summative assessment.

4.2.7 References

[1] M. Goldenberg and J. Y. Lee, "Surgical Education, Simulation, and Simulators-Updating the Concept of Validity," (in eng), *Curr Urol Rep,* vol. 19, no. 7, p. 52, May 17 2018. https://doi.org/10.1007/s11934-018-0799-7

[2] M. Pfandler, M. Lazarovici, P. Stefan, P. Wucherer, and M. Weigl, "Virtual reality-based simulators for spine surgery: A systematic review," *The Spine Journal,* vol. 17, 05/01 2017. https://doi.org/10.1016/j.spinee.2017.05.016

[3] K. El-Monajjed and M. Driscoll, "Analysis of Surgical Forces Required to Gain Access using a Probe for Minimally Invasive Spine Surgery via Cadaveric-based Experiments towards use in Training Simulators," *IEEE Transactions on Biomedical Engineering,* pp. 1-1, 2020. https://doi.org/10.1109/TBME.2020.2996980

[4] S. Alkadri, "Kinematic Study and Layout Design of a Haptic Device Mounted on a Spine Bench Model for Surgical Training," Undergraduate Honours Program - Mechanical Engineering, Mechanical Engineering, McGill University, 2018.

[5] N. Ledwos, N. Mirchi, V. Bissonnette, A. Winkler-Schwartz, R. Yilmaz, and R. F. J. O. N. Del Maestro, "Virtual reality anterior cervical discectomy and fusion simulation on the novel sim-ortho platform: validation studies," *Operative Neurosurgery,* 2020.

[6] N. Mirchi *et al.*, "Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance," *Operative Neurosurgery,* vol. 19, no. 1, pp. 65-75, 2019. https://doi.org/10.1093/ons/opz359

[7] F. E. Alotaibi *et al.*, "Assessing Bimanual Performance in Brain Tumor Resection With NeuroTouch, a Virtual Reality Simulator," *Operative Neurosurgery,* vol. 11, no. 1, pp. 89-98, 2015. https://doi.org/10.1227/NEU.0000000000000631

[8] H. Azarnoush *et al.*, "Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection," (in eng), *Int J Comput Assist Radiol Surg,* vol. 10, no. 5, pp. 603-18, May 2015. https://doi.org/10.1007/s11548-014-1091-z

[9] R. Sawaya *et al.*, "Development of a performance model for virtual reality tumor resections," (in English), *Journal of Neurosurgery,* vol. 131, no. 1, p. 192, 2018. https://doi.org/10.3171/2018.2.Jns172327

[10] A. Winkler-Schwartz *et al.*, "Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation," (in eng), *J Surg Educ,* vol. 76, no. 6, pp. 1681-1690, Nov-Dec 2019. https://doi.org/10.1016/j.jsurg.2019.05.015

[11] N. Mirchi, V. Bissonnette, R. Yilmaz, N. Ledwos, A. Winkler-Schwartz, and R. F. Del Maestro, "The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine," *PLOS ONE,* vol. 15, no. 2, 2020. https://doi.org/10.1371/journal.pone.0229596

[12] A. Winkler-Schwartz *et al.*, "Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation," *JAMA Network Open,* vol. 2, no. 8, 2019. https://doi.org/10.1001/jamanetworkopen.2019.8363

[13] J. D. Olden, M. K. Joy, and R. G. Death, "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data," *Ecological modelling,* vol. 178, no. 3-4, pp. 389-397, 2004.

[14] N. M. J. J. o. e. i. Nasrabadi, "Pattern recognition and machine learning," vol. 16, no. 4, p. 049901, 2007.

[15] J. Heaton, S. McElwee, J. Fraley, and J. Cannady, "Early stabilizing feature importance for TensorFlow deep neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN),* 2017, pp. 4618-4624.

[16] O. Ibrahim, "A comparison of methods for assessing the relative importance of input variables in artificial neural networks," *Journal of Applied Sciences Research,* vol. 9, no. 11, pp. 5692-5700, 2013.

[17] J. D. Olden and D. A. Jackson, "Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks," *Ecological Modelling,* vol. 154, no. 1, pp. 135-150, 2002/08/15/ 2002. https://doi.org/10.1016/S0304-3800(02)00064-9

[18] A. Reich *et al.*, "Artificial Neural Network Approach to Competency-Based Training " 2020.

[19] C. Huang, H. Cheng, Y. Burreau, H. M. Ladak, and S. K. Agrawal, "Automated Metrics in a Virtual-Reality Myringotomy Simulator: Development and Construct Validity," (in eng), *Otol Neurotol*, vol. 39, no. 7, 2018. https://doi.org/10.1097/mao.0000000000001867

[20] R. M. Kwasnicki, R. Aggarwal, T. M. Lewis, S. Purkayastha, A. Darzi, and P. A. Paraskeva, "A Comparison of Skill Acquisition and Transfer in Single Incision and Multi-port Laparoscopic Surgery," *Journal of Surgical Education,* vol. 70, no. 2, pp. 172-179, 2013/03/01/ 2013. https://doi.org/10.1016/j.jsurg.2012.10.001

[21] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint*, 2019.

[22] S. Chintala. *DEEP LEARNING WITH PYTORCH: A 60 MINUTE BLITZ*. Available: https://pytorch.org/tutorials/beginner/deep_learning_60min_blitz.html#deep-learning-with-pytorch-a-60-minute-blitz

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[24] M. W. Beck, "NeuralNetTools: Visualization and Analysis Tools for Neural Networks," *Journal of statistical software,* vol. 85, no. 11, p. 20, 2018-07-30 2018. https://doi.org/10.18637/jss.v085.i11

[25] S. Xie, A. T. Lawniczak, and J. Hao, "Modelling Autonomous Agents' Decisions in Learning to Cross a Cellular Automaton-Based Highway via Artificial Neural Networks," *Computation,* vol. 8, no. 3, p. 64, 2020.

[26] L. J. M. l. Breiman, "Random forests," vol. 45, no. 1, pp. 5-32, 2001.

[27] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," *Journal of Machine Learning Research,* vol. 20, no. 177, pp. 1-81, 2019.

[28] A. Winkler-Schwartz *et al.*, "Bimanual Psychomotor Performance in Neurosurgical Resident Applicants Assessed Using NeuroTouch, a Virtual Reality Simulator," *Journal of Surgical Education,* vol. 73, no. 6, pp. 942-953, 2016/11/01/ 2016. https://doi.org/10.1016/j.jsurg.2016.04.013

[29] A. S. Rao, A. L. R. Michael, and J. Timothy, "Surgical Technique of Anterior Cervical Discectomy and Fusion (ACDF)," in *Practical Procedures in Elective Orthopedic Surgery: Upper Extremity and Spine*, P. V. Giannoudis, Ed. London: Springer London, 2012, pp. 189-193.

[30] R. Sawaya *et al.*, "Virtual reality tumor resection: the force pyramid approach," *Operative Neurosurgery,* vol. 14, no. 6, pp. 686-696, 2018.

[31] A. Cheng *et al.*, "Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements," *Advances in Simulation,* vol. 1, no. 1, pp. 1-13, 2016.

[32] H. Azarnoush *et al.*, "The force pyramid: a spatial analysis of force application during virtual reality brain tumor resection," *Journal of Neurosurgery*, vol. 127, no. 1, pp. 171-181, 2016.

[33] N. Choudhury, N. Gélinas-Phaneuf, S. Delorme, and R. Del Maestro, "Fundamentals of neurosurgery: virtual reality tasks for training and evaluation of technical skills," *World Neurosurgery,* vol. 80, no. 5, 2013.

[34] S. Delorme, D. Laroche, R. DiRaddo, and R. F. Del Maestro, "NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training," *Operative Neurosurgery*, vol. 71, no. suppl_1, 2012.

4.3 Article 4: Unveiling Surgical Expertise Through Machine Learning in a Novel VR/AR Spinal Simulator: A Multilayered Approach Using Transfer Learning and Connection Weights Analysis

Sami Alkadri, Rolando Del Maestro, Mark Driscoll

Author names:

Sami Alkadri B.Eng., Ph.D. Student (1,3), Rolando F. Del Maestro MD, PhD (3), Mark A. Driscoll, PEng., Ph.D., Associate Professor (1,2).

Institutional affiliations:

Musculoskeletal Biomechanics Research Lab, Department of Mechanical Engineering, McGill University, Macdonald Engineering Building, 815 Sherbrooke St W, Montreal, H3A 2K7, QC, Canada.

Orthopaedic Research Lab, Montreal General Hospital, 1650 Cedar Ave (LS1.409), Montreal, Quebec, Canada, H3G 1A4

Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology & Neurosurgery, Montreal Neurological Institute, McGill University, 2200 Leo Pariseau, Suite 2210, Montreal, Quebec Canada, H2X 4B3.

Corresponding author:

Mark Driscoll

Mailing Address: Macdonald Engineering Building, RM 153, 817 Sherbrooke St W, Montreal, Quebec H3A 2K7, Canada

Phone: 514-398-6299

Fax: 514-398-7365

Email Address: mark.driscoll@mcgill.ca

ABSTRACT

Background

Virtual and augmented reality surgical simulators are being recognized as safe and efficient for training psychomotor skills. The integration of machine learning in these simulators has enhanced the analysis and classification of surgical performance, extracting valuable insights into the composites of surgical expertise. While methods, such as the Connection Weights Algorithm, have shown promise in analyzing these machine learning models, challenges like the small sample size (small number of participants (N)) in surgical simulator trials persist. The small N problem impacts the generalizability and therefore the robustness of the models. Potential solutions, such as data augmentation and transfer learning from models trained on similar surgical tasks, offer a way to address this limitation.

Objective

This study aims to demonstrate the efficacy of artificial neural network and transfer learning algorithms in evaluating virtual surgical performances, applied to a simulated oblique lateral lumbar interbody fusion technique in an augmented and virtual reality simulator.

Design

The study developed and integrated an artificial neural network algorithm within a novel simulator platform, using data from the simulated tasks to generate 276 performance metrics across motion, safety, and efficiency.

Setting

Musculoskeletal Biomechanics Research Lab; Neurosurgical Simulation and Artificial Intelligence Learning Centre, McGill University, Montreal, Canada.

Participants

Twenty-seven participants were recruited and divided into 3 groups: 9 post-residents, 6 senior and 12 junior residents.

Results

Two models, a stand-alone model trained from scratch and another leveraging transfer learning, were trained on nine selected surgical metrics achieving 75% and 87.5% testing accuracy respectively.

Conclusions

This study outlines the benefits of integrating transfer learning with artificial neural networks to surgical simulators in understanding composites of expertise performance.

Keywords

Multilayered artificial neural network, transfer learning, data augmentation, feature importance, virtual reality, surgical simulation, surgical education, performance metric, surgical expertise

Funding

NSERC Collaborative Research Development (CRD) Grant; Franco Di Giovanni Foundation; Brain Tumour Foundation of Canada Brain Tumour Research Grant; a Medical Education Research Grant from the Royal College of Physicians and Surgeons of Canada; and the Montreal Neurological Institute and Hospital.

Conflict of interest statement

No competing interests to declare.

4.3.1 Introduction

The use of virtual (VR) and augmented reality (AR) surgical simulators in training and evaluating surgical skills is gaining popularity supported by studies highlighting their effectiveness [1]. The integration of haptic technology, providing real-time force-feedback, enhances the authenticity of the training programs [2]. Haptics in surgical simulations allow trainees to develop a tactile understanding of procedures before being involved with patient surgical procedures, leading to improved learning outcomes, even when using non-realistic voxel-based gaming engine forces. However, our group strives to show the added benefits of incorporating realistic physics-based haptic feedback on learning outcomes through detailed quantification of surgical forces from cadaver studies [3, 4]. This aspect is deemed crucial in the development of new surgical simulator platforms, particularly for challenging and tactile-dependent minimally invasive spinal surgeries (MISS) [5, 6]. One such platform is the physics-based VR/AR spinal surgical simulator developed by our group to simulate the Oblique Lateral Lumbar Interbody Fusion (OLLIF) surgery.

VR/AR simulators generate extensive data of user psychomotor interactions in simulations. Our group has demonstrated that converting this data into performance metrics effectively classifies individuals by expertise level and aids in enhancing their performance [7-10]. This naturally gave rise to the utility of machine learning (ML) – a subset of artificial intelligence (AI) – in exploiting these large data sets for more detailed classification and to enhance the training capabilities of simulators [11]. Multilayered perceptron (MLP) artificial neural networks (ANNs), a deeper subset of ML, has shown promise in the domain of surgical simulation due to their ability to learn and model complex non-linear patterns within the data collected during simulated tasks [12]. ANNs resemble biological neural networks; they consist of multiple interconnected neurons organized into layers, with each layer processing data and transferring it to the next layer [12]. Despite the effectiveness of ML algorithms in classifying surgical simulation performance, there are limitations. One limitation is the focus on classification, while neglecting to delve deeper into the underlying reasons for the classifications or quantify the relative importance of performance metrics used by the ML models [13-15]. Our previous study, on VR anterior cervical discectomy and fusion simulation, addressed this limitation by introducing a novel application of the

Connection Weights Algorithm (CWA) on multi-layered ANNs [16]. The CWA, originally created by Olden and Jackson [15], provided an improved understanding of the contributions of individual performance metrics to the classification task in one-layered ANNs. By employing this novel approach on a multi-layered ANN, this study aimed to demonstrate the usefulness of the approach in identifying the relative importance of each metric in complex models.

Another limitation associated with deploying ML algorithms with surgical simulations is the small dataset (small N) due to difficulties in recruiting participants, especially for simulators of less common surgical procedures. A potential solution to address this issue is data augmentation, which introduces slight variations in the form of jittering (i.e. noise) or scaling to the original dataset to increase the size, thus aids in preventing overfitting and improving generalizability of the model [17]. Transfer learning is another effective strategy, where the insights from a model trained on a similar, but distinct task are utilized [18]. By applying transfer learning, one may build on existing models developed for similar surgical simulators to create more robust systems.

To that end, the novelty of the current study lies in two key areas: 1) Classify surgical performance and identify the key performance metrics essential in determining surgical expertise using a novel physics-based VR/AR spinal surgical simulator. This approach builds on our previous work, further enriched by examining the advantages of data augmentation and transfer learning in surgical simulators. Specifically, we adapt the learning from an ANN model, previously developed for a similar spinal simulator, to our new model, and rigorously assess its performance. 2) Examine the novel CWA approach developed by the authors by applying it to both the newly developed ANN and the ANN based on transfer learning. These models are further validated using the permutation feature importance, a well-established technique for interpreting ML models.

4.3.2 Material and Methods

4.3.2.1 The Simulator Platform & The Simulated Scenario

The platform used in this study is a novel VR/AR surgical simulator developed by McGill University in affiliation with CAE Healthcare and Depuy Synthes part of Johnson & Johnson. The platform consists of a high-performance gaming laptop (i7-8750H), two flat panel monitors to match the interface in the operating room, and a haptic ENTACT W3D device generating realistic force feedback, (Figure 4-9(a)). The simulation focusses on three phases of an OLLIF surgery: gaining access through the back muscles, removing the intervertebral disc, and inserting graft and a spinal cage. The detailed steps along with the surgical tools used at each phase are shown in Figure 4-9(b).

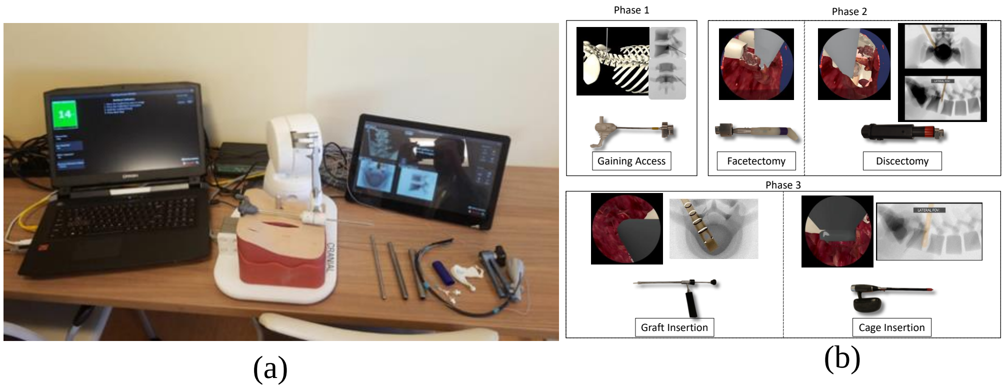


Figure 4-9 (a) Simulator layout. Laptop (left) indicates the instruction of the surgery process. The haptic device and benchtop model are in the middle. External display (right) indicates the four cameras that demonstrate the surgical area. (b) The three phases of the simulated surgery: Phase 1 includes gaining access to the disc using a Multitool; Phase 2 includes facetectomy using a Burr Tool followed by a discectomy using a Concord Tool; Phase 3 includes graft and cage insertions using the respective tools.

Phase 1 of the simulated surgery includes gaining access to the surgical area using a multiprobe tool. Phase 2 requires the participant to first use a Burr tool for drilling and performing a facetectomy, followed by using the Concord tool's suction mechanism to remove the disc. In

Phase 3, the participant is required to insert a graft and a cage using the graft and cage insertion tools. The force feedback replicates the resistance provided by the instruments when penetrating through the muscles during an actual surgery using tailored empirical response curves extracted during cadaver experiments [4]. The empirical curves have implicitly incorporated the non-linearity and viscoelasticity of realistic physiological tissue responses [4]. The current study focuses on the first two phases, gaining access and facetectomy & discectomy. Prior to the start of the simulation, participants were made aware of all steps and instruments needed to complete the procedure via verbal and written instructions. No time limit was imposed on participants.

4.3.2.2 Participants

This study utilized participant data previously collected for the face, content, and construct validation study of this simulator platform. Thirty-four participants were initially recruited to perform the virtual OLLIF scenario. Seven expert orthopedic surgeons out of the 34 participants were recruited in a side-by-side cadaver trial, where participants completed a minimally invasive spinal fusion surgery on a cadaver, then immediately repeated the identical procedure on the surgical trainer/simulator. The remaining participants completed the trial without performing the cadaver surgery. Due to errors during the simulation runs 7 participant data could not be utilized. Therefore 27 individuals were included in the current analysis: 12 post-residents, 6 senior residents, and 9 junior residents. Table 4-16 and Table 4-17 outline the demographics and the difference in experiences and knowledge of the 27 participants. The participants were divided into three groups: A post-resident group (3 neurosurgeons, 5 spine surgeons, 2 spine fellows, and 2 neurosurgical fellows), a Senior-Resident group (4 PGY 4-6 neurosurgery and 2 PGY 4-5 orthopaedics residents), and a Junior-Resident group (4 PGY 1-3 neurosurgery and 5 PGY 1-3 orthopaedics residents).

This study was approved by the Institutional Review Board (IRB) of the Faculty of Medicine and Health Sciences at McGill University. All participants signed an approved written consent form prior to providing demographic and other information and beginning the simulation of the virtual reality spine surgery simulation which took on average 60 minutes to complete. This article follows the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and Best Practices for Machine Learning to Assess Surgical Expertise [19, 20].

Table 4-16 Demographics of the post-resident, senior-resident, and junior-resident groups.

	Junior Residents	Senior Residents	Post-Residents
No. of individuals	9	6	12
Sex			
Male	8	5	11
Female	1	1	1
Surgical Specialty Level of Training	Neurosurgery	Orthopaedic Surgery	
PGY 1-3	4	5	
PGY 4-6	4	2	
Fellows	2	2	
Consultants	3	5	

Table 4-17 Differences in previous experience, knowledge, and comfort level of the groups.

	Junior Residents	Senior Residents	Post-Residents
No. of individuals in each group who:			
Have previous experience using a surgical simulator	2 (22%)	5 (83%)	10 (83%)
Assisted on a TLIF	7 (77%)	6 (100%)	10 (83%)
Performed a TLIF solo	0 (0%)	0 (0%)	7 (58%)
Medina self-rating on 5-point Likert scale:			
Textbook Knowledge of a TLIF	3.0 (3.0 – 4.0)	3.0 (3.0 – 4.0)	3.5 (1.0 – 5.0)

Surgical Knowledge of a TLIF	3.0 (2.0 – 4.0)	3.0 (3.0 – 4.0)	3.5 (1.0 – 5.0)
Comfort level performing a TLID with a consultant in the room	3.0 (1.0 – 4.0)	4.0 (2.0 – 5.0)	4.5 (2.0 – 5.0)
Comfort level performing a TLIF solo	1.0 (1.0 – 2.0)	2.0 (1.0 – 4.0)	3.0 (1.0 – 5.0)

4.3.2.3 Machine Learning Analysis

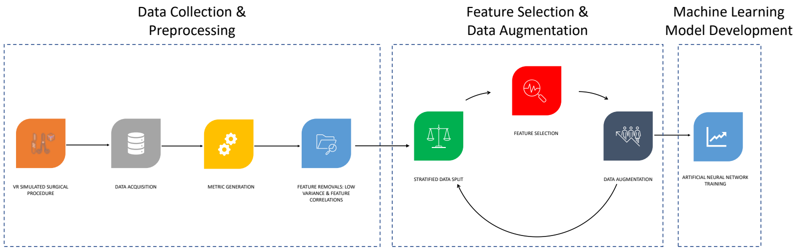


Figure 4-10 The study methodology consists of three main steps: Data Collection & Preprocessing, Feature Selection & Data Augmentation, and Machine Learning Model Development

A systematic approach was used in integrating a MLP ANN in classifying the virtual surgical performance. As illustrated in Figure 4-10, the methodology can be divided into three main steps: Data collection & Preprocessing, Feature Selection & Data Augmentation, and Machine Learning Model Development. While the first two steps of the methodology were implemented only once, this study develops and compares two distinct MLP ANN architectures: a MLP ANN constructed from scratch and another leveraging transfer learning from a previously trained two layered ANN model. The current study expands on the methodology developed in our previous publication to include data augmentation at the feature selection phase and the use of transfer learning in the model development phase [16].

4.3.2.3.1 Data Collection & Preprocessing

During each virtual reality surgical simulation, the platform tracked tool use, converting this data into metrics to evaluate participant performance, as previously detailed in our validity studies [3]. Initially, 276 features were identified through expert opinions, literature on spinal fusion surgery, and novel data-derived metrics. However, this extensive feature set risked overfitting due to the "curse of dimensionality", leading to a less interpretable model [17]. This is further exacerbated in cases of small datasets as in the current context. The current study utilized a combination of feature reduction, data augmentation, and transfer learning in a carefully constructed methodology to overcome these limitations.

All generated metrics were assigned into one of three main categories: motion, safety, or efficiency. The performance metrics were then normalized using z-score normalization to reduce impact of outliers. Data extraction, metrics generation and z-score normalization were done in Python (Version 3.7, OR USA). An initial feature reduction removed features with zero or near-zero variance and those highly correlated, reducing the feature count to 168.

4.3.2.3.2 Feature Selection & Data Augmentation

Developing a machine learning model involves key steps for optimal and generalizable outcomes. This study’s iterative approach, depicted in Figure 4-10, refined the feature space to essential metrics, addressing the "curse of dimensionality" and removing unimportant features. Initially, the dataset, with underrepresented classes, underwent a stratified split into training, validation, and testing sets for class balance (Table 4-18). Following the data split, a sequential forward selection (SFS) algorithm with a built-in machine learning model was used to remove irrelevant metrics that may not be useful in distinguishing surgical performance. The SFS algorithm iteratively builds and evaluates optimal feature subsets, continuing until identifying the optimal subset. This study employed a 6-fold cross validation Neural Network model as part of the SFS algorithms for feature selection. The data split was firstly passed into the SFS algorithm, which reduced the feature space from 168 features to 16 features (Table 4-19).

Table 4-18 First stratified split of the original dataset into training, validation, and testing sets.

Classes	Original Dataset	Training Dataset	Validation Dataset	Testing Dataset
Junior	9	5	2	2
Senior	6	4	1	1
Post	12	7	2	3
Total	27	16	5	6

Table 4-19 SFS average 6-fold validation accuracy during the 2 passes of the Feature Selection & Data Augmentation Step.

Features Prior to SFS	Features Post SFS	Avg. SFS 6-Fold Validation Accuracy
168	16	82.5%
16	9	92.5%

With the refined feature set of 16, data augmentation in the form of data jittering was used to address the limitations of small dataset as well as imbalanced classes. This was specifically used to balance the underrepresented Junior and Senior Resident classes, achieving an equal distribution of 12 data points per class. Data jittering introduces small variations or "noise" to the existing data by randomly sampling from a group of participants and applying a slight random noise. In this study, a random gaussian noise centered at 0 with a standard deviation of 0.01 was used. Although scaling and jittering were both potential augmentation methods, jittering was more appropriate than scaling in the context of surgical performance metrics. As compared to data scaling, data jittering provides: (1) a natural variability in the data that may arise from hand tremors, tool handling errors, and dexterity control; (2) preserves realistic values of surgical performance features – for example scaling forces might lead to unrealistic values; (3) avoids skewing feature distributions; and (4) aligns well with the pre-normalized data.

To prevent information leakage from the testing set during model development, datapoints from the original test set were isolated directly after applying data augmentation. The rest were re-split, allocating 78% to training and validation sets. These subsets were then passed through the SFS algorithm yielding a final of 9 surgical performance metrics. With the refined and augmented data, the machine learning model development was initiated. The split dataset and the nine features selected in the final step are shown in Table 4-20 and Table 4-21 respectively.

Table 4-20 Final stratified split of the dataset into training, validation, and testing sets.

Classes	Original Dataset	Training Dataset	Validation Dataset	Testing Dataset
Junior	12	7	2	3
Senior	12	8	2	2
Post	12	7	2	3
Total	36	22	6	8

Table 4-21 Nine final metrics resulted from the second pass into the SFS algorithm used in this study.

Metric Category	Metric Description	Metric Abbreviation
Motion	Sign changes of the Multitool acceleration in the X direction	$sign_{a_{x_{Multitool}}}$
	Mean jerk in the Y direction while using the Burr Tool	$J_{Y_{BurrTool_{mean}}}$
	Mean velocity while using the Burr Tool	$v_{BurrTool_{mean}}$
	Mean velocity during the Discectomy Surgical Step	$v_{Discectomy_{mean}}$
Safety	Mean torque exerted by the Burr Tool	$T_{BurrTool_{mean}}$
	Mean force exerted on the NP during the Gaining Access Surgical Step	$F_{NP_{GainingAccess_{mean}}}$
	Mean force exerted on the M5 Muscle during the Discectomy Surgical Step	$F_{M5_{Discectomy_{mean}}}$

	Mean force exerted on the M6 Muscle while using the Concorde tool	$F_{M6_{ConcTool_{mean}}}$
	Mean force exerted on the SAP while using the Burr tool	$F_{SAP_{BurrTool_{mean}}}$

4.3.2.3.3 Machine Learning Model Development

Following the feature selection & data augmentation step, building and training the MLP ANNs were initiated. The same methods for training and optimizing hyperparameters were applied to both models: the MLP ANN built from scratch and the one developed using transfer learning.

A PyTorch framework was used for building and training our MLP models, as detailed in our prior publication [16], and inspired by frameworks outlined by Paszke, et al. [21] and Chintala [22]. The models were trained using cross-entropy loss and stochastic gradient descent with momentum (SGD with momentum). The ReLu activation function, along with Lecun weights initialization, was implemented as per PyTorch's default settings. To avoid overfitting, early stopping was incorporated based on the validation set's loss and accuracy: training stopped if validation loss increased, or accuracy decreased consistently over 200 epochs. Our algorithm also saved model parameters upon validation loss improvement and kept a record of training and validation accuracies and loss values.

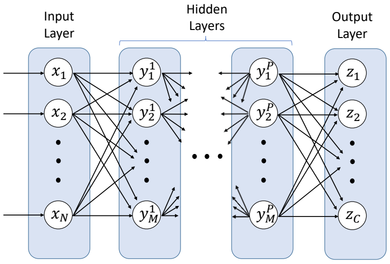


Figure 4-11 A general MLP diagram showing the input layer, the hidden layers and the interconnected hidden units, and the output layer.

An MLP architecture consists of multiple interconnected hidden neurons within multiple layers as presented in Figure 4-11. Optimizing an MLP involves tuning various hyperparameters related to both the architecture and the training process. For the model architecture, key hyperparameters include the number of hidden layers and hidden units. For training the MLP with SGD, important hyperparameters are the learning rate and momentum of the SGD algorithm. Table 4-22 presents a provides a comprehensive list of potential hyperparameter values, selected based on best practices in literature for using SGD with momentum in MLP neural networks [17]. This study advances beyond the manual, semi-systematic grid search approach of our previous publication, implementing a systematic grid search algorithm to evaluate all possible models created from the hyperparameter combinations. This approach was used for both the standalone MLP and the MLP with transfer learning. The grid search was aimed to identify the best performing models, using model performance on the validation set as the primary criterion, similar to our approach with early stopping.

Table 4-22 Hyperparameters potential values.

No. of Hidden Layers	1	2	3		
No. of Hidden Units	6	10	20	40	100
Learning Rate	0.0001	0.0005	0.001	0.005	0.01
Momentum	0.6	0.7	0.8	0.9	1

To enhance performance and mitigate the limitations of a small dataset, transfer learning was implemented, using a 2-layered ANN model previously developed for the Sim-Ortho simulator, a VR simulator for an annulus incision task in anterior cervical discectomy and fusion (ACDF) scenarios by OSSimTech [16]. The hyperparameters and architecture of this model are detailed in Table 4-23 and Figure 4-12. Transfer learning extracts knowledge from models trained on similar tasks [18]. Multiple approaches exist to transfer the knowledge learnt by a previously built ML model. Two main methods are frequently highlighted in the literature: fine-tuning a pre-trained model or using it as a feature generator [17, 18]. Fine tuning the model to adapt to the new dataset is seen as a continuation of the model's training phase on the new dataset. This method is extensively used in deep learning applications where firstly the outmost layers are fine tuned (shallow tuning) before incrementally engaging and fine tuning the entirety of the layers (deep tuning). This process leverages the idea that an ANN's last layers hold task-specific high-level features, while the initial layers contain low-level features common to many tasks [18]. However, overfitting is an important drawback of this method when dealing with ANNs with few layers applied on small datasets, as in the current application.

Table 4-23 Pre-Trained Model in the side study performed on the Sim-Ortho VR simulator developed by OSSimTechTM

Hidden Inputs Per Layer	Hidden Layers	SGD Learning Rate	SGD Momentum
40	**2**	**0.001**	**0.7**

Another approach is to leverage the knowledge stored in the trained model by freezing its layers and appending new set of layers to the output of the learnt model. This method is also known as the feature extractor method as the learnt layers act as a sophisticated filter that transforms the input data into high-level features that result in better classifications, especially in small datasets. This approach mitigates overfitting and improves model generalizability. In this study, this method was adopted by freezing the pre-trained layers of the previously developed model and appending new, trainable layers. This was done by loading the old model and setting it into evaluation mode, before accessing the output of the second hidden layer to append the new and trainable layers. The training of the new layers followed the same approach described above for the stand alone MLP, including the systematic grid search to find the optimal combination of hyperparameters.

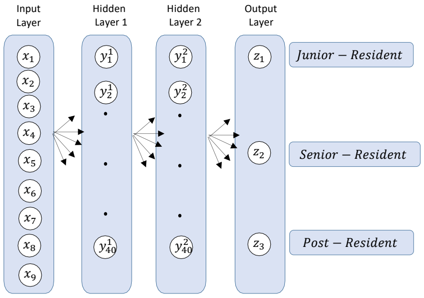


Figure 4-12 Pre-trained model architecture

Table 4-6 displays the top-performing one-layer, two-layer, and three-layer standalone ANNs, as well as those using transfer learning, determined by our search criteria. Notably, the three-layered standalone ANN and the one appended layer transfer learning model showed superior performance on the validation set. The table also details the optimal hyperparameters for each model. Figure 4-13 illustrates their training progress, where validation accuracy and loss were assessed after each training epoch. Early stopping was frequently employed, training stopped at 3500 epochs for the standalone model and 890 epochs for the transfer model (Figure 4-13).

Table 4-24 The best performing models in each of the one-layered, two-layered, and three-layered ANNs.

Model	Hidden Inputs Per Layer	Hidden Layers	SGD Learning Rate	SGD Momentum	Validation Accuracy	Validation Loss
Stand Alone Model	20	**1**	**0.001**	**0.8**	**66.67%**	**0.32**
	40	**2**	**0.001**	**0.7**	**83.33%**	**0.26**
	20	**3**	**0.0005**	**0.8**	**100%**	**0.14**
Transfer Learning Model	6	**1Ŧ**	**0.0005**	**0.6**	**100%**	**0.01**
	6	**2Ŧ**	**0.001**	**0.6**	**83.33%**	**0.04**
	20	**3Ŧ**	**0.005**	**0.6**	**83.33%**	**0.05**

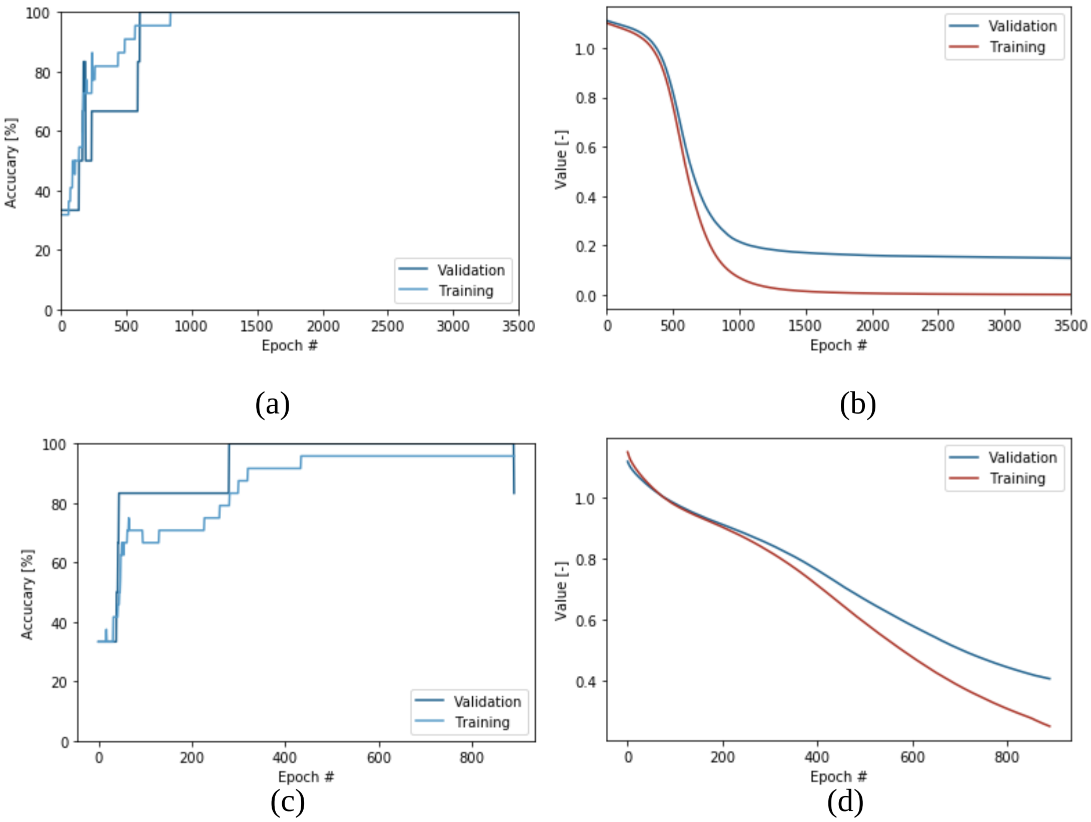
Ŧ The hidden layers indicated in the MLP ANN with transfer learning are the new appended layers after the 2 pre-trained hidden layers.

Table 4-25 Best performing model found within the grid search.

Model		Hidden Inputs Per Layer	Hidden Layers	SGD Learning Rate	SGD Momentum
Stand Alone Model		20	3	0.0005	0.8
Transfer Learning Model	**New Layers**	6	1	0.0005	0.6
	Pre-Trained Layers	40	2	N/A[Ŧ]	N/A[Ŧ]

Ŧ The Pre-Trained Layers are frozen and therefore not updated during training.

The Connection Weights Algorithm, originally developed by Olden and Jackson [15], was used to understand and quantify the relative impact of each metric on the classification task. The algorithm was developed for one-hidden layer networks and assigns a distinct weight for each feature-class pair by summing the products of all the connection weights that relate an input to an output, as demonstrated by Figure 4-14 and Equation (20). In our previous publication, the Algorithm was adapted to a multilayer neural network to calculate the Connection Weights Product (CWP) [16]. More specifically, as demonstrated by Figure 4-5 and Equation (21), the study adapted the algorithm to a two hidden layer network – the model used as the basis of the transfer learning model in the current study.



(a)

(b)

(c)

(d)

Figure 4-13 The performance of the models at each training epoch: (a) the accuracy of the optimal stand-alone model on the training and validation sets at each training epoch; (b) the value of the loss function of optimal stand-alone model on the training and validation sets at each training epoch; (c) the accuracy of the optimal model with transfer learning on the training and validation sets at each training epoch; (d) the value of the loss function of optimal model with transfer learning on the training and validation sets at each training epoch.

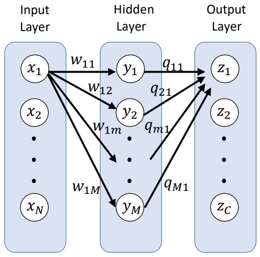


Figure 4-14 Schematic of a one hidden layer network demonstrating the weights that connect the first input node to the first output node.

$$CWP_{x,z} = \sum_{m=1}^{M} w_{xm} q_{mz} \qquad \text{Equation (20)}$$

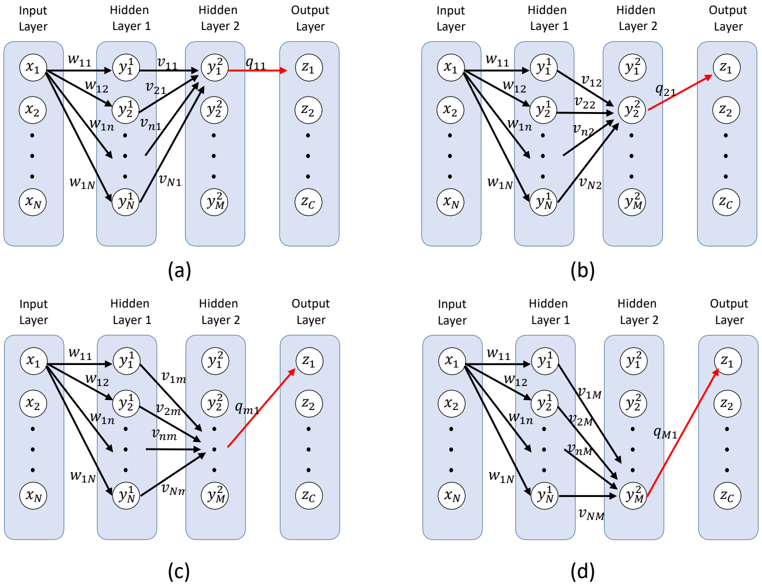


Figure 4-15 Schematic of a two hidden layer network demonstrating the weights that connect the first input node to the first output node. To simplify the illustration, the connection weights are broken into multiple schematics (a-d) by varying the last hidden layer m from 1 to M.

$$CWP_{x,z} = \sum_{m=1}^{M} \sum_{n=1}^{N} w_{xn} v_{nm} q_{mz} \qquad \text{Equation (21)}$$

Where $CWP_{x,z}$ is the connection weight product of an input metric x to a class output z, w_{xn} is the weight connecting an input metric x to a first hidden layer neuron n, v_{nm} is the weight connecting a first hidden layer neuron n to a second hidden layer neuron m, and q_{mz} is the weight connecting a second hidden neuron m to an output z. As demonstrated in Figure 4-5 and Equation (21), the new adaptation of the algorithm can be seen as computing and subsequently adding the original algorithm M times. Similarly, the calculation can be expanded to a general MLP ANN with L hidden layers as follows:

$$CWP_{x,z} = \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} \cdots \sum_{i_L=1}^{N_L} w_{xi_1}^{(0)} w_{i_1 i_2}^{(1)} \cdots w_{i_{L-1} i_L}^{(L-1)} w_{i_L z}^{(L)} \qquad \text{Equation (22)}$$

Where $w_{ij}^{(l)}$ is the weight connecting the i^{th} neuron in the l^{th} hidden layer to the j^{th} neuron in the $(l+1)^{th}$ layer. As with the original algorithm, the CWP can attain both positive and negative values, outlining the relative contribution of each input feature to each output in both magnitude and sign. The sign of the CWP indicates whether a high or a low feature value results in a higher probability of a certain class. CWPs can be further leveraged to obtain the relative importance of the features to each class by determining the ratio of the magnitude of a feature CWP to the sum of the magnitudes of all the features CWPs for that certain class.

In this study, the novel adaptation of the Connection Weights Algorithm was further validated by comparing its results with the permutation feature importance method, as previously outlined [16]. This method evaluates feature importance by observing the impact on model performance when a feature's values are randomly shuffled [23]. A feature is deemed important if model performance, assessed by the loss function and prediction accuracy, significantly worsens after permutation. Conversely, a negligible impact indicates a less important feature. This analysis, similar to a sensitivity analysis in engineering, was conducted using both training and testing sets for both the standalone ANN and the transfer learning ANN.

4.3.3 Results

4.3.3.1 Surgical Performance Metrics

The surgical performance metrics were categorized into motion, safety, and efficiency. Initially, 276 metrics were generated for each participant, but after feature selection and data augmentation, only 9 important metrics remained, primarily from the motion and safety categories (Table 4-21). This differs from the construct validity analysis in our validation studies [3]. These nine surgical performance metrics served as inputs for the developed ANNs, which had the following architectures:

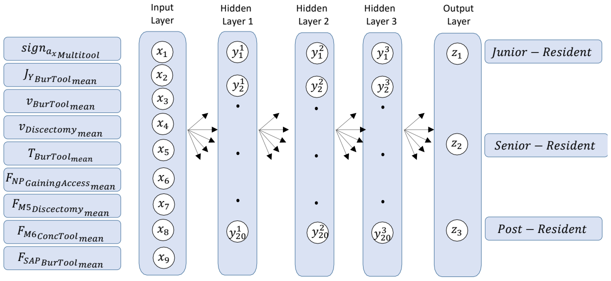

Input Layer
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Output Layer
$sign_{a_{x\,Multitool}}$
$J_{Y\,BurTool_{mean}}$
$v_{BurTool_{mean}}$
$v_{Discectomy_{mean}}$
$T_{BurTool_{mean}}$
$F_{NP\,GainingAccess_{mean}}$
$F_{M5\,Discectomy_{mean}}$
$F_{M6\,ConcTool_{mean}}$
$F_{SAP\,BurTool_{mean}}$
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9
y_1^1 y_2^1 y_{20}^1
y_1^2 y_2^2 y_{20}^2
y_1^3 y_2^3 y_{20}^3
z_1 z_2 z_3
$Junior - Resident$
$Senior - Resident$
$Post - Resident$

Figure 4-16 Model architecture of the final stand-alone MLP ANN model developed from scratch demonstrating the input surgical metrics, the number of hidden units and layers, as well as the output variables.

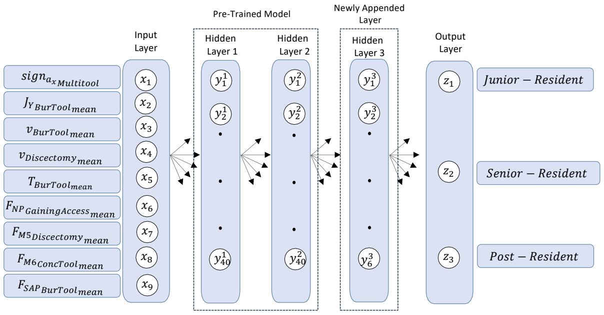


Figure 4-17 Model architecture of the final MLP ANN model developed from transfer learning demonstrating the input surgical metrics, the number of hidden units and layers, as well as the output variables.

4.3.3.2 Accuracy in Classification of Surgical Performance

The final standalone MLP model and the MLP with transfer learning were trained for 3500 and 890 epochs, respectively. Their classification accuracies are detailed in Table 4-26, with performance visualized in confusion matrices (Figure 4-18 and Figure 4-19). A confusion matrix provides a visual representation of an ANN's performance. For both models, matrices were generated for training (22 participants), validation (6 participants), and testing sets (8 participants). The standalone MLP achieved 100%, 100%, and 75% accuracies across these sets, while the MLP with transfer learning attained 95.45%, 100%, and 87.5%, respectively.

Table 4-26 Accuracy performance of the trained model on the training set, validation set, and testing set.

Model	No. of Training Epochs	Training Accuracy (%)	Validation Accuracy (%)	Testing Accuracy (%)
Stand Alone Model	3500	**100**	**100**	**75**
Transfer Learning Model	890	**95.45**	**100**	**87.5**

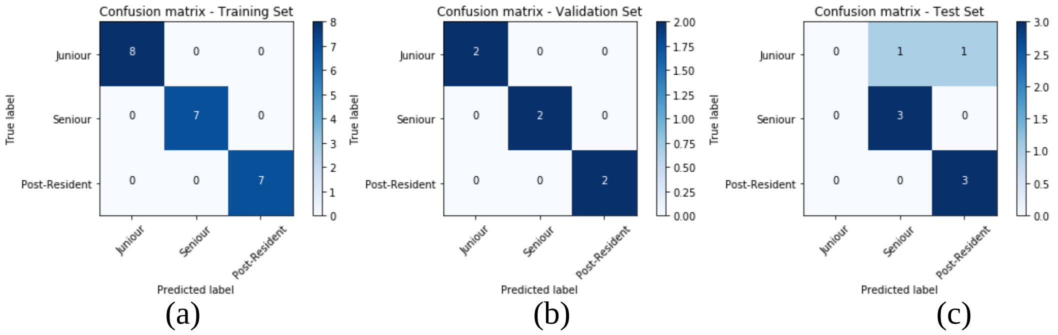
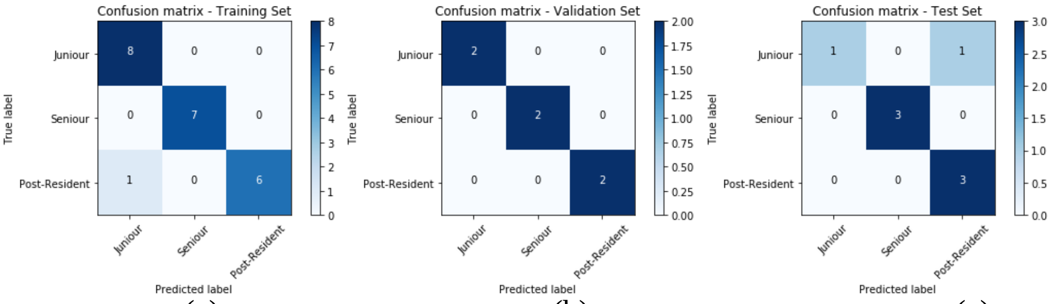


Figure 4-18 Confusion matrices highlighting the performance of the stand alone MLP ANN model trained from scratch on the: (a) training set, (b) validation set, and (c) testing set.



(a) (b) (c)

Figure 4-19 Confusion matrices highlighting the performance of the MLP ANN model with transfer learning on the: (a) training set, (b) validation set, and (c) testing set.

4.3.3.3 Surgical Performance Metrics Importance

This study adapted the Connection Weights Algorithm for multilayered ANNs and applied it to two MLP ANN architectures: one built from scratch and the other using transfer learning. The results were then compared to the permutation feature importance method. Table 4-27, Table 4-28, and Table 4-29 present the relative importance of the nine surgical performance metrics for both the standalone MLP ANN and the transfer learning MLP ANN. They detail the CWPs rankings and permutation feature importance results for both test and train sets across post-resident, senior-resident, and junior-resident groups. Notably, the CWP importance order varies for each surgical level. Table A 1 –Table A 10 in Appendix provide detailed CWP values, feature relative importance, and permutation feature importance for the training and testing sets for each surgical class group. Figure 4-20 presents the learning patterns that are exhibited in each input feature for the stand alone model, illustrating the CWPs for each feature across the three surgical levels.

Table 4-27 Surgical Performance Metrics Ranking for each model: CWPs & Permutation Importance for Junior Residents.

Stand-Alone MLP ANN Model				Transfer Learning MLP ANN Model		
Rank	CWP Rel. Imp.	Perm. Feat. Import. - Test Set	Perm. Feat. Import. - Train Set	CWP Rel. Imp	Perm. Feat. Import. - Test Set	Perm. Feat. Import. - Train Set
1	$F_{M5_{Discectomy_{m}}}$	$F_{M5_{Discectomy_{mean}}}$	$F_{SAP_{BurTool_{mean}}}$	$F_{NP_{GainingAccess_{n}}}$	$F_{SAP_{BurTool_{mean}}}$	$F_{SAP_{BurTool_{mean}}}$
2	$v_{Discectomy_{mear}}$	$F_{NP_{GainingAccess_{mec}}}$	$F_{M5_{Discectomy_{mean}}}$	$F_{M5_{Discectomy_{mec}}}$	$F_{M6_{ConcTool_{mean}}}$	$F_{M6_{ConcTool_{mean}}}$
3	$F_{SAP_{BurTool_{mea}}}$	$F_{M6_{ConcTool_{mean}}}$	$F_{M6_{ConcTool_{mean}}}$	$v_{Discectomy_{mean}}$	$F_{M5_{Discectomy_{mec}}}$	$F_{M5_{Discectomy_{mea}}}$
4	$v_{BurTool_{mean}}$	$F_{SAP_{BurTool_{mean}}}$	$F_{NP_{GainingAccess_{mea}}}$	$F_{SAP_{BurTool_{mean}}}$	$F_{NP_{GainingAccess_{n}}}$	$F_{NP_{GainingAccess_{n}}}$
5	$J_{Y_{BurTool_{mean}}}$	$T_{BurTool_{mean}}$	$T_{BurTool_{mean}}$	$sign_{a_{x_{Multitool}}}$	$J_{Y_{BurTool_{mean}}}$	$T_{BurTool_{mean}}$
6	$sign_{a_{x_{Multitool}}}$	$v_{Discectomy_{mean}}$	$v_{Discectomy_{mean}}$	$J_{Y_{BurTool_{mean}}}$	$T_{BurTool_{mean}}$	$v_{Discectomy_{mean}}$
7	$T_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$
8	$F_{NP_{GainingAcces}}$	$J_{Y_{BurTool_{mean}}}$	$J_{Y_{BurTool_{mean}}}$	$F_{M6_{ConcTool_{mean}}}$	$v_{Discectomy_{mean}}$	$J_{Y_{BurTool_{mean}}}$
9	$F_{M6_{ConcTool_{mea}}}$	$sign_{a_{x_{Multitool}}}$	$sign_{a_{x_{Multitool}}}$	$T_{BurTool_{mean}}$	$sign_{a_{x_{Multitool}}}$	$sign_{a_{x_{Multitool}}}$

Table 4-28 Surgical Performance Metrics Ranking for each model: CWPs & Permutation Importance for Senior-Residents

Stand-Alone MLP ANN Model				Transfer Learning MLP ANN Model		
Rank	CWP Rel. Imp.	Perm. Feat. Import. - Test Set	Perm. Feat. Import. - Train Set	CWP Rel. Imp	Perm. Feat. Import. - Test Set	Perm. Feat. Import. - Train Set
1	$F_{M5_{Discectomy_{m}}}$	$F_{M5_{Discectomy_{mean}}}$	$F_{SAP_{BurTool_{mean}}}$	$sign_{a_{x_{Multitool}}}$	$F_{SAP_{BurTool_{mean}}}$	$F_{SAP_{BurTool_{mean}}}$
2	$F_{SAP_{BurTool_{mea}}}$	$F_{NP_{GainingAccess_{mec}}}$	$F_{M5_{Discectomy_{mean}}}$	$J_{Y_{BurTool_{mean}}}$	$F_{M6_{ConcTool_{mean}}}$	$F_{M6_{ConcTool_{mean}}}$
3	$J_{Y_{BurTool_{mean}}}$	$F_{M6_{ConcTool_{mean}}}$	$F_{M6_{ConcTool_{mean}}}$	$F_{NP_{GainingAccess_{n}}}$	$F_{M5_{Discectomy_{mec}}}$	$F_{M5_{Discectomy_{mea}}}$
4	$F_{NP_{GainingAcces}}$	$F_{SAP_{BurTool_{mean}}}$	$F_{NP_{GainingAccess_{mea}}}$	$F_{SAP_{BurTool_{mean}}}$	$F_{NP_{GainingAccess_{n}}}$	$F_{NP_{GainingAccess_{n}}}$
5	$F_{M6_{ConcTool_{mea}}}$	$T_{BurTool_{mean}}$	$T_{BurTool_{mean}}$	$F_{M5_{Discectomy_{mec}}}$	$J_{Y_{BurTool_{mean}}}$	$T_{BurTool_{mean}}$
6	$v_{BurTool_{mean}}$	$v_{Discectomy_{mean}}$	$v_{Discectomy_{mean}}$	$v_{Discectomy_{mean}}$	$T_{BurTool_{mean}}$	$v_{Discectomy_{mean}}$
7	$T_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$
8	$sign_{a_{x_{Multitool}}}$	$J_{Y_{BurTool_{mean}}}$	$J_{Y_{BurTool_{mean}}}$	$T_{BurTool_{mean}}$	$v_{Discectomy_{mean}}$	$J_{Y_{BurTool_{mean}}}$
9	$v_{Discectomy_{mear}}$	$sign_{a_{x_{Multitool}}}$	$sign_{a_{x_{Multitool}}}$	$F_{M6_{ConcTool_{mean}}}$	$sign_{a_{x_{Multitool}}}$	$sign_{a_{x_{Multitool}}}$

Table 4-29 Surgical Performance Metrics Ranking for each model: CWPs & Permutation Importance for Post-Residents

	Stand-Alone MLP ANN Model			Transfer Learning MLP ANN Model		
Rank	CWP Rel. Imp.	Perm. Feat. Import. - Test Set	Perm. Feat. Import. - Train Set	CWP Rel. Imp	Perm. Feat. Import. - Test Set	Perm. Feat. Import. - Train Set
1	$v_{Discectomy_{mea}}$	$F_{M5_{Discectomy_{mean}}}$	$F_{SAP_{BurTool_{mean}}}$	$J_{Y_{BurTool_{mean}}}$	$F_{SAP_{BurTool_{mean}}}$	$F_{SAP_{BurTool_{mean}}}$
2	$F_{NP_{GainingAcces}}$	$F_{NP_{GainingAccess_{me}}}$	$F_{M5_{Discectomy_{mean}}}$	$v_{Discectomy_{mean}}$	$F_{M6_{ConcTool_{mean}}}$	$F_{M6_{ConcTool_{mean}}}$
3	$v_{BurTool_{mean}}$	$F_{M6_{ConcTool_{mean}}}$	$F_{M6_{ConcTool_{mean}}}$	$F_{NP_{GainingAccess_{n}}}$	$F_{M5_{Discectomy_{mea}}}$	$F_{M5_{Discectomy_{mea}}}$
4	$J_{Y_{BurTool_{mean}}}$	$F_{SAP_{BurTool_{mean}}}$	$F_{NP_{GainingAccess_{mea}}}$	$F_{M5_{Discectomy_{mea}}}$	$F_{NP_{GainingAccess_{n}}}$	$F_{NP_{GainingAccess_{n}}}$
5	$sign_{a_{x_{Multitool}}}$	$T_{BurTool_{mean}}$	$T_{BurTool_{mean}}$	$sign_{a_{x_{Multitool}}}$	$J_{Y_{BurTool_{mean}}}$	$T_{BurTool_{mean}}$
6	$F_{M6_{ConcTool_{mea}}}$	$v_{Discectomy_{mean}}$	$v_{Discectomy_{mean}}$	$v_{BurTool_{mean}}$	$T_{BurTool_{mean}}$	$v_{Discectomy_{mean}}$
7	$T_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$T_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$
8	$F_{SAP_{BurTool_{mea}}}$	$J_{Y_{BurTool_{mean}}}$	$J_{Y_{BurTool_{mean}}}$	$F_{SAP_{BurTool_{mean}}}$	$v_{Discectomy_{mean}}$	$J_{Y_{BurTool_{mean}}}$
9	$F_{M5_{Discectomy_{m}}}$	$sign_{a_{x_{Multitool}}}$	$sign_{a_{x_{Multitool}}}$	$F_{M6_{ConcTool_{mean}}}$	$sign_{a_{x_{Multitool}}}$	$sign_{a_{x_{Multitool}}}$

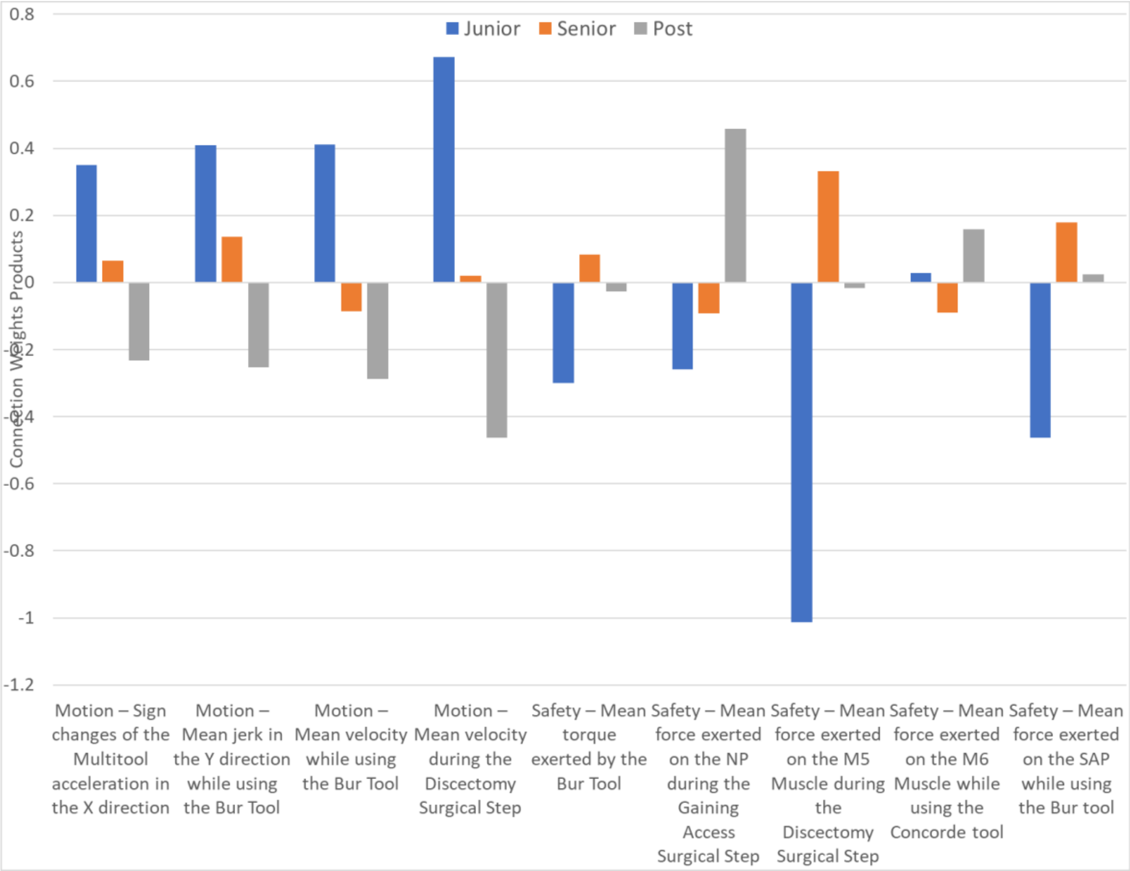


Figure 4-20 Learning patterns of the Connection Weights Products for each input feature on the Stand-Alone MLP ANN.

4.3.4 Discussion

4.3.4.1 Performance of the MLP ANN Models

The first objective of the study was to classify surgical performance and identify the relative importance of surgical performance metrics on the novel OLLIF AR/VR simulator. Focusing on the “gaining access” and “facetectomy and discectomy” steps of the OLLIF simulation, this study identified nine critical features for neural network development. Using the methodology shown in Figure 4-10, two MLP neural networks were successfully trained: one from scratch and another using transfer learning. Both models achieved high accuracy in classifying the three surgical classes, performing well on training (standalone: 100%, transfer learning: 95.45%), validation (both models: 100%), and testing sets (standalone: 75%, transfer learning: 87.5%). These results are within the 65% to 97.6% accuracy range reported in previous studies using machine learning for virtual surgical performance classification [8, 11, 16, 24, 25].

Analysis of the misclassified points in both models revealed some insights pertaining to the general applicability of the Connection Weights Algorithm on multilayered neural networks. More specifically, the developed equation was extended for three-layered neural networks to be applied on both the model developed from scratch and the one using transfer learning. The serendipitous fact that the optimal models in both cases led to three layered networks allow for a better comparison of the algorithm by removing the number of hidden layers as an influential factor. Both models share one misclassified junior-resident participant as a post-resident, while the stand-alone model had another misclassified junior-resident as a senior-resident. Using the CWPs from the standalone model (Table 4-9 – A 3), it was observed that the two misclassified junior-resident individuals exhibited performance traits that resembled senior and post-residents in key overlapping features (Table 4-14). For the junior participant that was misclassified as a senior, the participant had positive scores in the mean force applied on the M5 muscle during discectomy (z-score of 0.93) and the mean force applied on the superior articular process (SAP) while using the Burr tool (z-score of 0.48). The participant that was misclassified as a post-resident had negative scores in the average velocity during the discectomy step (z-score of -1.10) and the average velocity while using the Burr tool (z-score of -1.19). The z-scores specify the number of standard deviations the surgical performance is from the mean values of each feature. Thus, the first individual applied higher than average forces on both the M5 muscle during discectomy and the SAP while using the Burr tool; while the other misclassified individual had lower than average velocities during the discectomy step and specifically while using the Burr tool. Based on the CWPs, one interpretation is that these values might increase the likelihood of a senior and post resident classification, respectively, while they reduce the likelihood of a junior resident classification (Table 4-14). A similar analysis was seen in our previous publication when trying to uncover reasons behind misclassifications in multilayered neural networks [16].

However, conducting a similar analysis with the transfer learning model revealed different insights. Despite the individual's z-scores aligning with the junior-resident group CWPs, a misclassification still occurred. A reasonable explanation may be the fact that the two pre-trained and transferred layers were frozen during training, thereby limiting the network to adapt to the actual input features in both sign and magnitude. Transfer learning models with frozen pre-trained layers typically act as feature generators, transforming input features into new high-level metrics. This would mean that the CWPs of such models adapt to the new generated features rather than the actual inputs. While the magnitude of the CWP still indicates the relative importance of input features in these models, as discussed in the next sections, the interpretation related to the sign of the CWPs becomes less clear.

Table 4-30 Misclassified Participants' Surgical Performance Scores: comparison using CWPs from Standalone and Transfer Learning Models, highlighting divergence from Junior Group and limitations in frozen-layers Transfer Learning Model.

Misclassified Participant	Model	Category	Metric	Score	Junior: CWP (%Importance)	Senior/Post: CWP (%Importance)
Junior as senior-resident	Stand-Alone	Safety	$F_{M5_{Discectomy_{mean}}}$	0.93	-1.01 (25.92%)	0.332 (30.68%)
		Safety	$F_{SAP_{BurTool_{mean}}}$	0.48	-0.463 (11.86%)	0.179 (16.5%)
Junior as post-resident	Stand-Alone	Motion	$v_{Discectomy_{mean}}$	-1.10	0.672 (17.20%)	-0.4631 (24.10%)
		Motion	$v_{BurTool_{mean}}$	-1.19	0.411 (10.53%)	-0.288 (15.00%)
Junior as post-resident	Transfer Learning	Motion	$v_{Discectomy_{mean}}$	-1.10	-0.45 (18.7%)	0.47 (17.20%)
		Safety	$F_{NP_{GainingAccess_{mean}}}$	-0.85	-0.49 (20.31%)	0.44 (16.37%)
		Safety	$F_{M5_{Discectomy_{mean}}}$	-0.63	-0.48 (19.97%)	0.36 (13.16%)

4.3.4.2 Insights and Surgical Performance Patterns Revealed by the ANNs

Table 4-27 to Table 4-29 summarize the selected surgical performance features used in training and testing the optimal models, ranking them by importance as determined by the Connection Weights Algorithm (CWA) and validated by the Permutation Feature Importance algorithm on both testing and training sets. This approach was applied to both stand-alone and transfer learning models, offering a comprehensive view of feature significance in classification. While the permutation feature importance rankings remain consistent across the tables, variations in the CWP columns reflect class-specific calculations for Junior, Senior, and Post-resident groups. This differentiation emphasizes the unique influence of each feature on the respective surgical classes as defined by the CWPs and highlights the importance of a detailed and nuanced approach in interpreting the results, given the inherent performance variability between the classes.

This study utilized the CWA to uncover insights from neural network models classifying virtual OLLIF surgical performance. This objective was accomplished by extending the previously developed method by the authors to apply the CWA on multilayered neural networks to further assess its validity. The CWA evaluates the impact of each surgical performance metric (input feature) on different surgical levels (classes) by assigning weights for each feature-class pair, calculated by summing the products of connection weights from inputs to outputs [14]. These weights, known as Connection Weights Products (CWPs), help determine the relative importance of features for each surgical class. The algorithm's value lies in its ability to quantify each input feature's contribution to each output, both in magnitude and sign. For example, a positive (or negative) CWP indicates that a higher (or lower) than average feature value correlates with a specific class. The detailed CWPs values and their percent of relative importance for both models are comprehensively summarized in the Appendix (Table 4-9 – A 6).

To verify the results and validate the applicability of the CWA on both model types, the permutation feature importance algorithm was developed and applied to each of the two models on both the training and testing sets. Permuting both the training and testing sets can give different insights on aspects of surgical performance and the associated classifications. When applied on the training set, the permutation feature importance underscores the performance metrics that are seen important during the learning phase of the models. It highlights the features that the model used in building the connections between surgical performance metrics and surgical classifications. Conversely, when applied on the testing set, the algorithm brings to light the pivotal features enabling the model to perform well on unseen data. It points out the features that the model relies on when formulating new predictions. This comparative approach of applying the algorithm on both the training and testing sets underscores the true importance of metrics in both the model's learning and prediction phases. The detailed results of the drop in accuracies of each of the models when the training and testing sets are permuted can be see in the Appendix (Table A 7– A 10).

4.3.4.2.1 Insights to the Identified Feature Importance

A number of insights can be drawn from the chosen analysis frameworks of feature importance applied on the models and defined by the CWA and the permutation feature importance algorithm. The following section starts with an overview of the commonality seen in the analyses and then delves into the intricacies of each model-algorithm combination.

Table 4-27 to Table 4-29 reveal a common thread of features ranked as the most important across each model (stand-alone vs transfer learning models) and method (CWA vs permutation feature importance), indicating robust findings. Force-related features such as $F_{SAP_{BurTool_{mean}}}$, $F_{M5_{Discectomy_{mean}}}$, $F_{M6_{ConcTool_{mean}}}$, $F_{NP_{GainingAccess_{mean}}}$, are consistently identified as crucial metrics, emphasizing their crucial role in differentiating surgical proficiency levels. Similarly, the velocity features, define by $v_{Discectomy_{mean}}$ and $v_{BurTool_{mean}}$, are also seen significant across different models and methods, highlighting their impact on surgical performance. This convergence of crucial features across diverse analytical frameworks not only underscores the reliability of our results but also sheds light on the interrelation between force and velocity metrics, offering a more comprehensive view on aspects of surgical composites that distinguishes expertise.

The permutation feature importance algorithm, applied to both training and testing sets, showed notable uniformity in feature rankings for both the Stand-Alone and Transfer Learning MLP ANN Models. This uniformity indicates a consistent representation of feature importance across different model configurations, demonstrating the robustness and critical role of the selected surgical performance features in accurate classification. Additionally, it further supports the overall reliability and validity of the models in classifying virtual OLLIF surgical performance. Furthermore, the consistent results reinforce the use of the permutation feature importance algorithm as a gold standard for comparing and validating the application of the CWA on multilayered neural networks.

Analyzing the CWPs, both models show consistency in identifying important features for each surgical resident group. For junior-residents, top features like ${F_{M5}}_{{Discectomy}_{mean}}$, ${v_{Discectomy}}_{mean}$, and ${F_{SAP}}_{{BurTool}_{mean}}$ were consistently recognized in CWP rankings. Senior-residents' key features included ${J_Y}_{{BurTool}_{mean}}$, ${F_{NP}}_{{GainingAccess}_{mean}}$, and ${F_{SAP}}_{{BurTool}_{mean}}$, while post-residents focused on ${v_{Discectomy}}_{mean}$ and ${F_{NP}}_{{GainingAccess}_{mean}}$. However, as outlined in Section 4.3.4.1, CWPs from the transfer learning model don't indicate the directional impact (positive or negative) of features. More specifically, one cannot infer from a positive or negative CWP whether a class is likely to have higher or lower values for that respective feature, a conclusion made evident by the analysis of the misclassified individual using the CWPs from the transfer learning model (Table 4-14). Despite this, the CWP magnitudes retain their importance, accurately reflecting feature relevance to each class. This relevance in magnitude, confirmed by the high consistency in key features identified by transfer learning CWPs, aligns with both the standalone model's CWPs and the permutation feature importance results. The consistency across the CWA results and the permutation feature importance affirms the reliability of insights acquired through the application of the CWA on both the stand-alone and transfer learning models.

4.3.4.2.2 Surgical Learning Patterns Through CWA

The CWP of the stand-alone model was pivotal in illustrating the distinctive aspects of surgical performances across the three surgical classes, as outlined in the previous sections. Not only did it accurately highlight the importance of performance features, verified by the permutation feature importance algorithm, but it was also able to justify the misclassifications, leveraging both the sign and magnitude of CWPs. Thus, the thorough insights from the CWPs may enhance the understanding of the complexities in surgical learning patterns and performance across various surgical proficiency levels, allowing for more informed and tailored instructional learning systems.

Figure 4-20 illustrates two learning patterns in surgical training: continuous and discontinuous, aligning with prior research [3, 8, 16, 26]. Continuous learning shows sequential skill improvement from junior to senior to post-resident levels, while discontinuous learning involves non-linear skill progression, with inconsistent senior resident performance [27}. The CWPs reveal that motion metrics and one safety metric, $F_{NP_{GainingAccess_{mean}}}$, follow a continuous learning pattern, whereas other safety metrics display a discontinuous pattern. In motion metrics, the junior resident group utilizes higher velocities during discectomy and specifically while using the Burr tool, as well as, using more sudden changes in direction while operating the multitool during access gaining and the Burr tool during discectomy. Post-residents, in contrast, use lower velocities and more controlled movements. This suggests trainees should aim for slower, controlled movements to enhance OLLIF surgical performance. The applied force to the nucleus pulposus (NP) during access $\left(F_{NP_{GainingAccess_{mean}}}\right)$ shows a continuous learning pattern, with post-residents exerting more force than senior and junior residents, indicative of more direct disc access. This suggests post-residents experience greater force at the end of the gaining access step, a crucial aspect since this phase lacks visual feedback, relying instead on tactile and somatosensory feedback for accurate navigation. Expert consultations confirm that the probe's goal during this step is to puncture through muscles and annulus, typically ending in the nucleus, aligning with post-residents' performances. This analysis emphasizes post-residents' approach as the performance benchmark. Therefore, for enhanced surgical performance, trainees may need to focus on developing somatosensory reflexes, using force feedback effectively during the gaining access step for precise disc navigation.

Figure 4-20 shows that the rest of the safety features display a discontinuous learning pattern, with variations in force and torque applications among junior, senior, and post-residents during OLLIF surgery. Compared to senior-residents, junior and post-residents apply lower forces on the M5 muscle and the SAP, and use lower torques when using the Burr tool during discectomy. Conversely, they exert more force, as compared to senior residents, on the M6 muscle using the Concorde tool. This pattern indicates an evolving surgical approach with experience gain. Post-residents, with more experience, use a refined approach, applying less force on the more superficial M5 muscle and SAP, and more force on the deeper M6 muscle [27]. This selective force use suggests an advanced understanding of anatomy and OLLIF procedural steps. Lower forces during early steps of the procedure on the M5 and SAP likely aim to preserve tissue integrity and to minimize tissue trauma while accessing deeper structures more precisely; on the other hand, the increased force using the Concorde tool on the M6 muscle in later surgery stages signifies a strategic approach to effectively navigate and manage tissue resistance while engaging deeper tissues effectively [27, 28]. This systematic and methodical approach, reflective of their advanced training and experience, contrasts sharply with the less nuanced strategies of junior and senior residents, highlighting expertise differences. The discontinuous learning patterns, particularly between senior and post-residents, underscores the transformative refinement in surgical methodology that is typically honed over years of deliberate practice and experiential learning.

Each surgical class in the OLLIF surgery demonstrates distinct characteristics. Junior-residents show fast, less precise movements with cautious force use, reflecting their reluctant and beginner level. Senior residents, in an intermediate skill phase, exhibit more controlled movements but with variable force application. Post-residents, showcasing surgical expertise, perform deliberate, slow, and controlled movements with targeted force application, developed from extensive experience and deep anatomical knowledge. Thus, data mirroring these specific class traits would likely be classified accordingly, as was shown previously by our group [3, 8, 16, 26]. This understanding explains the misclassifications in the stand-alone model, where one individual's higher force application resembled senior residents and another's lower velocities mirrored post-residents.

4.3.4.2.3 Intelligent AI Surgical Tutors

The trend towards developing AI-based intelligent tutor systems has emerged as an ideal complement to the proven ability of ML algorithms in accurately classifying performance as demonstrated in this study. Our group has highlighted the effectiveness of such systems in efficiently training residents by offering real-time performance feedback [10, 29]. These systems are designed to replicate the guidance of expert surgeons by providing immediate, action-specific assessments and addressing the associated risks. Building these systems can follow two strategies, as shown by Mirchi, et al. [10] and Yilmaz, et al. [29]: one employs an offline pre-trained ML model for assessment and feedback, while the other uses an algorithm that learns continuously from new data while giving feedback to trainees. However, a potential issue is 'negative training,' where residents might be trained to incorrect skill levels [30]. One method of overcoming this issue is validating the skills benchmarked by the ML algorithm, for instance, by using realistic physics-based forces, similar to those in our newly developed simulator.

4.3.5 Limitations

4.3.5.1 Overcoming Small Data Set Limitation

Addressing the limitations of a relatively small dataset collected from one university center was pivotal for the accuracy and generalizability of the models developed in this study. This study addressed this limitation by using a combination of data augmentation, feature selection, and transfer learning techniques.

Initially, the feature set was pruned, reducing it from 276 to 168 features, by removing those with zero or near-zero variance and those having high correlation. Subsequently, a first pass through the SFS algorithm fine-tuned the feature space once more from 168 to a focused set of highly relevant 16 features. Afterwards, data augmentation, specifically through data jittering, was integrated, designed to address both the small dataset limitation as well as the imbalanced classes.

A subsequent round of SFS was then applied, refining the feature set to the final nine key metrics, each critical in distinguishing surgical performances.

In this study, data jittering was chosen for its ability to introduce natural variability, reflecting variances like hand tremors and dexterity control seen in actual surgical scenarios. It also preserved the realistic values of surgical performance features, avoiding distribution skew and aligning well with pre-normalized data. This approach was more suited to the realistic dynamics of surgical performance than other methods like data scaling, which could introduce unrealistic force values due to haptic limitations. Combining data augmentation with the removal of redundant features significantly improved the model’s predictive accuracy, raising the validation accuracy on the SFS algorithm from 82% to 92%. This improvement underscored the efficacy of using data augmentation with feature selection to enhance model precision and reliability in applications where data is scarce.

Transfer learning acts as a strategic leverage, harnessing previously acquired knowledge from related tasks, refining and extending the utility of machine learning models especially in cases of limited data scenarios. This methodology is commonly manifested in two predominant forms: fine-tuning of pre-trained models and utilizing pre-trained models as feature generators. Fine-tuning involves adapting the pre-trained model to new data, progressively optimizing all layers, starting from the outermost layers to the deeper ones, based on the basis that the initial layers contain generic features applicable to related tasks. However, this approach often encounters overfitting issues especially when applied to shallow networks with constrained datasets. Conversely, the feature extractor method freezes the pre-existing layers of a trained model and appends new layers, acting as sophisticated filters, transforming input data into high-level features to enhance classifications. Given the specificity of the current application and the constraints in dataset size, this study embraced the feature extractor methodology, which allowed robust generalization, effectively mitigating the risk of overfitting associated with the relatively small dataset and less complex network architecture. In fact, the transfer learning model resulted in a lower training accuracy than the stand-alone model, implying that the model did overfit on the training set.

The incorporation of transfer learning improved the accuracy on the testing set, emphasizing its significant contribution in surgical simulation classifications, especially in situations where the novelty of the surgery limits participant availability. This enhancement was evident as one of the participants, who was misclassified in the stand-alone model, achieved accurate classification with the transfer learning model. A plausible interpretation is that the model, via transfer learning, generated a subtle, novel feature offering a more complex analysis of performances, although with reduced interpretability on the hidden insights. It's possible that this improved method of analysis helped detect the subtle differences in performances, leading to more correct classifications even when the performances are quite similar. This balance between accuracy and detailed insight highlights how important transfer learning can be in improving the exactness and trustworthiness of prediction models, especially when dealing with limited and specific datasets, like the ones used in advanced surgical simulations.

4.3.5.2 Connection Weights Algorithm Limitations

The study's findings reveal that while CWPs effectively determined feature impact in both sign and magnitude for the standalone model, they only indicated the magnitude of relative importance without discerning the sign for the transfer learning model. This discrepancy becomes clear when analyzing misclassified instances, highlighting the difficulty in applying the CWA to multilayered ANNs with frozen layers transferred from other models. The limited adaptability of the transfer learning model, due to its reliance on these frozen layers for feature generation, hindered its ability to adjust to novel surgical features. To test this observation, a future research direction could involve unfreezing and deeply fine-tuning all layers of the transfer learning model, enabling a more comprehensive comparison of its CWPs in both signs and magnitudes with those from the standalone model.

4.3.6 Conclusion

This study demonstrates the advantages of using MLP ANNs for classifying and analyzing surgical performance on a novel OLLIF surgical simulator. It highlighted the effectiveness of data augmentation and transfer learning in overcoming the challenges posed by small datasets typical of surgical simulators. Additionally, the study expanded on the authors' previous work by comparing the new approach's analyses with the gold standard permutation feature importance algorithm. Results indicate that this method is adaptable to deeper networks for determining feature importance, including assessing feature impact in both sign and magnitude. However, its effectiveness is limited to identifying feature importance when applied to transfer learning with frozen layers.

4.3.7 References

[1] M. Goldenberg and J. Y. Lee, "Surgical Education, Simulation, and Simulators-Updating the Concept of Validity," (in eng), *Curr Urol Rep,* vol. 19, no. 7, p. 52, May 17 2018. https://doi.org/10.1007/s11934-018-0799-7

[2] M. Alaker, G. R. Wynn, and T. Arulampalam, "Virtual reality training in laparoscopic surgery: A systematic review & meta-analysis," *International Journal of Surgery,* vol. 29, pp. 85-94, 2016/05/01/ 2016. https://doi.org/10.1016/j.ijsu.2016.03.034

[3] S. Alkadri, R. F. Del Maestro, and M. Driscoll, "Face, Content, and Construct Validity of a Novel VR/AR Surgical Simulator of a Minimally Invasive Spine Operation," *Medical & Biological Engineering & Computing,* In Press.

[4] K. El-Monajjed and M. Driscoll, "Analysis of Surgical Forces Required to Gain Access using a Probe for Minimally Invasive Spine Surgery via Cadaveric-based Experiments towards use in Training Simulators," *IEEE Transactions on Biomedical Engineering,* pp. 1-1, 2020. https://doi.org/10.1109/TBME.2020.2996980

[5] S. Alkadri, "Kinematic Study and Layout Design of a Haptic Device Mounted on a Spine Bench Model for Surgical Training," Undergraduate Honours Program - Mechanical Engineering, Mechanical Engineering, McGill University, 2018.

[6] N. Ledwos, N. Mirchi, V. Bissonnette, A. Winkler-Schwartz, R. Yilmaz, and R. F. J. O. N. Del Maestro, "Virtual reality anterior cervical discectomy and fusion simulation on the novel sim-ortho platform: validation studies," *Operative Neurosurgery,* 2020.

[7] H. Azarnoush *et al.*, "Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection," (in eng), *Int J Comput Assist Radiol Surg,* vol. 10, no. 5, pp. 603-18, May 2015. https://doi.org/10.1007/s11548-014-1091-z

[8] N. Mirchi *et al.*, "Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance," *Operative Neurosurgery,* vol. 19, no. 1, pp. 65-75, 2019. https://doi.org/10.1093/ons/opz359

[9] R. Sawaya *et al.*, "Development of a performance model for virtual reality tumor resections," (in English), *Journal of Neurosurgery,* vol. 131, no. 1, p. 192, 2018. https://doi.org/10.3171/2018.2.Jns172327

[10] N. Mirchi, V. Bissonnette, R. Yilmaz, N. Ledwos, A. Winkler-Schwartz, and R. F. Del Maestro, "The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine," *PLOS ONE,* vol. 15, no. 2, 2020. https://doi.org/10.1371/journal.pone.0229596

[11] A. Winkler-Schwartz *et al.*, "Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation," *JAMA Network Open,* vol. 2, no. 8, 2019. https://doi.org/10.1001/jamanetworkopen.2019.8363

[12] N. M. J. J. o. e. i. Nasrabadi, "Pattern recognition and machine learning," vol. 16, no. 4, p. 049901, 2007.

[13] J. Heaton, S. McElwee, J. Fraley, and J. Cannady, "Early stabilizing feature importance for TensorFlow deep neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 4618-4624.

[14] O. Ibrahim, "A comparison of methods for assessing the relative importance of input variables in artificial neural networks," *Journal of Applied Sciences Research,* vol. 9, no. 11, pp. 5692-5700, 2013.

[15] J. D. Olden and D. A. Jackson, "Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks," *Ecological Modelling,* vol. 154, no. 1, pp. 135-150, 2002/08/15/ 2002. https://doi.org/10.1016/S0304-3800(02)00064-9

[16] S. Alkadri *et al.*, "Utilizing a multilayer perceptron artificial neural network to assess a virtual reality surgical procedure," *Computers in Biology and Medicine*, vol. 136, p. 104770, 2021/09/01/ 2021. https://doi.org/10.1016/j.compbiomed.2021.104770

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[18] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," vol. 35, no. 5, pp. 1299-1312, 2016.

[19] A. Winkler-Schwartz *et al.*, "Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation," (in eng), *J Surg Educ*, vol. 76, no. 6, pp. 1681-1690, Nov-Dec 2019. https://doi.org/10.1016/j.jsurg.2019.05.015

[20] E. Von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. J. T. L. Vandenbroucke, "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies," vol. 370, no. 9596, pp. 1453-1457, 2007.

[21] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint*, 2019.

[22] S. Chintala. *DEEP LEARNING WITH PYTORCH: A 60 MINUTE BLITZ*. Available: https://pytorch.org/tutorials/beginner/deep_learning_60min_blitz.html#deep-learning-with-pytorch-a-60-minute-blitz

[23] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1-81, 2019.

[24] J. Chan *et al.*, "A systematic review of virtual reality for the assessment of technical skills in neurosurgery," vol. 51, no. 2, p. E15, 2021.

[25] E. Bilgic *et al.*, "Exploring the roles of artificial intelligence in surgical education: A scoping review," 2021.

[26] A. Reich *et al.*, "Artificial Neural Network Approach to Competency-Based Training " 2020.

[27] J. W. Park, H. S. Nam, S. K. Cho, H. J. Jung, B. J. Lee, and Y. J. A. o. r. m. Park, "Kambin's triangle approach of lumbar transforaminal epidural injection with spinal stenosis," vol. 35, no. 6, pp. 833-843, 2011.

[28] P. Kambin, L. J. C. O. Zhou, and R. Research, "Arthroscopic discectomy of the lumbar spine," vol. 337, pp. 49-57, 1997.

[29] R. Yilmaz *et al.*, "Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation," *npj Digital Medicine*, vol. 5, no. 1, p. 54, 2022/04/26 2022. 10.1038/s41746-022-00596-8

[30] A. M. Fazlollahi *et al.*, "Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial," *JAMA Network Open*, vol. 5, no. 2, pp. e2149008-e2149008, 2022. 10.1001/jamanetworkopen.2021.49008 %J JAMA Network Open

4.3.8 Appendix

Table A 1 Ranked Surgical Performance Metrics with Corresponding Weights and Relative Importance for Junior-Residents as defined by the Stand-Alone MLP ANN Model.

Rank	Category	Metric	Connection Weights Product	Relative Importance (%)
1	Safety	$F_{M5_{Discectomy_{mean}}}$	-1.0126	25.92%
2	Motion	$v_{Discectomy_{mean}}$	0.6722	17.20%
3	Safety	$F_{SAP_{BurTool_{mean}}}$	-0.4633	11.86%
4	Motion	$v_{BurTool_{mean}}$	0.4113	10.53%
5	Motion	$J_{Y_{BurTool_{mean}}}$	0.4087	10.46%
6	Motion	$sign_{a_{x_{Multitool}}}$	0.3507	8.97%
7	Safety	$T_{BurTool_{mean}}$	-0.2992	7.66%
8	Safety	$F_{NP_{GainingAccess_{mean}}}$	-0.2598	6.65%
9	Safety	$F_{M6_{ConcTool_{mean}}}$	0.0281	0.71%

Table A 2 Ranked Surgical Performance Metrics with Corresponding Weights and Relative Importance for Senior-Residents as defined by the Stand-Alone MLP ANN Model.

Rank	Category	Metric	Connection Weights Product	Relative Importance (%)
1	Safety	$F_{M5_{Discectomy_{mean}}}$	0.3327	30.68%
2	Safety	$F_{SAP_{BurTool_{mean}}}$	0.1790	16.50%
3	Motion	$J_{Y_{BurTool_{mean}}}$	0.1370	12.63%
4	Safety	$F_{NP_{GainingAccess_{mean}}}$	-0.0915	8.43%
5	Safety	$F_{M6_{ConcTool_{mean}}}$	-0.0903	8.32%
6	Motion	$v_{BurTool_{mean}}$	-0.0854	7.87%
7	Safety	$T_{BurTool_{mean}}$	0.0830	7.65%
8	Motion	$sign_{a_{x_{Multitool}}}$	0.0655	6.04%
9	Motion	$v_{Discectomy_{mean}}$	0.0200	1.85%

Table A 3Ranked Surgical Performance Metrics with Corresponding Weights and Relative Importance for Post-Residents as defined by the Stand-Alone MLP ANN Model.

Rank	Category	Metric	Connection Weights Product	Relative Importance (%)
1	Motion	$v_{Discectomy_{mean}}$	-0.4631	24.10%
2	Safety	$F_{NP_{GainingAccess_{mean}}}$	0.4581	23.84%
3	Motion	$v_{BurTool_{mean}}$	-0.2880	15%
4	Motion	$J_{Y_{BurTool_{mean}}}$	-0.2526	13.15%
5	Motion	$sign_{a_{x_{Multitool}}}$	-0.2322	12.08%
6	Safety	$F_{M6_{ConcTool_{mean}}}$	0.1586	8.25%
7	Safety	$T_{BurTool_{mean}}$	-0.0275	1.43%
8	Safety	$F_{SAP_{BurTool_{mean}}}$	0.0239	1.24%
9	Safety	$F_{M5_{Discectomy_{mean}}}$	-0.0171	0.89%

Table A 4 Ranked Surgical Performance Metrics with Corresponding Weights and Relative Importance for Junior-Residents as defined by the Transfer Learning MLP ANN Model.

Rank	Category	Metric	Connection Weights Product	Relative Importance (%)
1	Safety	$F_{NP_{GainingAccess_{mean}}}$	-0.4946	20.31%
2	Safety	$F_{M5_{Discectomy_{mean}}}$	-0.4862	19.97%
3	Motion	$v_{Discectomy_{mean}}$	-0.4553	18.7%
4	Safety	$F_{SAP_{BurTool_{mean}}}$	-0.3877	15.92%
5	Motion	$sign_{a_{x_{Multitool}}}$	0.2721	11.17%
6	Motion	$J_{Y_{BurTool_{mean}}}$	-0.213	8.75%
7	Motion	$v_{BurTool_{mean}}$	-0.1178	4.83%
8	Safety	$F_{M6_{ConcTool_{mean}}}$	-0.0059	0.24%
9	Safety	$T_{BurTool_{mean}}$	-0.0023	0.095%

Table A 5 Ranked Surgical Performance Metrics with Corresponding Weights and Relative Importance for Senior-Residents as defined by the Transfer Learning MLP ANN Model.

Rank	Category	Metric	Connection Weights Product	Relative Importance (%)
1	Motion	$sign_{a_{x_{Multitool}}}$	-0.5302	22.98%
2	Motion	$J_{Y_{BurTool_{mean}}}$	-0.3611	15.65%
3	Safety	$F_{NP_{GainingAccess_{mean}}}$	0.3484	15.10%
4	Safety	$F_{SAP_{BurTool_{mean}}}$	0.3409	14.78%
5	Safety	$F_{M5_{Discectomy_{mean}}}$	0.245	10.62%
6	Motion	$v_{Discectomy_{mean}}$	0.1582	6.86%
7	Motion	$v_{BurTool_{mean}}$	-0.1367	5.92%
8	Safety	$T_{BurTool_{mean}}$	-0.0992	4.30%
9	Safety	$F_{M6_{ConcTool_{mean}}}$	-0.0866	3.75%

Table A 6Ranked Surgical Performance Metrics with Corresponding Weights and Relative Importance for Post-Residents as defined by the Transfer Learning MLP ANN Model.

Rank	Category	Metric	Connection Weights Product	Relative Importance (%)
1	Motion	$J_{Y_{BurTool_{mean}}}$	0.5348	19.58%
2	Motion	$v_{Discectomy_{mean}}$	0.4699	17.20%
3	Safety	$F_{NP_{GainingAccess_{mean}}}$	0.4471	16.37%
4	Safety	$F_{M5_{Discectomy_{mean}}}$	0.3594	13.16%
5	Motion	$sign_{a_{x_{Multitool}}}$	0.3458	12.67%
6	Motion	$v_{BurTool_{mean}}$	0.3098	11.34%
7	Safety	$T_{BurTool_{mean}}$	0.1189	4.35%
8	Safety	$F_{SAP_{BurTool_{mean}}}$	0.0769	2.81%
9	Safety	$F_{M6_{ConcTool_{mean}}}$	-0.0677	2.47%

Table A 7 Permutation Feature Importance applied on the training set with Stand-Alone MLP ANN Model.

Rank	Category	Metric	Prediction Accuracy(%)
1	Safety	$F_{SAP\,BurTool_{mean}}$	39.63%
2	Safety	$F_{M5\,Discectomy_{mean}}$	46.54%
3	Safety	$F_{M6\,ConcTool_{mean}}$	57.06%
4	Safety	$F_{NP\,GainingAccess_{mean}}$	61.83%
5	Safety	$T_{BurTool_{mean}}$	71.26%
6	Motion	$v_{Discectomy_{mean}}$	75.38%
7	Motion	$v_{BurTool_{mean}}$	91.07%
8	Motion	$J_{Y\,BurTool_{mean}}$	95.44%
9	Motion	$sign_{a_x\,Multitool}$	95.84%

Table A 8 Permutation Feature Importance applied on the testing set with Stand-Alone MLP ANN Model.

Rank	Category	Metric	Prediction Accuracy(%)
1	Safety	$F_{M5\,Discectomy_{mean}}$	36.36%
2	Safety	$F_{NP\,GainingAccess_{mean}}$	43.76%
3	Safety	$F_{M6\,ConcTool_{mean}}$	48.44%
4	Safety	$F_{SAP\,BurTool_{mean}}$	50%
5	Safety	$T_{BurTool_{mean}}$	50.01%
6	Motion	$v_{Discectomy_{mean}}$	57.86%
7	Motion	$v_{BurTool_{mean}}$	60.94%
8	Motion	$J_{Y\,BurTool_{mean}}$	62.61%
9	Motion	$sign_{a_x\,Multitool}$	73.45%

Table A 9 Permutation Feature Importance applied on the training set with Transfer Learning MLP ANN Model.

Rank	Category	Metric	Prediction Accuracy(%)
1	Safety	$F_{SAP\,BurTool_{mean}}$	28.11%
2	Safety	$F_{M6\,ConcTool_{mean}}$	45.15%
3	Safety	$F_{M5\,Discectomy_{mean}}$	49.57%
4	Safety	$F_{NP\,GainingAccess_{mean}}$	50.15%
5	Safety	$T_{BurTool_{mean}}$	50.54%
6	Motion	$v_{Discectomy_{mean}}$	53.93%
7	Motion	$v_{BurTool_{mean}}$	64.91%
8	Motion	$J_{Y\,BurTool_{mean}}$	71.67%
9	Motion	$sign_{a_x\,Multitool}$	87.16%

Table A 10 Permutation Feature Importance applied on the testing set with Transfer Learning MLP ANN Model.

Rank	Category	Metric	Prediction Accuracy(%)
1	Safety	${F_{SAP}}_{{BurTool}_{mean}}$	21.88%
2	Safety	${F_{M6}}_{{ConcTool}_{mean}}$	25%
3	Safety	${F_{M5}}_{{Discectomy}_{mean}}$	32.9%
4	Safety	${F_{NP}}_{{GainingAccess}_{mean}}$	53.18%
5	Safety	${J_{Y}}_{{BurTool}_{mean}}$	62.49%
6	Motion	$T_{{BurTool}_{mean}}$	67.12%
7	Motion	$v_{{BurTool}_{mean}}$	70.44%
8	Motion	$v_{{Discectomy}_{mean}}$	71.85%
9	Motion	$sign_{{a_x}_{Multitool}}$	75%

4.4 Conclusions on Articles Three & Four

Articles Three and Four were instrumental in attaining Objective 2 of this thesis, focusing on an ML study designed to accurately classify surgical performance on the new surgical simulator platform. This objective aimed not only to reinforce the construct validity established in Objective 1 but, more critically, to delve into aspects of surgical performance that defines surgical expertise. This exploration aligns with the overarching goal of transitioning from competency-based to expertise-based surgical training within the broader validation framework of the thesis.

Article Three laid the framework for the attainment of Objective 2, investigating the application of multilayered ANNs for classifying the complexities within the surgical performance domain, as well as, identifying the significance and impact of various performance features. A novel method was developed, extending the Connection Weight Algorithm to encompass deeper neural networks. This methodology was rigorously validated against the gold standard of permutation feature importance, marking a significant improvement in understanding feature importance in surgical performance. Furthermore, this article facilitated the development and training of an ANN, which, through the application of transfer learning, proved vital in Article Four, particularly in overcoming challenges associated with limited sample sizes.

Addressing Objective 2, Article Four expanded on the innovative method introduced in Article Three, revealing both its potential and its limitations. Importantly, it further demonstrated the ability of the novel simulator to distinguish between surgical levels. It also provided deeper insights into surgical performances related to the simulated OLLIF. Furthermore, Article Four outlined a comprehensive strategy for navigating the small sample size issue inherent in ML applications within VR/AR surgical simulations. This strategy involved a combination of data augmentation, feature selection, and the strategic use of transfer learning, providing a robust blueprint for future research in this field.

These studies resulted in the development and training of two distinct three-layered ANN models in Article Four to classify surgical performances on the simulator employed in this thesis: the first was a stand-alone model constructed from scratch; and the other leveraged transfer learning from the two-layered ANN model developed in Article Three. These models were pivotal in advancing towards achieving Objective 3, by providing an objective measure of the impact of using physics-based forces on surgical training as discussed in the next chapter.

Chapter 5. Study to Evaluate Importance of Physics-Based Force Feedback on Surgical Training

5.1 Background of Fifth Article

The manuscript presented in this chapter was designed to establish a methodology for objectively assessing the impact of physics-based haptic feedback on MI surgical training. It utilizes the ML models developed and refined in Chapter 4 to evaluate the influence of accurate haptic feedback during the "gaining access" step of the OLLIF surgical procedure outlined in this thesis. Notably, the "gaining access" step is unique within the surgical approach as it lacks visual feedback, necessitating surgeons to depend extensively on tactile sensations to navigate to and accurately reach the target disc. Prior to this study, the effect of employing precise physics-based haptic feedback in MI surgical simulations had not been directly quantified. To address this gap, this study analyzed cadaver-based force profiles during the "gaining access" phase and adjusted them to create new force profiles based on common biomechanical modeling assumptions. Subsequent recruitments allowed for the collection of data from surgeons experiencing both sets of force profiles before their performances were evaluated using the previously developed ML algorithms. The attainment of Objective 3 and hypothesis 3 are presented in the manuscript entitled "Impact of Physics-Based Force Feedback on Surgical Training and Performance in VR/AR Simulations", for which the contribution of the first author is considered to be 85%. This manuscript was submitted to the journal of Computers in Biology and Medicine.

5.2 Article 5 Impact of Physics-Based Force Feedback on Surgical Training and Performance in VR/AR Simulations

Sami Alkadri, Rolando Del Maestro, Mark Driscoll

Author names:

Sami Alkadri B.Eng., Ph.D Student [(1,3)], Rolando F. Del Maestro MD, PhD [(3)], Mark Driscoll, PEng., Ph.D., Associate Professor [(1,2)]

Institutional affiliations:

(1) Musculoskeletal Biomechanics Research Lab, Department of Mechanical Engineering, McGill University, Macdonald Engineering Building, 815 Sherbrooke St W, Montreal, H3A 2K7, QC, Canada.

(2) Orthopaedic Research Lab, Montreal General Hospital, 1650 Cedar Ave (LS1.409), Montreal, Quebec, Canada, H3G 1A4

(3) Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology & Neurosurgery, Montreal Neurological Institute, McGill University, 2200 Leo Pariseau, Suite 2210, Montreal, Quebec Canada, H2X 4B3.

Corresponding author:

Mark Driscoll

Mailing Address: Macdonald Engineering Building, RM 153, 817 Sherbrooke St W, Montreal, Quebec H3A 2K7, Canada

Phone: 514-398-6299

Fax: 514-398-7365

Email Address: mark.driscoll@mcgill.ca

ABSTRACT

Background

Minimally invasive spinal surgery requires precise haptic feedback, such as the sensation of puncturing different tissue layers, to accurately distinguish between tissues and access surgical sites. The complexity of biological tissues' non-linearity and viscoelasticity complicates this process. While current VR/AR surgical simulators aim to replicate complex force profiles to improve training realism, the haptic feedback's fidelity is often limited. Recent advances highlight the superiority of physics-based haptic feedback over geometric models, emphasizing the need for enhanced training realism.

Objective

To assess the impact of implementing physics-based haptic feedback, derived from cadaveric experiments, in a novel VR/AR surgical simulator for Oblique Lateral Lumbar Interbody Fusion.

Design

A controlled experimental design was used, involving the modification of cadaveric-based haptic feedback profiles using biomechanical theory, validated through the variance-accounted-for method. Six neurosurgical participants (2 post residents, 2 senior residents, 2 junior residents) participated, performing simulated surgeries with both original and altered force profiles. Artificial Neural Network models – previously trained for this VR/AR surgical simulator – were used to analyze and test classification robustness against variations in haptic feedback.

Setting

Neurosurgical Simulation and Artificial Intelligence Learning Centre, McGill University, Montreal, Canada.

Results

Realistic force profiles significantly improved classification accuracy of surgical performance. A one-sided paired t-test confirmed significant discrepancies in accuracy between realistic and modified non-realistic force profiles, especially in the transfer learning model.

Conclusions

This study supports the importance of realistic, physics-based haptic feedback in surgical simulation training, particularly in the “gaining access” phase of minimally invasive spinal surgery. Despite limitations, findings advocate for integrating accurate physics-based feedback to enhance training efficacy and prevent negative training.

Keywords

Physics-based haptic feedback, haptic fidelity, surgical simulation, minimally invasive spinal surgery, artificial neural network, transfer learning, surgical education.

Funding

NSERC Collaborative Research Development (CRD) Grant; Franco Di Giovanni Foundation; Brain Tumour Foundation of Canada Brain Tumour Research Grant; a Medical Education Research Grant from the Royal College of Physicians and Surgeons of Canada; and the Montreal Neurological Institute and Hospital.

5.2.1 Introduction

In the field of minimally invasive surgery (MIS), the role of force feedback, particularly the feeling of puncturing tissue layers, is fundamental to the surgeon's understanding of the environment [1]. This form of feedback allows surgeons to identify the unique feel associated with different types of tissues, providing valuable information that contributes to the safe and effective execution of surgical procedures [2]. More precisely, in minimally invasive spinal surgeries, surgeons rely on the somatosensory feedback in the access gaining step, to uniquely identify anatomical structures allowing them to reach and access the desired surgical location within soft and hard tissues of the spine [3]. The importance of these puncturing sensations stems from the inherent complexity of biological tissues, which exhibit nonlinear viscoelastic mechanical properties [4]. Nonlinearity and viscoelasticity in this context refer to the varying nonlinear responses of biological tissues to puncture forces at various depths and speeds [4]. For example, during a puncturing event, the exhibited tissue stiffness felt by the surgeon is not constant and leads to non-linear responses in the strain with the depth of penetration. Additionally, the same tissue can exhibit different stress-strain responses depending on the speed of puncture, indicating the viscoelastic nature of biological tissues [4]. In addition to nonlinearity and viscoelasticity, there are subtle yet sudden changes in the forces as the tool transitions and punctures both within a certain tissue layer and through different tissue layers. Mechanistically, the above takes form through tool tip tissue deformation until failure per layer followed by the combination of tool body friction. The combinations of these variations in forces provide surgeons with critical tactile cues about their progress in accessing the surgical site [2, 3, 5]. To that end, the realistic reproduction of these force profiles in surgical simulators, particularly virtual and augmented realty (VR/AR) simulators, is essential for training surgeons for real-world surgical scenarios. By accurately replicating these force sensations, VR/AR simulators can provide a more comprehensive and practical learning experience for surgical trainees, better preparing them for the diverse range of tactile sensations they will encounter in the operating room.

Haptic feedback is generally described as providing the user with both kinesthetic (forces and torques sensed by muscles, tendons, and joints) and tactile (vibrations sensed by mechanoreceptors on the skin) feedback resulting from the interactions between the virtual tool and components in the virtual scene [6]. The employment of haptic-feedback in surgical simulators has substantially impacted surgical practice learning curves as demonstrated by numerous studies [6, 7]. However, not all haptic-based simulators provide realistic force outputs [7]. Some state-of-the-art simulators utilize advanced voxel-based gaming engines and leverage haptic and auditory feedback based on geometric models to augment the experiential realism of the simulation [7]. Recent studies have highlighted the importance of using physics-based haptics rather than geometric models to ensure the accuracy and reliability of the generated force feedback and reliability of imparted training reactions. These studies highlighted the benefits of using cadaveric experiments to extract and implement realistic physics-based feedback [5]. Forces produced by simulators that employ discrete or heuristic methods, rather than those derived from constitutive modeling in continuum mechanics, may not accurately reflect or record the forces encountered in actual patient surgeries. This discrepancy could lead to participants reacting with forces that are not typically used in real-world situations [7, 8]. Naturally, this error presents a further limitation when utilizing training metrics, such as the forces applied by surgical tools, to evaluate and train surgical residents. In such cases, using the benchmark values identified by the simulator might be different to reality and thus resulting in negative training of junior surgical residents to wrong skill levels [9]. This discrepancy is further exacerbated in MIS whereby the applied forces are crucial for guidance as explained [6]. Nevertheless, even with the current evidence supporting the integration of physics-based haptic feedback in minimally invasive surgical training, there are no studies that directly measure the influence of physically accurate force profiles on simulation training. Thus, an objective measure of the impact of accurate physics-based haptic force feedback on surgical training is required.

The objective of this study is to assess the impact and importance of using physics-based haptic feedback derived from cadaveric experiments in a novel VR/AR Oblique Lateral Lumbar Interbody and Fusion (OLLIF) surgical simulator. A novel analytical method was used in this study that leverages trained machine learning algorithms previously developed to classify surgical performance levels with high accuracies on the mentioned simulator. To achieve the objectives, the cadaveric-based force profiles encoded within the given platform were firstly analyzed and significantly altered based on biomechanics principles. This was followed by a verification process to ensure sufficient and significant alterations were made prior to using the developed and pre-trained artificial neural networks (ANNs).

5.2.2 Material and Methods

The overall methodology of the current study revolves around measuring the impact of using accurate physics-based force profiles derived from cadaveric experiments on surgical training. The study utilizes a novel method to quantify this impact by leveraging previously developed machine learning algorithms. More precisely, artificial neural networks were previously developed and trained to classify surgical performances on a novel VR/AR surgical simulator with high accuracies [10]. One can examine the robustness of the identified connections made by the algorithm – in addition to measuring the impact of physics-based force metrics on surgical training – by varying the force profiles and recruiting new participants. More specifically, by varying the force-feedback generated by the haptic device, new surgical participants can be recruited to perform the virtual procedure and then subsequently inputting their new performance metrics to the machine learning algorithm. The change in the accuracy of the machine learning model can be a measure of the robustness of the model. No significant change in model accuracy would indicate that the use of physics-based force-feedback has no significant effect on virtual surgical performance. It would also indicate the robustness of the machine learning model. Conversely, a significant impact in the model accuracy would underscore the importance of the force metrics in identifying surgical expertise and hence the importance of using physics-based haptics in surgical training.

5.2.2.1 The VR/AR Simulator & The Simulated Scenario

This study utilized a novel VR/AR surgical training platform developed by our group at McGill University in collaboration with CAE Healthcare and Depuy Synthes part of Johnson & Johnson. Validation studies were performed and published for the simulator platform [11]. The simulator system comprises of a high-performance gaming laptop (i7-8750H) with Windows 10 operating system that displays the surgical site laparoscopic view, a flat panel monitor displaying the MRI view as well as the Anterior Oblique and the Lateral fluoroscopic views, a physics-based six-degrees of freedom ENTACT W3D haptic device, and a benchtop model (Figure 5-1 (a)). The surgical simulation under consideration is the OLLIF spinal surgery, in which the simulation focusses on three phases: gaining access through the back muscles, removing the intervertebral disc, and inserting graft and a spinal cage. The specific phases with the used surgical tools are displayed in Figure 5-1(b).

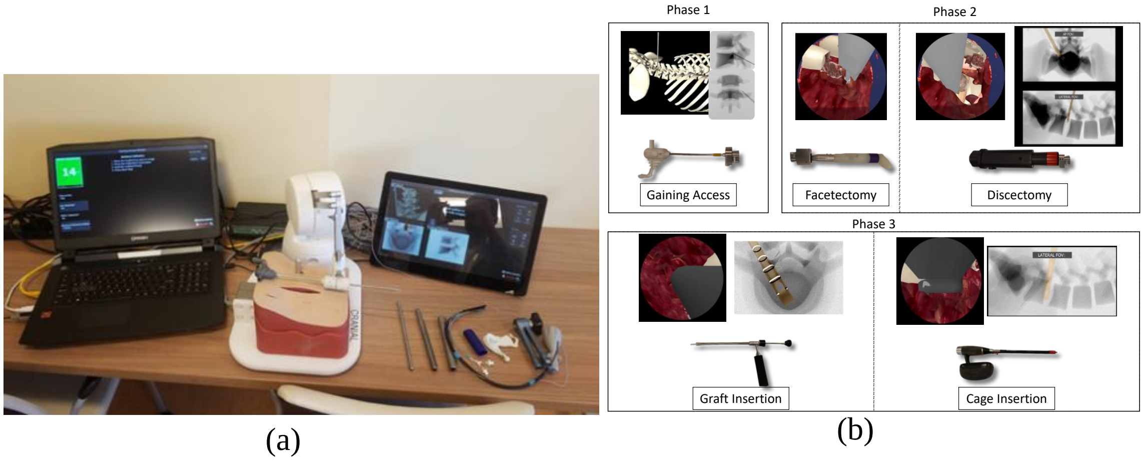


Figure 5-1 The summarized simulator layout. Left is the laptop displaying the laparoscopic view of the surgical site and indicates the instruction of the surgery process. The haptic device and benchtop model are in the middle. And right is the external display that displays the MRI view as well as the Anterior Oblique and the Lateral fluoroscopic views. The surgeon operates the haptic device based on the visual feedback from both monitors. (b) The three phases of the simulated surgery: Phase 1 includes gaining access to the disc using a Multitool; Phase 2 includes facetectomy using a Burr Tool followed by a discectomy using a Concord Tool; Phase 3 includes inserting graft followed by inserting a cage using the respective tools.

The current study focuses on the first phase of the OLLIF simulation demonstrated in Figure 5-1(b). The first phase of the simulated surgery requires the operator to use a multiprobe tool to gain access to the surgical area by puncturing through the tissue layers. The first step is the only step where no visual feedback is given to the surgeons, and therefore they heavily rely on the somatosensory feedback to identify the anatomical landmarks that aids them in reaching the target surgical area within the soft and hard tissue complex [12]. Consequently, this phase presents an optimal scenario to examine the hypothesis of the current study. Specifically, it offers a valuable opportunity to determine whether employing accurate physics-based force feedback is impactful in surgical training, with special focus on minimally invasive spinal surgeries.

5.2.2.1.1 The Gaining Access Phase

The force feedback generated by the haptic device in the simulation replicates the resistance provided by the instruments when penetrating through the muscles and connective tissues during an actual surgery. The force profiles incorporated in the current simulation are based on cadaveric experiments conducted by our group that extracted the tissue force responses [5]. The study explored the force responses of spinal tissue layers at the L4-L5 level, resulting in a comprehensive dictionary of forces during tool-tissue interactions [3]. More specifically, experiments were conducted to extract and generate a general haptic force framework that consisted of different force components namely, linear insertion, linear extraction, lateral resistance, and forces generated due to moments. The developed forces were carefully extracted into piece-wise functions, and subdivided wherever there were sharp drops due to puncturing events of the tissue layers to simulate the realistic behavior of physiological tissues during the interaction with the surgical tool. The force-displacement and force-time output were curve fitted using second-order polynomials (Equation (23) and the extracted coefficients are presented in Table 5-1. At each iteration the individual force components are calculated and combined to generate the appropriate force feedback [3]. In essence the force profiles attempt to simulate the physiological tissues' force responses, as well as the force drops associated with the puncturing of the different muscle layers as demonstrated in Figure 5-2 [13].

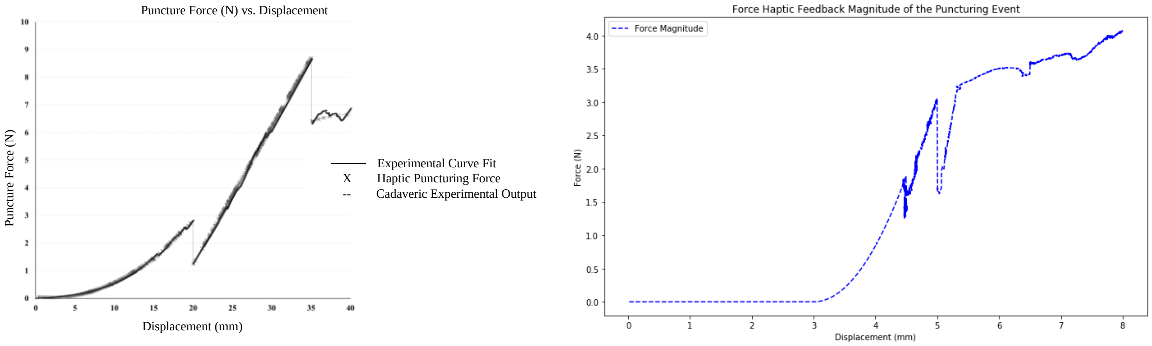
$$F(d) = \begin{cases} c_2^{(1)} d^2 + c_1^{(1)} d + c_0^{(1)};\ a_0 < d < a_1 \\ \vdots \\ c_2^{(n)} d^2 + c_1^{(n)} d + c_0^{(n)};\ a_n < d < a_{n+1} \end{cases} \qquad \text{Equation (23)}$$

Piece-wise second order formulation of the Force-Displacement curves used in developing the puncturing event force feedback (adapted from [3]).

Table 5-1 Second-order polynomial curve fitting coefficients (adapted from [3]).

Component	**2nd Order Polynomial Coefficients**			**Range (mm)**
	c_2	c_1	c_0	
Linear Insertion	0.0084	0	0	0-15
	0	0.2864	-2.8951	15-22
	0	0.4975	-8.7087	22-30
	-0.0108	0.859	-10.44	30-40
Lateral Resistance	0.0141	0.1877	0	0-30
Linear Extraction	-0.02	0	0	0-50

Figure 5-2 (a) presents the puncturing forces as measured in the cadaveric experiments, which are then used to firstly generate the curve fits experimentally before using the equations to generate the haptic force feedback in the simulation. Figure 5-2 (b) presents the force feedback generated by a previous participant using the equations described above. It can be seen that the force-displacement curves generated by the haptic device during a simulation run highly match the forces generated and extracted from the cadaveric experiments.



(a) (b)

Figure 5-2(a) The extracted force-displacement curves for the puncturing event for the actual cadaveric experiments, curve fitted output that was fed to the haptic device, and the resulting recorded haptic feedback force (adapted from [3]); (b) The resulting recorded haptic feedback force magnitude from a simulator run of the gaining access phase.

To achieve the objectives of this study, significant modifications to the cadaveric-based force profiles within the simulator are necessary. The aim is to simplify the force profiles by adopting assumptions commonly applied in creating simplified finite element models for biomechanical simulations. By using the small strains assumption, the linear approximation of the second-order piece-wise polynomial curves for each force component can be derived using the Taylor series expansion, stopping after the second term as follows:

$$f(x) = f(a_0) + f'(a_0)(x - a_0) \qquad \text{Equation (24)}$$

General formulation of the linear approximation of a general function f(x) about x=a_0.

As a demonstration, when applying Equation (24) to the linear insertion forces as described by Equation (23) and Table 5-1, and expanding the functions about the midpoint of each muscle layer, the following modified functions emerge:

$$F(d) = \begin{cases} 0.126d - 0.4725;\ about\ d = 7.5 \\ 0.2864d - 2.8951;\ N/A \\ 0.4975d - 8.7087;\ N/A \\ 0.103d + 2.79;\ about\ d = 35 \end{cases} \quad \text{Equation (25)}$$

Linear approximation of the linear insertion functions using Taylor series expansion about the middle point of each muscle layer.

In addition to the small strain assumption, another frequently used basis in biomechanical modelling is the homogeneity of mechanical properties in biological tissues [14]. By serendipity, leveraging this assumption in the current context would also eliminate the puncturing sensation, often considered crucial for guiding surgeons in the gaining access step of a MIS. As a result, a single equation is employed for each force component throughout the entire puncturing process, as illustrated in Table 5-2. For linear insertion, the steepest curve in Equation (25) was selected.

Table 5-2 Modified curve fitting coefficients to generate linear force feedback without puncturing sensations.

Component	**1st Order Linear Expansion Coefficients**				**Range (mm)**
	c_2	c_1	c_0	Expanded about	
Linear Insertion	0	0.4975	-8.7087	N/A	0-40
Lateral Resistance	0	0.6107	-3.1725	15	0-30
Linear Extraction	0	-1	12.5	25	0-50

5.2.2.2 Verification of Force Profile Modifications

A verification process is required to establish that the newly generated force profile has changed sufficiently. A previous study conducted by our group examined the ability of the operator to distinguish between different dynamic models that were developed based on the original force profile programmed in the simulator [15]. More specifically, the study computed the percent-variance-accounted-for (%VAF) between the newly generated dynamic models and the original cadaveric-based force profile model. When applying %VAF to compare force profiles, a high %VAF would suggest that the forces produced by the modified model are very similar to the forces produced by the original model [16]. It can be interpreted as the modified model being able to accurately replicate or reproduce the force characteristics of the original model. A lower %VAF value would imply that the modified model cannot adequately account for the variance in the original model, suggesting that there are significant differences between the force profiles, indicating a poor fit or dissimilarity [16]. Based on the results of the previous study, it can be shown that a %VAF of approximately 63% is enough for operators to perceive a force profile difference. The paper highlighted that the higher order (HO) and the Kelvin-Boltzmann (KB) models, each having a minimum 70% VAF to the original force profile, have significant differences as rated by the operators to the Maxwell (MW) model, having a 6.9% VAF. Therefore, one can extrapolate that a %VAF difference of 63% is sufficient for operators to qualitatively differentiate between force models. Thus, a verification methodology was constructed and applied to determine the extent to which the newly generated force profile in the current study differed to the original force profile. Specifically, the %VAF for the newly generated force profiles was calculated as follows [16]:

$$\%VAF = \left(1 - \frac{var\left(F_{original} - F_{modified}\right)}{var\left(F_{original}\right)}\right) \times 10 \qquad \text{Equation (26)}$$

Percent variance accounted for (%VAR) formula.

The force data recorded during the gaining access step were preprocessed and plotted for visual inspections in Python. As a first step, data were filtered to extract the range of datapoints once a significant change in the force magnitudes was observed to highlight the start of the trial, prior to the activation of the haptic forces. This was followed by a visual plotting of the force-time and displacement-time graphs to establish a threshold height where the forces first were activated as demonstrated by the following figures:

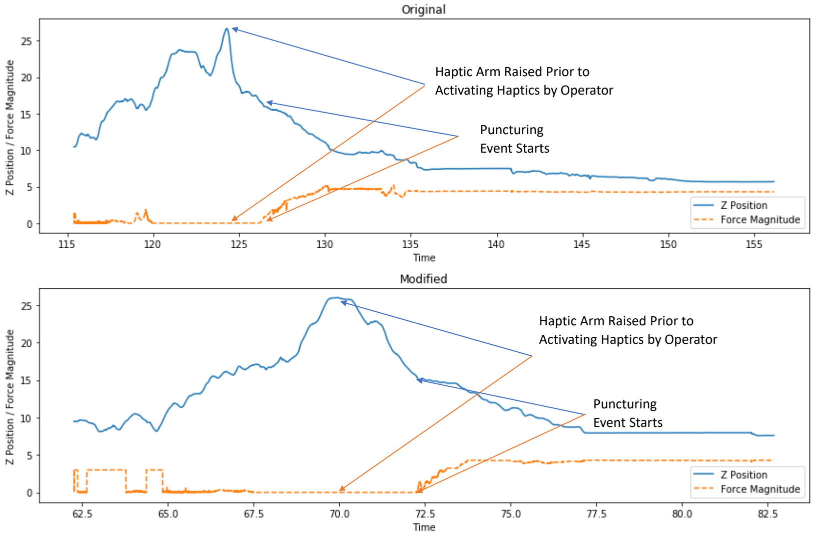


Figure 5-3 Force-time and displacement-time graphs of the original force profile based on cadaveric experiments (top), and the modified force profile generated in this study (bottom). The graphs show the gaining access step from the start of the simulation (prior to haptic feedback activation) to the reaching of the surgical area.

The above plots narrowed the start of the puncturing events to have a more accurate computation of the %VAF. As a result, the following force-time, displacement-time, and force-displacement graphs were generated in Figure 5-4 and Figure 5-5. To have a better comparison with the plots in Figure 5-2, the displacement array in Figure 5-5 was linearly transformed to measure the displacement starting from the top position of the puncturing point.

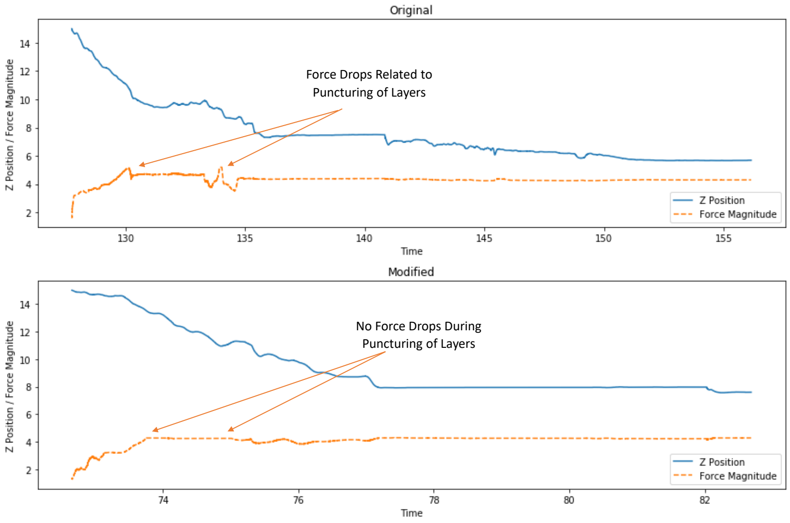


Figure 5-4 The force-time and displacement-time graphs of the puncturing events for the original (top) and the modified (bottom) force profiles. The graphs clearly highlight the presence of force drops associated with the original force profile and the absence of force drops in the modified force profile during puncturing of the layers.

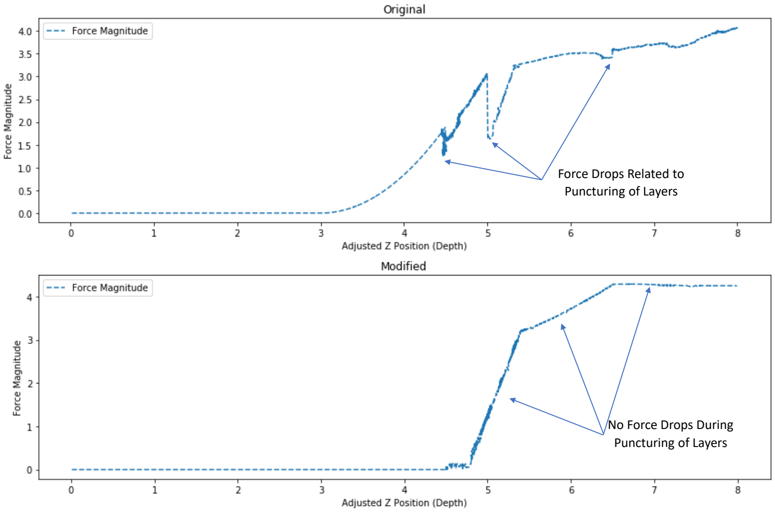


Figure 5-5 The force-displacement graphs of the puncturing event in the original force profile (top) and the modified force profile (bottom). The graphs highlight the change from the second-order piecewise curves to a first-order force response.

Using Equation (26), the %VAF was calculated using Python:

$$\%VAF = 53\%$$

Based on the above calculation in light of the previous study results, it can be deduced that the newly generated force profile is distinct enough for operators to detect a difference qualitatively. Hence, the task shifts to determine quantitatively if such changes in the force profiles are significant enough to impact virtual surgical performances.

5.2.2.3 Study Participants

Six participants were recruited to perform the virtual reality OLLIF scenario: 2 post-residents, 2 senior residents, and 2 junior residents. Table 3-1 presents the demographics of the participants. The participants were divided into three groups: A post-resident group (1 spine surgeon and 1 spine fellow), a Senior-Resident group (2 PGY 4-6 neurosurgery), and a Junior-Resident group (2 PGY 1-3 neurosurgery). This study was approved by an appropriate Research Ethics Board. All participants signed an approved written consent form prior to completing the simulation of the virtual spine surgery which took on average 30 minutes to complete. The recruited participants performed the simulation twice: once with the original physics-based force profiles derived from the cadaveric experiments and once with the non-realistic modified force profiles. Subsequently, the surgical performance metrics are split into two groups: one group for the simulations with realistic force profiles and one group for the simulations with non-realistic force profiles.

Table 5-3 Demographics of the post-resident, senior-resident, and junior-resident groups.

	Junior Residents	**Senior Residents**	**Post-Residents**
No. of individuals	2	2	2
Sex			
Male	1	2	2
Female	1	0	0

Surgical Specialty Level of Training	Neurosurgery	Orthopaedic Surgery
PGY 1-3	2	0
PGY 4-6	2	0
Fellows	1	0
Consultants	1	0

5.2.2.4 Machine Learning Model & Statistical Analysis

The current study leverages previously developed multilayered perceptron (MLP) artificial neural networks (ANNs) trained in classifying surgical performances on the current VR/AR simulator [10]. As demonstrated in multiple texts such as Bishop [17] and Goodfellow, et al. [18], MLP ANNs are a deeper subset of machine learning, which is described as the ability of algorithms to make classifications or decisions by identifying and learning from hidden patterns within datasets, without the need for explicit instructions. MLP ANNs have displayed potential in surgical simulation, owing to their capability to capture and replicate complex non-linear patterns in data collected during simulation tasks [8]. ANNs mirror biological neural networks in structure; they are composed of numerous linked neurons arranged in layers. Each layer processes data and relays it to the subsequent layer. The algorithm adaptively learns the weights linked to connections between nodes across different layers, aiming for a closer approximation of the true model. When integrated with VR/AR surgical simulators, this algorithm can enhance the precision of surgical performance classification.

Two distinct three-layered networks were trained: one was a three-layered MLP ANN built from scratch on the current simulator, while the other employed transfer learning from a pre-existing two layered ANN model. This latter model was developed and trained in a side study using the Sim-Ortho simulator – a VR simulator of an annulus incision task during an anterior cervical discectomy and fusion (ACDF) scenario developed by OSSimTech [8]. The models use 9 surgical performance features as inputs to the model to assign the performance as one of three surgical classes: junior, senior, or post-resident. The architectures and hyperparameters of the trained models, as well as their accuracies across training, validation, and testing sets, are displayed in Figure 5-6, Figure 5-7, Figure 5-8, and Figure 5-9.

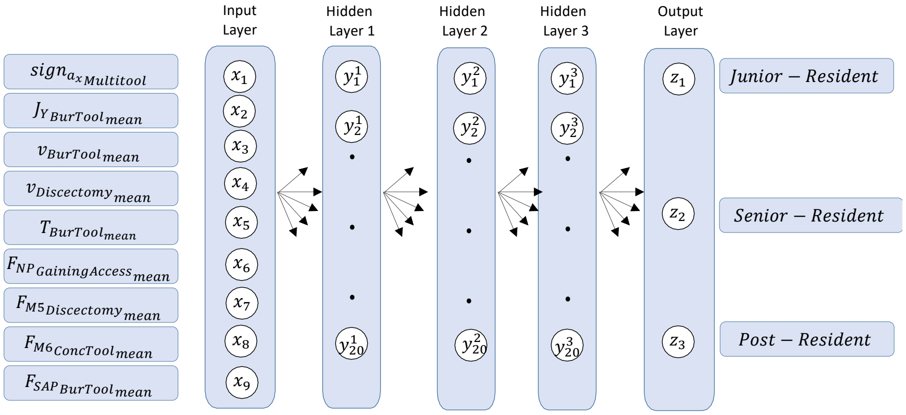


Figure 5-6 Model architecture of the final stand-alone MLP ANN model developed from scratch demonstrating the input surgical metrics, the number of hidden units and layers, as well as the output variables.

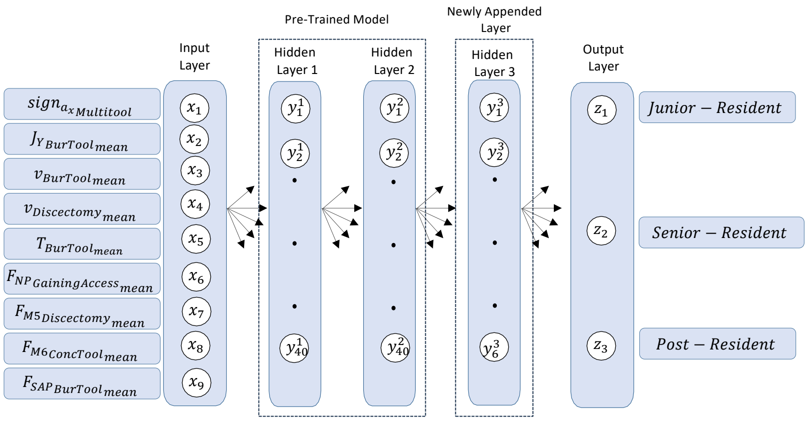


Figure 5-7 Model architecture of the final MLP ANN model developed from transfer learning demonstrating the input surgical metrics, the number of hidden units and layers, as well as the output variables.

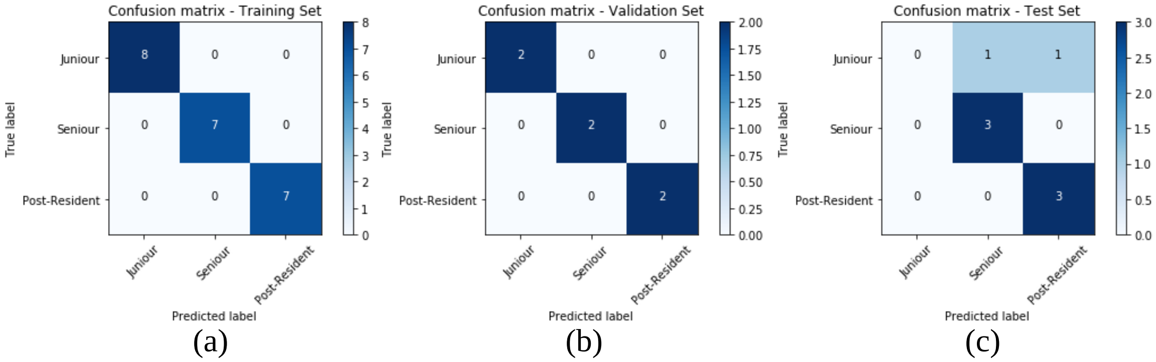


Figure 5-8 Confusion matrices highlighting the performance of the stand alone MLP ANN model trained from scratch on the: (a) training set, (b) validation set, and (c) testing set.

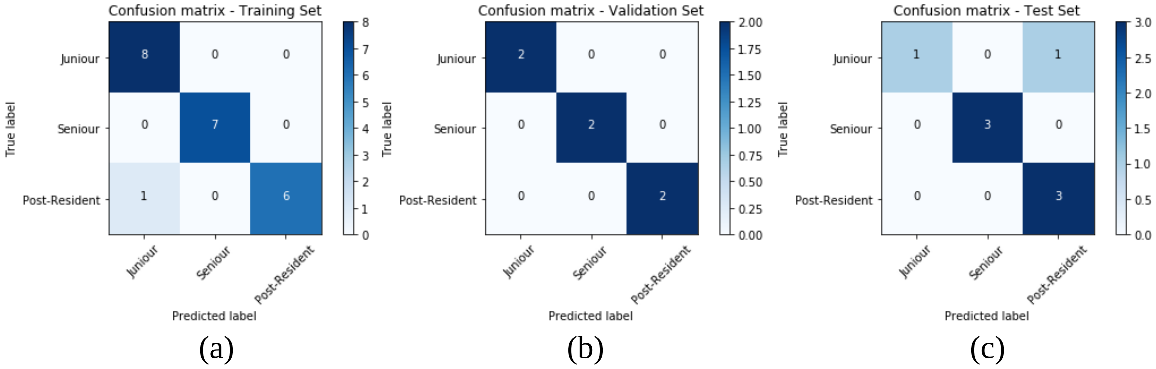


Figure 5-9 Confusion matrices highlighting the performance of the MLP ANN model with transfer learning on the: (a) training set, (b) validation set, and (c) testing set.

Table 5-4 The top three ranked surgical performance metrics for each of the surgical classes as measured by the CWPs and Permutation Feature Importance applied on both the testing and training sets on both trained models (adapted from [10]).

Stand-Alone MLP ANN Model				
Class	**Rank**	**CWP Rel. Imp.**	**Perm. Feat. Import. - Test Set**	**Perm. Feat. Import. - Train Set**
Junior-Resident	1	$F_{M5\,Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{SAP\,BurTool_{mean}}$
	2	$v_{Discectomy_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{M5\,Discectomy_{mean}}$
	3	$F_{SAP\,BurTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$
Senior-Resident	1	$F_{M5\,Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{SAP\,BurTool_{mean}}$
	2	$F_{SAP\,BurTool_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{M5\,Discectomy_{mean}}$
	3	$J_{Y\,BurTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$
Post-Resident	1	$v_{Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{SAP\,BurTool_{mean}}$
	2	$F_{NP\,GainingAccess_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{M5\,Discectomy_{mean}}$
	3	$v_{BurTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$
		Transfer Learning MLP ANN Model		
Junior-Resident	1	$F_{NP\,GainingAccess_{mean}}$	$F_{SAP\,BurTool_{mean}}$	$F_{SAP\,BurTool_{mean}}$
	2	$F_{M5\,Discectomy_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$
	3	$v_{Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$
Senior-Resident	1	$sign_{a_{x}\,Multitool}$	$F_{SAP\,BurTool_{mean}}$	$F_{SAP\,BurTool_{mean}}$
	2	$J_{Y\,BurTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$
	3	$F_{NP\,GainingAccess_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$
Post-Resident	1	$J_{Y\,BurTool_{mean}}$	$F_{SAP\,BurTool_{mean}}$	$F_{SAP\,BurTool_{mean}}$
	2	$v_{Discectomy_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$
	3	$F_{NP\,GainingAccess_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$

Features derived from the force profile during the gaining access step significantly influence the trained models’ capacity to classify surgical performances as illustrated in Table 5-4. Specifically, the number of sign changes in the x-direction of the MultiTool acceleration and the forces exerted on the Nucleus Pulposus (NP) by the MultiTool stand out as pivotal features in the classification procedure. These metrics appear at least once in the top three ranked features for each of the surgical classes as determined by the feature importance analyses used in the previous study, namely, the Connection Weights Algorithm and the Permutation Feature Importance. The

Connection Weights Algorithm determines the influence of individual metrics on the classification task in both magnitude and sign. Initially devised for single-layer networks, it has been adapted for multilayer networks, as detailed in our earlier works, to calculate the Connection Weights Product (CWP). This adaptation allowed for its application to the three-layer networks used in this study. On the other hand, the Permutation Feature Importance method assesses feature importance by measuring the degradation of model performance after random shuffling of a feature's values.

As outlined, six participants each performed the simulation twice: initially with the original force profiles derived from cadaveric experiments and then with the modified, non-realistic ones. Following the data collection from these participants, the dataset was divided into two groups: one representing the original, realistic force profile and the other, the modified non-realistic force profile. This data underwent identical preprocessing steps as those in our previous study, including z-score normalization based on the means and standard deviations of the features from the main trial. The normalized data was subsequently input into the developed machine learning algorithms, and accuracies and details of any misclassifications for both sets of performances were recorded. For the statistical analysis, aimed at identifying significant changes in the model accuracies between realistic and non-realistic force profiles, a tailored approach was employed to mitigate the limitations posed by the small sample size. This involved generating pseudo-independent, paired samples for paired statistical testing between the two performance groups. Bootstrap resampling with replacement was chosen for this purpose, ensuring that each resampling instance included the same participants for both the original and modified force profiles. This method facilitated a valid paired comparison. The resampled data from both realistic and non-realistic simulations were then reprocessed through the trained ANN models to produce the necessary number of accuracy points for the statistical analysis. A power study indicated the need for 6 to 18 accuracy points; hence, 18 bootstrap samples were executed for each force profile group. This power study, conducted for a paired one-sided t-test, was designed to detect a minimum 16.67% difference in model accuracy for classifying performances under realistic versus non-realistic force profiles. This was based on a standard deviation in model accuracy differences ranging from 10% to 23%. The chosen threshold for minimum accuracy change was determined by the number of recruited participants, ensuring at least one improved classification in a 6-participant per sample accuracy calculation.

To apply the paired one-sided t-test, specific assumptions must be satisfied, including the normality of the measured variable, the pairing of observations, and the independence of measurements, whereby the outcome of one pair does not influence another. To this end, the normality of the distribution of differences in accuracies was first assessed using the Shapiro-Wilk test. Additionally, the pairing of observations was ensured by selecting the same participants for each random resampling from both the original and modified force profile datasets. Concerning the independence assumption, it is acknowledged that while the data points in each bootstrap sample are not entirely independent due to their origin from the same limited dataset, the process of independently drawing each bootstrap sample helps to mitigate this concern to some degree.

Once normality is established, the one-sided paired t-test was used to statistically identify any significant changes in the accuracy of the models between datasets using the original realistic force profiles and those using the modified non-realistic force profiles. More precisely, for each the two models, the null and alternative hypotheses are as follows:

H_0: The decrease in the model's accuracy is less than 16.67% between datasets using the realistic force profiles and those using non-realistic force profiles.

$$\mu(Acc_{Realisitc\ Force} - Acc_{Non-Realisitc\ Force}) \leq 16.67\%$$

H_1: The decrease in the model's accuracy is more than 16.67% between datasets using the realistic force profiles and those using non-realistic force profiles.

$$\mu(Acc_{Realisitc\ Force} - Acc_{Non-Realisitc\ Force}) > 16.67\%$$

5.2.3 Results

5.2.3.1 Feature Distribution & MLP ANNs Accuracies

The distributions of the input feature values of the new collected data using both the original realistic and the modified non-realistic force profiles are compared to that of the training dataset used in developing the models in our previous study (Figure 5-10). The performance of the new models on both datasets are outline in Table 5-5 and the corresponding confusion matrices are highlighted in Figure 5-11 and Figure 5-12.

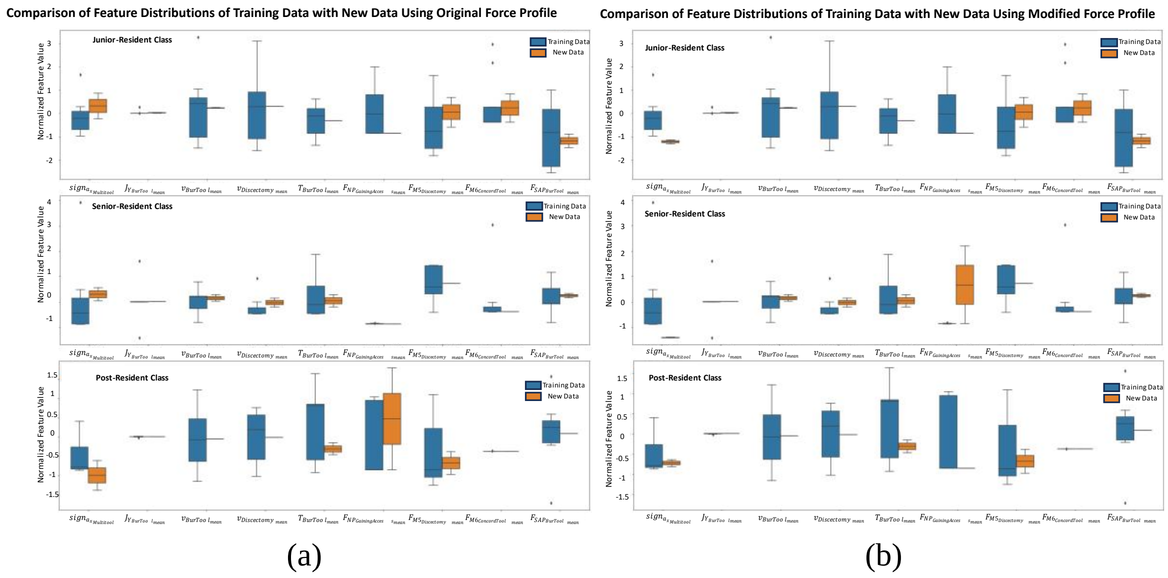


Figure 5-10 Box plots of the 9 surgical performance features used as inputs to the MLP ANNs, comparing the distribution of the feature values as compared to the data used to train the models with (a) new data based on the original force profile, and (b) new data based on the modified force profile.

Table 5-5 Accuracy performance of the trained models on the testing set used when developing the model, the new data collected using the original force profiles, and the new data collected using the modified force profile.

Model	Testing Accuracy in Previous Study (%)	Accuracy on New Data with Original Force Profile (%)	Accuracy on New Data with Modified Force Profile (%)
Stand Alone MLP ANN Model	**75**	**66.67**	**50**
MLP ANN with Transfer Learning	**87.5**	**83.33**	**50**

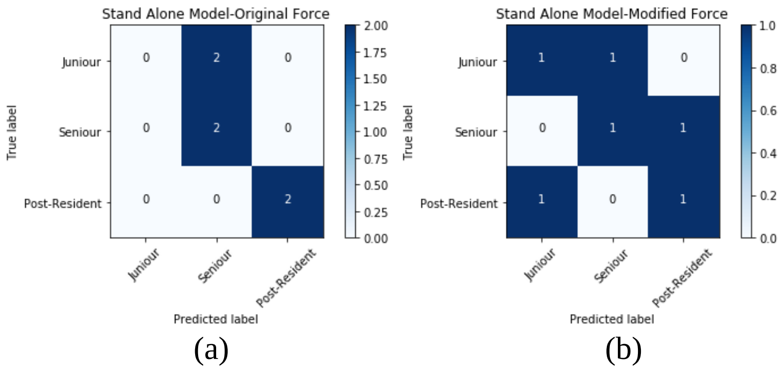
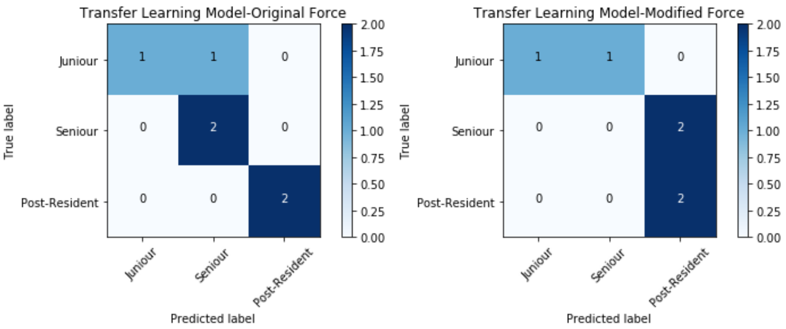


Figure 5-11 Confusion matrices highlighting the performance of the stand alone MLP ANN model trained from scratch on the newly collected data based on: (a) original force profile, and (b) modified force profile.



(a) (b)

Figure 5-12 Confusion matrices highlighting the performance of the MLP ANN model with transfer learning on the newly collected data based on: (a) original force profile, and (b) modified force profile.

5.2.3.2 Statistical Analysis

The change in the model's accuracy is used as a measure to establish statistical significance on the impact of the force profile realism on surgical classifications. The Shapiro Wilk test was used to test normality of the measured variable followed by conducting a one-sided paired t-test using the hypothesis outlined in Section 5.2.2.4 (Table 5-6).

Table 5-6 Statistical analysis results.

Model	Data Distribution Based on Shapiro Wilk Test	One-sided paired t-test		
Stand Alone MLP ANN Model	**Normal**	$\boldsymbol{t_{statisitic} = -0.14}$	$\boldsymbol{P_{value} = 0.886}$	**Fail to reject $\boldsymbol{H_0}$**
MLP ANN with Transfer Learning	**Normal**	$\boldsymbol{t_{statistic} = 3.21}$	$\boldsymbol{P_{value} = 0.005}$	**Reject $\boldsymbol{H_0}$**

5.2.4 Discussion

The study successfully met its primary objective of quantitatively assessing the impact of using accurate physics-based force profiles in surgical simulation training. This was done by demonstrating the influence of force profile fidelity on the accuracy of machine learning models in classifying surgical performance. The 'gaining access' phase of a discectomy approach was specifically chosen for analysis because it is the stage where surgeons most heavily rely on force feedback. The force profiles were modified using biomechanical principles and assumptions commonly employed in finite element modeling for simulations. The modified force profiles were confirmed to be sufficiently different by calculating the %VAF and verifying these results with those from a previous study that assessed qualitative detection sensitivity among expert surgeons.

The use of MLP ANNs revealed a discernible variation in classification accuracy when the models were applied to datasets with original, realistic force profiles versus those with modified, non-realistic force profiles. Notably, the stand-alone MLP ANN model, trained from scratch on the current simulator, demonstrated a decrease in accuracy from 66.67% to 50% when transitioning from the original to the modified force profiles. Similarly, the MLP ANN model employing transfer learning exhibited a substantial accuracy decline from 83.33% to 50% under the same conditions. In fact, statistical analysis demonstrated that transfer learning model performed significantly worse when altering the force profiles, with the accuracy dropping significantly below the 16.67% threshold. These results underscore the critical role of haptic feedback realism in virtual surgical training and its direct impact on the efficacy of machine learning models used within these training platforms.

5.2.4.1 Robustness of Models

The robustness of the machine learning models employed in this study is evident from their performance when utilizing the original, realistic force profiles – profiles that align with the data on which the models were initially trained. Both models performed relatively well, achieving an accuracy approximating the testing accuracies reported in our previous study. In alignment with findings from the previous study, the model employing transfer learning notably outperformed the stand-alone model, achieving an accuracy of 83.33% compared to the latter's 66.67%. This level of consistency not only highlights the robust nature of these models but also affirms their ability to generalize effectively, particularly when confronted with new datasets that bear a strong resemblance to their original training data.

A closer examination of the misclassified instances from the original realistic force profile data offers additional insights into the models' performance. Notably, the stand-alone model misclassified two junior participants as seniors, while the transfer learning model misclassified one junior as a senior. An in-depth analysis of the performance metrics of these misclassified individuals’ z-score performance scores, in relation to the CWPs of each model, offers insights that are consistent with prior findings. Z-score values define the number of standard deviations the surgical performance is from the mean values of each feature with the sign specifying whether the score is higher or lower than the average score among recruited participants. CWPs, indicative of the significance of each feature within the models, assign impact of a feature based on both the magnitude and the sign for each specific class. Our previous research indicated that while CWPs effectively explain misclassifications in the stand-alone model, their utility is somewhat limited in transfer learning models, particularly those with fixed initial layers. In these models, CWPs accurately assign relative importance to features but may not fully capture the influence of the sign – a critical aspect that determines whether higher or lower values of a feature affect the probability assigned by the model to a specific class. In the current study, for the stand-alone model, the CWPs were able to rationalize the misclassifications observed. The performance metrics of the junior participants classified as seniors revealed a similarity to the metrics typically associated with the senior class. This was especially evident in metrics deemed most influential by the CWPs, where the juniors' scores paralleled those of the seniors, potentially leading to the classification overlap. Conversely, for the transfer learning model, the performance metrics did not provide a satisfactory explanation for the misclassification, confirming our previous conclusions regarding the need for further refinement in understanding and interpreting the CWPs, particularly in the context of transfer learning with fixed initial layers. More specifically, it has been determined that employing transfer learning with fixed layers functions similarly to a high-level filter. This approach effectively combines and transforms the original input features into more complex, higher-level features, thereby making their interpretation substantially more intricate. Essentially, it is these newly combined features that are instrumental in capturing subtle performance cues, delineating surgical classes based on these nuanced indicators.

5.2.4.2 Impact of Physics-Based Force Profiles

To fully understand the effects of modifying the force profile, it's essential to examine the resultant differences in performance that led to changes in classification from the model's perspective. In this regard, the CWPs of the stand-alone model are particularly useful, given their proven interpretative power as evidenced in this and previous studies. Altering the force profiles, especially in the gaining access step, impacts key metrics like the number of sign changes in the multitool's acceleration and the average force exerted on the nucleus pulposus. The multitool's acceleration sign changes, representing the directional consistency of the tool's movements, can be viewed as a stability measure – a higher number of sign changes might indicate less stable, more tremor-like movements. The average force exerted on the nucleus pulposus is a pivotal aspect of minimally invasive lumbar interbody fusions. According to expert surgeons, successful execution of these procedures is marked by effectively penetrating through the annulus to reach the surgical site within the nucleus, consequently impacting forces on the nucleus pulposus. This step is a critical indicator of proficiency and precision in the surgical process. In fact, our previous study has shown that expert post-residents typically exhibit fewer directional changes and exert higher forces on the nucleus pulposus compared to their less experienced counterparts. These observations are substantiated by the CWPs across different surgical classes.

Focusing on the misclassifications in the stand-alone model with the modified force profiles, it was noted that two individuals, previously classified correctly, were misclassified following the profile change. These individuals deviated from their class's typical performance metrics to align more closely with the class they were incorrectly assigned to. A senior resident, for instance, was misclassified as a post-resident due to reduced Multitool directional changes and higher forces on the nucleus pulposus, aligning with post-resident benchmarks. A post-resident was misclassified as a junior resident due to slightly higher Multitool directional changes and lower than average force exertion on the nucleus pulposus, which are more characteristic of the junior resident class.

Although the CWP cannot fully elucidate the misclassifications in the transfer learning model, there is clear evidence that the features significantly influenced the model's accuracy. This is evident not only from the confusion matrices but more compellingly from the results of the statistical analysis. The one-sided paired t-test established statistical significance, showing that the accuracies dropped below the 16.67% threshold set, when the transfer learning model was evaluated using data derived from the modified force profiles. This finding is critical, as it demonstrates that while the direct interpretative impact of the features on the model might be elusive, the features related to the force profiles, especially during the gaining access step, significantly affect the overall performance of the participants. The statistical significance underscores the profound influence of these specific force profile features on the model's ability to classify surgical performance accurately.

5.2.4.3 Surgical Training Implications

The above analysis demonstrates the ramifications of lacking accurate, physics-based force feedback, particularly in the 'gaining access' step of minimally invasive lumbar interbody and fusion surgical procedures. The alteration of the force profile had a marked impact on the performance of expert participants. Notably, these experts, who are typically benchmarks in surgical simulators for junior and senior residents, failed to puncture and reach the nucleus pulposus effectively with the modified profiles, despite having successfully completed this critical step using the original force profile. This phenomenon highlights a significant risk in simulator surgical training: the potential for 'negative training', where trainees may develop incorrect skillsets due to inaccuracies in simulation feedback.

This study also underscores the challenges in capturing such complicated aspects of surgical performance, even with advanced validation methodologies like concurrent validation. These methods typically compare simulation results against a gold standard, such as expert ratings of surgical procedure videos. However, accurately measuring the intricacies of applied force in such contexts can be extremely challenging. Therefore, studies like the present one are invaluable, offering a unique and essential perspective for evaluating and ensuring the fidelity of surgical simulators. They serve as a crucial system check, ensuring that the training provided aligns accurately with the demands and realities of actual surgical procedures.

5.2.4.4 Limitations

The primary limitation of this study is the small dataset and the limited scope of the data pool utilized to generate the necessary observations for statistical analysis. While bootstrapping facilitated the creation of additional data points, it's crucial to acknowledge that these are not new, independent observations, but rather extrapolations from the existing dataset. As such, the interpretative power of the paired t-test applied to bootstrapped data may not fully correspond to what would be observed in a larger, independent dataset. This aspect constrains the generalizability of the statistical findings, suggesting that the current study might best serve as a pilot, providing a foundation for future research aimed at more comprehensively quantifying the impact of force profiles on surgical training.

Moreover, the nature of bootstrapping, involving resampling from a limited dataset, inherently challenges the assumption of independent paired observations. However, the independent drawing of each bootstrap sample, coupled with the maintained pairing across resampling, potentially mitigates this concern to some degree.

Another limitation arises in the application of the Connection Weights Product (CWP) for result interpretation. As previously evidenced and reiterated in this study, employing CWP in models that utilize transfer learning with fixed initial layers has its constraints. Yet, the current study reinforces the utility of CWP in standalone models, successfully rationalizing misclassifications in both original and modified force profile datasets. Despite CWP providing insights into classifications, it does not encapsulate the full breadth of decision-making processes in neural networks. Neural networks are adept at uncovering not just direct correlations between features and classes, but also intricate interdependencies among features themselves, where the value of one feature can influence and modulate the impact of another. Thus, CWP, while useful, offers only a partial perspective on classification rationale and should be applied with caution, acknowledging that it captures only one facet of the neural network's complex decision-making landscape.

5.2.5 Conclusion

This study represents a significant step towards the quantification of the impact of realistic, physics-based force feedback on surgical training within VR/AR environments. The findings demonstrated that the fidelity of force profiles plays a crucial role in the accuracy of machine learning models in classifying surgical performance. This was evident from the discernible variations in classification accuracy between original and altered force profiles, highlighting the importance of realistic haptic feedback in surgical simulations. The robustness of the employed machine learning models was validated by their consistent performance with the original data, closely resembling their training conditions. The alterations in force profiles resulted in significant misclassifications, even among expert surgeons, emphasizing the critical need for precise replication of realistic surgical environments. This accuracy is essential for creating valid performance benchmarks that trainees can realistically aspire to achieve. Despite the study’s limitations, this research not only sheds light on the significance of realistic force feedback in surgical training but also serves as a foundation for future studies. It paves the way for more comprehensive research to further explore and quantify the impact of force profiles in VR/AR surgical training, ultimately aiming to enhance the training and skills of future surgeons.

5.2.6 References

[1] A. M. J. C. o. i. u. Okamura, "Haptic feedback in robot-assisted minimally invasive surgery," vol. 19, no. 1, p. 102, 2009.

[2] E. M. Overtoom, T. Horeman, F.-W. Jansen, J. Dankelman, and H. W. J. J. o. s. e. Schreuder, "Haptic feedback, force feedback, and force-sensing in simulation training for laparoscopy: A systematic overview," vol. 76, no. 1, pp. 242-261, 2019.

[3] K. El-Monajjed and M. J. J. o. C. S. Driscoll, "Haptic integration of data-driven forces required to gain access using a probe for minimally invasive spine surgery via cadaveric-based experiments towards use in surgical simulators," vol. 60, p. 101569, 2022.

[4] G. A. J. T. h. o. m. b. m. Holzapfel, "Biomechanics of soft tissue," vol. 3, no. 1, pp. 1049-1063, 2001.

[5] K. El-Monajjed and M. Driscoll, "Analysis of Surgical Forces Required to Gain Access using a Probe for Minimally Invasive Spine Surgery via Cadaveric-based Experiments towards use in Training Simulators," *IEEE Transactions on Biomedical Engineering,* pp. 1-1, 2020. https://doi.org/10.1109/TBME.2020.2996980

[6] M. Hong, J. W. Rozenblit, and A. J. Hamilton, "Simulation-based surgical training systems in laparoscopic surgery: a current review," *Virtual Reality,* vol. 25, no. 2, pp. 491-510, 2021/06/01 2021. 10.1007/s10055-020-00469-z

[7] J. Zhang, Y. Zhong, and C. Gu, "Deformable Models for Surgical Simulation: A Survey," *IEEE Reviews in Biomedical Engineering,* vol. 11, pp. 143-164, 2018. 10.1109/RBME.2017.2773521

[8] S. Alkadri *et al.*, "Utilizing a multilayer perceptron artificial neural network to assess a virtual reality surgical procedure," *Computers in Biology and Medicine,* vol. 136, p. 104770, 2021/09/01/ 2021. https://doi.org/10.1016/j.compbiomed.2021.104770

[9] A. M. Fazlollahi *et al.*, "Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial," *JAMA Network Open,* vol. 5, no. 2, pp. e2149008-e2149008, 2022. 10.1001/jamanetworkopen.2021.49008 %J JAMA Network Open

[10] S. Alkadri, R. F. Del Maestro, and M. Driscoll, "Unveiling Surgical Expertise Through Machine Learning in a Novel VR/AR Spinal Simulator: A Multilayered Approach Using Transfer Learning and Connection Weights Analysis," *Computers in Biology and Medicine,* In Press.

[11] S. Alkadri, R. F. Del Maestro, and M. Driscoll, "Face, content, and construct validity of a novel VR/AR surgical simulator of a minimally invasive spine operation," *Medical & Biological Engineering & Computing,* 2024/02/26 2024. 10.1007/s11517-024-03053-8

[12] K. El-Monajjed, "Implementation of a virtual reality module for gaining surgical access via planned oblique lateral lumbar interbody fusion," 2021.

[13] W. K. Durfee and K. I. Palmer, "Estimation of force-activation, force-length, and force-velocity properties in isolated, electrically stimulated muscle," *IEEE Transactions on Biomedical Engineering,* vol. 41, no. 3, pp. 205-216, 1994. 10.1109/10.284939

[14] B. Stott *et al.*, "A Critical Comparison of Comparators Used to Demonstrate Credibility of Physics-Based Numerical Spine Models," vol. 51, no. 1, pp. 150-162, 2023.

[15] H. V. Chorney, J. R. Forbes, M. J. C. i. B. Driscoll, and Medicine, "System identification and simulation of soft tissue force feedback in a spine surgical simulator," vol. 164, p. 107267, 2023.

[16] D. T. Westwick and R. E. Kearney, *Identification of nonlinear physiological systems*. John Wiley & Sons, 2003.

[17] C. Bishop, *Pattern recognition and machine learning*. Springer, 2006, pp. 5-43.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

5.3 Conclusion

This chapter explored the final integration of chapters 3 and 4 towards achieving the global objective of the thesis in validating the current physics-based surgical simulator. The manuscript supported the thesis proposition regarding the importance of using accurate physics-based force profiles in surgical simulations. It introduced a novel methodological plan for simulator studies, aimed at gauging the importance of utilizing accurate physics-based forces in simulation. Employing the ML algorithms developed in the previous chapter, the manuscript offered a quantifiable and objective methodology to assess the impact of employing accurate force feedback during the gaining access phase of surgery Findings revealed a notable impact on the accuracy of ML models when using altered, non-realistic force profiles. Further, a detailed analysis of surgical performance demonstrated a deviation in expert performance from their established benchmarks when utilizing non-accurate force feedback.

Chapter 6. General Discussion

The overall objective of this thesis was to establish the comprehensive validity of a physics-based VR/AR spinal surgical simulator as a novel tool for the training and assessment of surgical trainees. At its core, the thesis sought to sequentially validate, via a component towards construct approach, various aspects of the simulator, ranging from its realism and potential assessment efficacy (face, content, and construct validity) to the exploration of surgical performance metrics and the usability of ML algorithms for enhancing the understanding of surgical expertise. Through thorough analyses, the study aimed to not only prove the simulator's effectiveness in training and assessment but also to shed light on the critical role of physics-based haptic feedback in the development of surgical skills, especially within MIS. This overarching objective guided the structured investigation across multiple dimensions of simulator validation, laying a solid framework that allowed the development of the manuscripts within the thesis, each contributing to the broader fields of surgical education, ML, and haptic-based VR/AR surgical simulation.

The first objective established the foundational validation of the newly developed simulator as defined by face, content, and construct validity. This phase not only validated the visual and skill realism of the VR/AR environment but also established innovative surgical metrics derived from psychomotor data. These metrics, pivotal for assessing construct validity, offered a quantitative lens through which surgical proficiency may be evaluated, distinguishing between three different skill levels, ranging from novice to expert. The positive reception of the simulator's realism by expert surgeons, alongside the critical analysis of tools like the Concorde clear tool and the identified limitations of the Burr tool, underscored the importance of accurate physics-based feedback in surgical training – the broad objective of this thesis. More precisely, feedback from LIF experts, especially those skilled in TLIFs and OLLIFs, reinforced Article 1’s conclusions. These experts collectively praised the Concorde clear tool, highlighting the authentic replication of forces and torques, including the tactile sensation of scraping, emphasizing its consequential significance for training surgical residents. These favorable feedback aligns with previous published findings by our group, underscoring the fidelity of the Concorde clear tool in mimicking the real surgical tool in active torque delivery and physical appearance [98]. Conversely, there was notable criticism directed at the Burr tool with regards to difficulty in maneuvering, reduced depth perception, and unrealistic force feedback. However, these criticisms served to support the central notion of this thesis. As elaborated in Article 1, critiques about the maneuverability and reduced depth perception linked to the Burr tool reinforce both face and content validity, aligning with Objective 1 and Hypothesis 1. Furthermore, feedback on the haptic forces of the Burr tool aligned with our group’s previous findings on the validation of the surgical tools [99]. The Burr tool was the only instrument to receive negative ratings on force feedback and also the sole tool programmed with unrealistic forces, strongly reinforcing the thesis's broad proposition regarding the importance of utilizing accurate physics-based and cadaver-derived force profiles, as explored in Chapter 5 (Objective 3).

Other findings related to Objective 1 revealed that a portion of recruited surgeons have limited experience in MI LIF surgeries, especially the innovative approach featured in the current simulation (Table 3-2). To standardize the analysis, every participating surgeon received identical information and guidelines during recruitment in the study’s trial. This methodology, however, inadvertently skewed in favor of seasoned surgeons already versed in the procedure. This variation was observed not only across different classes but also within the same category. Specifically, within the post-resident group, which included both fellows and staff surgeons, some fellows lacked expertise in the specific technique considered in this thesis due to the intricate nature of this surgery. Similarly, the junior-resident group showcased variations in previous knowledge, with some junior residents having limited prior experience in LIF surgeries. These observations slightly affected the data distribution of some surgical performance metrics, increasing the variance in the observed measurements. Nevertheless, this effect was seen in only two metrics among the eight statistically significant metrics in the construct validity analysis (Multitool Pathlength and Maximum Force on IAP in Figure 3-3) highlighting the limited impact of the observations.

The foundational validation established by Article 1 paves the way for subsequent phases of assessment, specifically targeting concurrent and predictive validity in future studies. There is a growing need to move towards the more advanced and more impactful validation steps defined by concurrent and predictive validation. These validation steps are linked to clinical outcomes and thus can further justify the investments of incorporating such simulations in training curriculums. Owing to the compactness and portability of the current simulator, it is possible to recruit surgeons from multiple centers, such as within the McGill hospital network, and track their progression over a set period, both before and after simulator training. For concurrent validity, validated surgical skill assessment tools can be employed, with expert-scored video recordings of simulation sessions as described in Section 1.1.4. For predictive validity, the same surgical skill assessment tools may be used to evaluate real surgical operations alongside patient outcome measures to gauge the effectiveness of the training.

Despite the promising findings of these validation approaches, a significant gap persists in the standardization of best practices for conducting validation studies on surgical simulations [100]. Variability in the use of subjective Likert-scale questionnaires across different studies has been noted [101]. Recent reviews and meta-analyses recommend the standardization of validation efforts by implementing a uniform Likert scale across questionnaires and ensuring that questions are consistently aimed at specific validation aspects [100]. Furthermore, there has been recent disputes in choosing the best validity framework for surgical simulations. Critiques have emerged against the use of the traditional face, content, construct, concurrent, and predictive validity, labeling these methods as outdated. Instead, a shift towards the Messick framework is recommended, which articulates five sources of evidence-based validity: content, response process, internal structure (reliability), relations with other variables, and consequences of assessment/test [102]. Upon closer examination, the Messick framework shows substantial overlap with the traditional model it seeks to replace. Both frameworks assess content validity, and the "response process" and "internal structure" components of the Messick framework mirror elements typically evaluated through concurrent validity. Similarly, the "relations with other variables" closely align with construct validity. Finally, the "consequences" step closely mirrors predictive validity, both highlighting their importance in evaluating clinical outcomes. The primary critique centers around the subjective nature of face validity and its perceived inadequacy as a standalone measure of validation. This critique often points to studies that rely solely on face validity for validation claims. However, as explained in Section 1.1.4 and exemplified in Article 1, the research in this thesis underscores the crucial simultaneous application of both face and content validity. This dual approach ensures the simulation's surface realism and its efficacy in measuring precisely what it is intended to measure, thereby addressing the core objectives of simulation-based surgical training.

Following the establishment of foundational validation of the developed novel simulator, the next objective focused on harnessing ML algorithms to not only accurately classify surgical performances but also to unravel deeper insights into surgical expertise for enhanced assessment and training. To approach this objective, an initial side study using a similar spinal surgical simulator laid the groundwork for establishing a methodology plan to conduct ML analyses in spinal surgery simulations. This initial side study facilitated the formulation of a novel methodology for determining feature importance, resulting in the development and training of a two-layered ANN focused on the incision task within an ACDF (anterior cervical discectomy and fusion) surgery scenario.

As highlighted in the literature review Section 1.2.2.1, a comprehensive framework for ML encompasses model representation, evaluation, and optimization. The utilization of adaptive bases within model representation is particularly valuable, enabling the model to accurately capture data's nonlinearity while evading overfitting – a critical benefit attributed to neural networks. These networks stand out for their adaptability across both small and large datasets, irrespective of data complexity, largely due to the potential for model architecture adjustments that foster enhanced generalization through deep learning. However, this adaptability often comes at the expense of model interpretability, a significant concern in applications such as surgical training where understanding feature impact is crucial. Despite this challenge, neural networks hold the potential of deciphering the reasoning behind classifications through analysis of the hidden weights. The CWA (Connection Weights Algorithm), pioneered by Olden and Jackson [67] represents a strategic tool in this endeavor for one-layered neural networks. Article 3 aimed to extend this tool's applicability to multilayered neural networks, thereby unveiling the model’s decision-making process with the aim of leveraging these insights in the surgical training of residents.

Article 4 advanced towards the realization of Objective 2 by applying the methodologies and insights garnered from Article 3 to the newly developed simulator. The article demonstrated the successful development of two distinct ANN models, which achieved high classification accuracies of 75% and 87.5%, respectively. This accomplishment not only reinforced the simulator's capability in differentiating among surgical skill levels – as affirmed in the construct validity analysis under Objective 1 – but also revisited and validated the innovative approach for determining feature importance initiated in Article 3. Furthermore, Article 4 defined a methodological blueprint for overcoming the challenges inherent in surgical simulation studies, particularly the small datasets typically collected from a single study center. This blueprint encompassed a strategic combination of data augmentation, feature selection, and transfer learning. Specifically, the article leveraged the ANN model developed from the side study on the SimOrtho platform, successfully employing transfer learning techniques that enhanced the classification accuracy from 75% to 87.5%. A notable challenge encountered in this process was assessing the applicability and potential benefits of transferring a model trained on the ACDF surgery scenario, characterized by an open approach, to a model designed for the MI OLLIF approach. Despite these differing surgical methodologies, both surgical tasks necessitate precise manipulation and deliberate force application on specified anatomical structures, with a concerted effort to minimize unnecessary interactions. Both the open ACDF approach and the MI OLLIF technique, have a shared emphasis on surgical precision, safety, and efficiency. In the ACDF incision task, surgeons are guided to perform incisions along the vertebrae's borders, avoiding critical anatomical structures [103]. This methodical approach mirrors post-residents' performance in MI OLLIF surgeries, where controlled movements and calculated force applications are pivotal [104]. Both surgical contexts also demand an acute reliance on tactile feedback for successful navigation. This commonality substantiates the transfer learning application, providing a logical extension of the model developed for the ACDF scenario to the MI OLLIF simulator.

Efforts to implement transfer learning involved experimenting with both deep tuning and freezing pre-trained layers as discussed in Article 4. Sequential tuning from shallow to deep layers led to overfitting, characterized by disproportionately high training accuracies contrasted with significantly lower validation accuracies. This phenomenon is consistent with findings in the literature when applying relatively shallow neural networks to small datasets [57, 63]. The alternative approach, which was ultimately adapted in the study, entailed freezing the pre-trained layers while introducing new, trainable layers in an attempt to refine model performance. In addition to selecting a transfer learning technique, an effort was made to correlate new model inputs with analogous metrics from the original model, such as mapping velocity features of the new model to velocity features in the old model. Contrary to expectations, this alignment did not yield improved performance on the validation set. This observation led to the realization that, in the context of fine-tuning, the model inherently relearns and adjusts to new input feature, obviating the need for manual feature mapping. Similarly, when employing frozen layers, the emergent higher-level features derived from input combinations adequately differentiated performance levels without necessitating explicit feature alignment. The phenomenon observed when using frozen layers made evident that interactions among features were similar across both surgical simulations. In fact, both Articles 3 and 4 highlighted the need for a combination of controlled movements and deliberate forces in delineating post-resident expertise from other levels. Therefore, it is evident that the feature extractor method used in frozen layers generates new high-level features that capture these interactions among original input features. As noted in Article 4, while this feature extractor methodology proved effective in classifying performance, it concurrently obscured the model's interpretability due to the generation of composite features. These observations imply that the application of transfer learning across surgical simulators must consider not only the overarching task similarities but also the intricate feature interrelations. Yet, upon confirming task and feature interaction parallels, transfer learning exhibits remarkable adaptability, obviating the need for precise feature mapping. This demonstrates the utility and flexibility of applying transfer learning within ML analyses on surgical simulations.

Building on the discussion of transfer learning's adaptability and its role in surgical simulation studies, it is imperative to delve into the raised criticisms surrounding the use of the CWA and its theoretical foundations. A critical aspect of this discourse is the correct application of the CWP. The consensus within the literature is that the application of CWP is primarily for comparing the relative magnitude and sign of CWPs within the same model, as the absolute values lack cross-model comparability [105]. Moreover, concerns have emerged regarding the feasibility of comparing CWPs' signs and magnitudes across different classification outputs within the same model. Such arguments have been mathematically challenged, especially when considering the computation of CWPs for input features before their normalization to individual groups (outputs). The process reveals that when CWPs are compared on a class-by-class basis, the vectors of CWPs undergo normalization for each output, yet the signs of these CWPs remain consistent with those derived prior to normalization. This calculation method not only sheds light on the relative significance of input features within each classification group but also maintains the utility of the sign in explaining the influence of a given feature on the classification outcome. This calculation facilitates an internal comparison within the model that highlights the sign’s indicative role: a positive sign suggests that values greater than the average are associated with a particular class, while a negative sign indicates lower than average values. However, within the domain of transfer learning, a significant constraint emerges regarding the interpretability of the CWP's sign, which compromises the algorithm's capacity to deduce how a specific feature influences class association. Despite this, it remains possible to ascertain the relative significance of that feature, a point thoroughly examined in Article 4. This limitation highlights a pivotal area for further investigation, particularly in enhancing the algorithm's adaptability to transfer learning contexts while maintaining the integrity of its interpretive capabilities. One recommendation includes assessing the improvements in the CWA analyses between a fine-tuned and a feature generator transfer learning algorithm.

Further points raised during the discourse on CWA is the debate over the interpretation of the CWPs' sign and magnitude, which presumably indicate the model’s classification tendencies rather than the actual performances observed during training. Essentially, CWPs articulate the model's understanding on class behaviors, offering insights into the algorithmic deductions on classification cues and transforming these observations into potentially teachable recommendations. In simpler terms, the CWPs for the input features reflect the model's understanding of how each class performed across the input features, rather than mirroring the actual performances observed in the training set for each class. Essentially, CWPs indicate the model's likelihood of predicting a class for the next input, rather than directly assessing the likelihood that a new input genuinely belongs to a specific class. This perspective underlines the CWPs' utility in unveiling the model's interpretative frameworks regarding trained performances. Therefore, it is crucial to focus on adequate model training and optimization to strengthen the reliability of the model's predictions. Apart from ensuring optimal training of the model on the data, this brings to light another critical concern: the fidelity of the training data. The efficacy of model-based interpretations and, consequently, the validity of training recommendations hinge on the representativeness and quality of the dataset. In scenarios where models are trained on limited or non-representative data, the risk of model limited generalizability and inaccurate classifications escalates, potentially leading to flawed training guidance. To circumvent these limitations, the employment of strategies such as data augmentation and transfer learning becomes crucial, particularly in situations where collecting expansive and diverse datasets via multicenter studies remains a challenge. This not only reinforces the significance of methodical data collection and preparation but also highlights transfer learning's role in enhancing model robustness and reliability in surgical simulation applications.

The study’s presented efficacy of data augmentation and transfer learning strengthens the foundation for applying similar algorithms to new simulators with limited datasets. The methodology applied in this thesis becomes more impactful with the expanding pool of data from existing simulations, transferring the knowledge learnt across models and simulators. This approach necessitates a collaborative effort between surgical experts and ML specialists to identify commonalities across different procedures or distinct phases within those procedures to ensure successful use of transfer learning. Apart from making use of this methodology on existing simulators, integrating this strategy from the outset in the development of new simulators may be pivotal. It may guide the engineering process towards optimizing simulator designs for complexity that is necessary and sufficient, identifying and enhancing surgical performance differentiators as directed by the ML algorithm. Such strategic design considerations, grounded in the principles of efficient production, do not undermine the essential validations of face and content but rather streamline the development process to prioritize the simulator's instructional and evaluative fidelity.

Integrating the findings from Objectives 1 and 2, an intriguing discrepancy emerges between the metrics identified as significant in the construct validity analysis (Table 3-5), which relies on traditional statistical tests, and those highlighted by the ML analysis employing the forward SFS algorithm (Table 4-21). This variance can be attributed to distinct characteristics inherent to the analytical approaches of these methods. Traditional statistical tests like ANOVA or Kruskal-Wallis are fundamentally univariate, focusing on the distribution of a single variable across groups. They are built on assumptions of linear relationships and come with specific prerequisites, such as homogeneity of variances and normally distributed data. These methods excel in identifying differences in isolated variables but may not adequately capture the complexities of multivariate interactions or non-linear dynamics that can exist both between inputs and outputs and within the interactions among the features themselves. Contrastingly, neural networks, as leveraged in the ML analysis, inherently accommodate multivariate data and are effective at identifying higher-order interactions between features. Their capacity to account for non-linearity – in both the relationships between inputs and outputs and among the features– marks a significant advantage in identifying patterns that might remain hidden under traditional statistical analyses. This distinction underlines why certain metrics that do not exhibit statistically significant differences across groups in a univariate linear context might still be crucial within the multivariate non-linear frameworks of neural networks. In fact, this advantage is particularly pertinent when considering the selection of features that enhance model performance rather than solely relying on statistical significance – a decision that motivated the preference for the wrapper method over filter methods in this study's feature selection process. This approach reflects the understanding that a metric's predictive power might emerge more from its interactions with other features than from its isolated statistical significance. This insight aligns with the phenomena observed during transfer learning, where the generation of new, high-level features through interactions among existing features was crucial for accurate classification.

Nevertheless, combining both the depth of neural networks with the breadth of traditional statistical analyses provide a more holistic understanding of the multidimensional interactions inherent in complex datasets. The combined results from the construct validity and ML analyses offer a more comprehensive understanding of the simulated OLLIF surgical performances. The construct validity analysis complements and further supports the analysis of expert performances detailed in Article 4. When combined, the analyses show that experts use a direct approach in gaining access to the disc characterized by short paths and minimal directional changes, puncturing through the annulus to precisely settle in the nucleus. Similarly, both analyses revealed that during Facetectomy and Discectomy, experts consistently use stable, deliberate movements, marked by reduced velocities and fewer directional changes, indicative of a focused and controlled surgical technique. Intriguingly, this comprehensive analysis also highlights a unique aspect of force application: experts tend to interact and exert considerable force on both the L5 SAP, a finding primarily surfaced through the ML analysis, and the L4 IAP, as indicated by the traditional statistical methods. This is contrasted with junior and senior residents who tend to interact with either structure in isolation. Combining the analyses provides comprehensive performance profiles for each surgical class, with the post-resident profile closely aligning with the detailed descriptions and expert recommendations outlined in Section 1.3.1. These recommendations emphasize minimizing manipulations and movements during gaining access to the IVD; avoiding interactions with neural components; creating sufficient space during Facetectomy to ensure optimal Discectomy and safe cage insertion; and advising against over-preparation of endplates to preserve the structural integrity of the vertebrae's bony structures for optimal outcomes. The post-resident profile, as derived from the combined analyses, is marked by minimal manipulations and exploratory movements during the gaining access step; a general avoidance of nerves and the cauda throughout the procedure; interaction with both the SAP and IAP as needed during Facetectomy; and careful preparation of the endplates by removing only what is necessary. In contrast, junior and senior residents were less direct in reaching the surgical site during the gaining access step, interacted with either the SAP or the IAP exclusively, and did not maintain sufficient distance from the cauda and nerves. They also overprepared the endplates by removing more than necessary compared to post-residents. The alignment of the post-residents’ performance with the expert description further supports using their performance as the benchmark for training. Moreover, these findings underscore the significance of combining both analyses in complex datasets for a refined analysis. This approach informs future and current studies employing ML algorithms to analyze surgical performances, both within simulation setting or more generally using real surgical videos. For instance, apart from VR/AR surgical simulators, one current commercial application includes the platform developed by Medtronic, Touch Surgery^{TM} (2022, London, UK), which uses ML to analyze real surgical videos for creating training tools.

The ML study in Article 4 illuminated the critical role of force metrics in distinguishing surgical proficiency levels, pinpointing the forces exerted on the NP during the gaining access phase as an important feature. This was demonstrated by both the permutation feature importance results and the post-residents' CWPs for both ANN models (Table 4-29). These results indicated that the forces impacted on the NP during gaining access are not only important for general classification but also specifically to identify post-resident expertise. Such insights were pivotal, especially since post-resident performances closely mirrored textbook descriptions of the MI OLLIF surgery. The gaining access phase, devoid of visual feedback, necessitates reliance on tactile and somatosensory feedback for precise navigation, highlighting the potential importance of using accurate physics-based forces. These findings align with the emphasis on the gaining access phase as a key differentiating step, further supporting Chapter 5's focus on that surgical step for examining the significance and impact of employing physics-based forces as defined by Objective 3.

Objective 3 integrated findings from Chapter 4, particularly the trained ANN models and the novel adaptation of the CWA, to highlight the significance of physics-based forces during the gaining access step. To evaluate hypothesis 3, modifications to the haptic forces used in the gaining access phase were informed by two primary considerations: adapting commonly used assumptions in biomechanical modeling and ensuring the new force profile differed sufficiently from the original to be noticeable by users. The former was guided by previous publications by our group demonstrating the prevalent use of linear approximations with homogenous mechanical properties in mechanical simulation and modelling [106]. Additionally, it was noted that most current surgical simulators on the market incorporate simplified unverified forces [96]. The latter involved a verification process to confirm the new force profile's distinctiveness. Our group has demonstrated a %VAF technique to quantify the minimal detectable change in force profiles, ensuring users could qualitatively discern the adjustments using the same simulator system explored in this thesis [107]. Article 5 extended this methodology to verify the new force profile's distinguishability.

The statistical analysis within Article 5 acknowledged limitations related to generalizability, mainly due to sampling a limited dataset with replacement. This weakness in the independence assumption highlighted a constraint in the study's statistical framework. However, the conclusions drawn from this study were not solely dependent on statistical testing. Equally, the study leveraged ML analysis to delve into the surgical performances under both original and modified force profiles, examining how the use of realistic cadaveric-derived force profiles influenced performance and classification. Furthermore, despite the dataset's limitations, the article introduced a novel approach to validation, aiming to mitigate the potential for negative training. As surgical training increasingly incorporates simulation, addressing the risk of trainees being guided towards inaccurate performance benchmarks becomes crucial [77]. The manuscript proposed a methodology for quantifying the impact of accurate force profiles on MIS performance, which could ultimately lead to trainees directed towards incorrect skill levels.

Objectively measuring the significance of employing accurate physics-based forces in surgical training poses a challenge, often necessitating to revert to subjective methodologies for evaluation. Generally, to objectively assess if a particular surgical metric could lead to negative training, one might require the implementation of predictive validation. This method would compare the training outcomes to actual clinical results, offering a direct measure of the training's impact on real-world surgical proficiency. However, relying solely on such methods isn't always feasible, given that such validation steps are usually conducted in the later stages of simulation validation, necessitating extensive recruitment and prolonged training periods. Therefore, within the scope of simulation studies, subjective assessments, such as the one conducted in Chapter 3, are mostly used to measure the haptics fidelity [96]. For instance, these assessments subjectively highlighted the low fidelity of the Burr tool, which received the lowest expert ratings for force accuracy compared to the Multitool and Concorde tool, both of which utilized cadaver-based force profiles. This signifies the importance and novelty of the approach presented in Article 5 in objectively assessing haptic fidelity. Nevertheless, this approach may not be useful in all simulation contexts. For example, in the specific context of the Burr tool, the objective approach used in Article 5 may not be useful to quantify the impact of using non-realistic force profiles. Despite having less accurate force feedback, both the ML and the construct validity analyses identified the forces applied by the Burr tool on the SAP and IAP as distinct across different skill levels. In surgical scenarios offering direct visual feedback, such as the Discectomy step, the accuracy of haptic feedback may not influence the surgeon's decision-making regarding which anatomical structures to engage with. As such, surgeons can selectively interact with specific anatomical structures and apply varying forces regardless of the haptic feedback's fidelity. Nevertheless, inaccurate haptic feedback in these cases could still inadvertently guide trainees towards incorrect skill levels – a risk that might not be fully understood or quantified until more advanced validation studies, like predictive validation, are conducted.

With the above in mind, the gaining access step in MIS scenarios offers a unique opportunity to employ the methodology developed in Article 5. This critical phase lacks visual feedback, necessitating surgeons to depend on tactile feedback as the surgical tool makes constant contact with tissue during puncturing. In contrast to situations where visual feedback might mitigate the impact of inaccurate forces, the gaining access step's reliance on precise force feedback directly influences surgeons' performances in the simulation. Surgeons count on the force feedback for accurate navigation to the surgical site. Article 5’s thorough analysis demonstrated that expert surgeons initially replicated benchmark performances established by previously recruited experts. However, their performance significantly deteriorated to novice levels upon the introduction of unrealistic force profiles, illustrating the direct impact of force feedback accuracy on expert performance in this phase. As noted, expert performance benchmarks guide simulation training. Therefore, this marked deterioration provides evidence that inaccurate force feedback may directly lead to negative training outcomes, highlighting the necessity of accurate force feedback to prevent the misdirection of trainees.

The robustness of the ANNs developed in this thesis is evidenced by their consistent accuracies when applied to data from newly recruited participants in Article 5. This consistency underscores the ANNs' capability to generalize effectively to new data affirming the reliability and stability of these models. Moreover, the overall robustness of the assessment methodology is highlighted by the strong alignment of expert performances, as determined by the novel CWA method developed in this thesis, with the OLLIF surgical recommendations. This alignment indicates that the assessment method accurately captures the nuanced performance characteristics of expert performances, thereby validating the practical applicability and reliability of the methodology in real-world surgical training and assessment contexts. To further evaluate the robustness of the methodology in future studies, beyond concurrent and predictive validation, future research could employ video assessments using simulation-generated videos as an intermediary step before the more complex concurrent and predictive validation steps, similar to real-life surgical video assessments.

Reflecting on the comprehensive analysis presented in this thesis, a crucial overarching recommendation emerges: the need for standardized guidelines in conducting surgical simulation validation. This encompasses all facets from foundational validation, through ML analyses, to the validation of haptic feedback to ensure avoidance of negative training. The Machine Learning to Assess Surgical Expertise (MLASE) checklist developed by Winkler-Schwartz, et al. [53] provides one framework for conducting ML studies to assess surgical expertise. MLASE offers a 20-point checklist spanning four main components related to the study's design, data structure, ML algorithm used, and discussion quality. This checklist is increasingly relied upon to assess the quality of ML studies. The framework presented in this thesis could act as a refinement and extension to such a framework, especially since the current thesis presents a more detailed and comprehensive analysis of surgical simulator validation, including the initial foundational steps and the assessment of haptic fidelity's impact on training outcomes. Addressing subjective validity limitations, future studies should aim to outline the essential elements required to comprehensively capture both face and content validity, thus mitigating perceptions of inadequate validation. The demonstrated method in this thesis of combining ML analyses with conventional statistical approaches offers a novel pathway for enhancing intelligent tutoring systems and broader skill assessment methodologies using ML. In addition to setting standard guidelines, future studies should investigate commonalities among different surgical simulations. Such studies would facilitate more reliable ML analyses of surgical performances through transfer learning as previously discussed. Furthermore, this strategy enhances the efficiency of future simulation studies; it fosters collaboration by facilitating the extraction and application of insights from existing multicenter predictive validation studies to the development of new simulators still in their early stages. By identifying and leveraging the commonalities across different surgical procedures and simulations, this approach streamlines the validation process, ensuring that new simulator studies can benefit from the comprehensive, real-world data already gathered. Such an initiative could directly address the challenges of objectively validating haptic fidelity, especially in tools like the Burr tool, where visual feedback complicates validation using the methodology developed in Objective 3. Concrete outcomes from predictive validation would offer a robust measure of haptic fidelity's real-world implications. Predictive validation may allow the opportunity to assess the impact of accurate physics-based forces on real surgical outcomes in the operating room. Coincidentally, this form of validation would concurrently assess the validity of the novel method introduced in Objective 3, designed to evaluate the impact of accurate physics-based forces on virtual surgical performance within simulations. This approach ensures that advancements in virtual surgical training are both reflective of and directly beneficial to actual surgical practice, thereby closing the gap between simulation environments and real surgery experiences.

In summary, the objectives outlined in this thesis have successfully led to the validation of the VR/AR physics-based spinal surgical simulation. The work presented a sequential and methodical approach to address each validation aspect necessary for VR/AR surgical simulations that requires validation prior to the widespread implementation into surgical curricula.

Furthermore, the work provided a comprehensive framework for advancing the validation of existing surgical simulations and for refining simulators currently in development. Significantly, the methodologies developed throughout this thesis have broader applications, extending beyond surgical simulations to other high-risk training tools that incorporate ML for performance classification.

Chapter 7. Conclusion

This doctoral dissertation has comprehensively validated the VR/AR OLLIF surgical simulator. Each of the three objectives contributed distinctively to the overarching goal by exploring various facets of simulator validation. Through this endeavor, significant contributions were made to the fields of surgical simulation validation, ML, and surgery biomechanics.

Objective 1 focused on foundational validation, involving a study with surgical experts and residents in orthopedics and neurosurgery. This phase validated both the visual and operational realism of the VR/AR environment and introduced innovative surgical metrics based on psychomotor data. The conclusions drawn from this study aligned with prior research conducted by our group and were further supported by expert consultations LIF surgeries. Additionally, subjective evaluations of the surgical tools underscored the value of physics-based feedback in MIS, a central theme of the thesis.

Objective 2 capitalized on the novel surgical metrics established in Objective 1 to showcase the application of ANNs in assessing and analyzing surgical proficiency. This objective introduced several innovations in the ML assessment of surgical performance, with potential applicability extending beyond surgical simulator validation to encompass analyses of real-life surgical performances. A novel methodology for identifying feature importance in multilayered ANNs was developed, along with a methodological blueprint designed to tackle the common challenge of small dataset sizes typically encountered in surgical simulation studies. Moreover, it offered insightful strategies for successfully transferring learned models across different surgical performance datasets. Additionally, the integration of Objectives 1 and 2 proposed a novel approach of blending the analytical strengths of neural networks with traditional statistical analyses to uncover the complex interactions within surgical performance data.

Objective 3 built on the insights from Objective 2, particularly leveraging the trained ANN models and the novel adaptation of the CWA, to emphasize the importance of physics-based force feedback during the “gaining access” step of surgery. It introduced a novel approach to validation aimed at mitigating the risk of negative training. It presented an innovative objective measure for evaluating haptic fidelity, presenting a significant addition to computational comparisons of haptic output profiles to cadaveric data. This approach stands out from customary subjective assessment methods by objectively demonstrating how the use of physically accurate forces impacts virtual surgical performances and, consequently, training effectiveness. The findings highlighted the necessity of realistic haptic feedback for surgical training in MIS; it also offered a methodical way to assess the fidelity of force feedback in surgical simulators, directly linking the accuracy of haptic cues to the quality of surgical training.

This thesis successfully validated the VR/AR physics-based spinal surgical simulator, presenting a sequential and thorough framework for enhancing surgical simulation validation, training, and assessment. The developed methodologies offer wider implications for advancing the development and validation of both surgical and medical training tools. This paves the path towards more effective and reliable methodologies for training in high-stakes fields.

References

[1] H. Abbasi and A. Abbasi, "Oblique Lateral Lumbar Interbody Fusion (OLLIF): Technical Notes and Early Results of a Single Surgeon Comparative Study," *Cureus*, vol. 7, no. 10, pp. e351-e351, 2015.

[2] R. A. Deyo, A. Nachemson, and S. K. Mirza, "Spinal-Fusion Surgery&#x2014;The Case for Restraint," *The Spine Journal*, vol. 4, no. 5, pp. S138-S142, 2004.

[3] A. J. Schoenfeld, L. M. Ochoa, J. O. Bader, and P. J. J. Belmont, "Risk Factors for Immediate Postoperative Complications and Mortality Following Spine Surgery: A Study of 3475 Patients from the National Surgical Quality Improvement Program," *JBJS*, vol. 93, no. 17, pp. 1577-1582, 2011.

[4] B. Bashankaev, "Review of available methods of simulation training to facilitate surgical education," (in eng), *Surgical Endoscopy*, vol. 25, no. 1, p. 28, 2011.

[5] I. Nisky, F. Huang, A. Milstein, C. M. Pugh, F. A. Mussa-Ivaldi, and A. Karniel, "Perception of stiffness in laparoscopy - the fulcrum effect," *Studies in health technology and informatics*, vol. 173, pp. 313-319, 2012.

[6] Y. W. Mun, "Getting the Most Out of Minimally Invasive Spine Surgery," 29 Nov 2018. Available: https://www.gleneagles.com.sg/healthplus/article/minimally-invasive-spine-surgery

[7] R. J. G. Stevens, M. P. Davies, and L. Hadfield-Law, "“Do One, Teach One”: The New Paradigm in General Surgery Residency Training," *Journal of Surgical Education*, vol. 69, no. 2, pp. 135-136, 2012/03/01/ 2012.

[8] N. Vaughan, "A review of virtual reality based training simulators for orthopaedic surgery," (in eng), *Medical Engineering and Physics*, vol. 38, no. 2, p. 59, 2016.

[9] M. Goldenberg and J. Y. Lee, "Surgical Education, Simulation, and Simulators-Updating the Concept of Validity," (in eng), *Curr Urol Rep*, vol. 19, no. 7, p. 52, May 17 2018.

[10] M. Pfandler, M. Lazarovici, P. Stefan, P. Wucherer, and M. Weigl, "Virtual reality-based simulators for spine surgery: A systematic review," *The Spine Journal*, vol. 17, 05/01 2017.

[11] S. Alkadri, "Kinematic Study and Layout Design of a Haptic Device Mounted on a Spine Bench Model for Surgical Training," Undergraduate Honours Program - Mechanical Engineering, Mechanical Engineering, McGill University, 2018.

[12] N. Ledwos, N. Mirchi, V. Bissonnette, A. Winkler-Schwartz, R. Yilmaz, and R. F. J. O. N. Del Maestro, "Virtual reality anterior cervical discectomy and fusion simulation on the novel sim-ortho platform: validation studies," *Operative Neurosurgery*, 2020.

[13] Y. Munz, "Laparoscopic virtual reality and box trainers: is one superior to the other?," (in eng), *Surgical Endoscopy*, vol. 18, no. 3, p. 485, 2004.

[14] C. B. Franzese and S. P. J. O. C. o. N. A. Stringer, "The evolution of surgical training: perspectives on educational models from the past to the future," vol. 40, no. 6, pp. 1227-1235, 2007.

[15] J. Hamdorf and J. J. B. J. o. S. Hall, "Acquiring surgical skills," vol. 87, no. 1, pp. 28-37, 2000.

[16] L. Nguyen *et al.*, "Education of the modern surgical resident: novel approaches to learning in the era of the 80-hour workweek," vol. 30, pp. 1120-1127, 2006.

[17] N. Gélinas-Phaneuf and R. F. J. N. Del Maestro, "Surgical expertise in neurosurgery: integrating theory into practice," vol. 73, pp. S30-S38, 2013.

[18] N. K. Choudhry, R. H. Fletcher, and S. B. J. A. o. I. m. Soumerai, "Systematic review: the relationship between clinical experience and quality of health care," vol. 142, no. 4, pp. 260-273, 2005.

[19] J. Martin *et al.*, "Objective structured assessment of technical skill (OSATS) for surgical residents," vol. 84, no. 2, pp. 273-278, 1997.

[20] M. C. Vassiliou *et al.*, "A global assessment tool for evaluation of intraoperative laparoscopic skills," vol. 190, no. 1, pp. 107-113, 2005.

[21] D. M. J. S. i. H. Gaba, "The future vision of simulation in healthcare," vol. 2, no. 2, pp. 126-135, 2007.

[22] W. R. Sherman and A. B. Craig, *Understanding virtual reality: Interface, application, and design*. Morgan Kaufmann, 2018.

[23] J. Gilbody, A. W. Prasthofer, K. Ho, and M. L. Costa, "The use and effectiveness of cadaveric workshops in higher surgical training: a systematic review," *Annals of the Royal College of Surgeons of England,* vol. 93, no. 5, pp. 347-352, 2011.

[24] M. W. Krueger, "An easy entry artificial reality," in *Virtual Reality*: Elsevier, 1993, pp. 147-161.

[25] J. Steuer, F. Biocca, and M. R. J. C. i. t. a. o. v. r. Levy, "Defining virtual reality: Dimensions determining telepresence," vol. 33, pp. 37-39, 1995.

[26] M. S. Fourman *et al.*, "Applications of augmented and virtual reality in spine surgery and education: A review," *Seminars in Spine Surgery,* vol. 33, no. 2, p. 100875, 2021/06/01/ 2021.

[27] L. M. Sutherland *et al.*, "Surgical simulation: a systematic review," vol. 243, no. 3, p. 291, 2006.

[28] R. Rehder, M. Abd-El-Barr, K. Hooten, P. Weinstock, J. R. Madsen, and A. R. J. C. s. N. S. Cohen, "The role of simulation in neurosurgery," vol. 32, pp. 43-54, 2016.

[29] J. S. Denson and S. J. J. Abrahamson, "A computer-controlled patient simulator," vol. 208, no. 3, pp. 504-508, 1969.

[30] R. M. J. W. j. o. s. Satava, "Historical review of surgical simulation—a personal perspective," vol. 32, pp. 141-148, 2008.

[31] D. Saddawi-Konefka and J. B. J. C. H. S. A. Cooper, "Anesthesia and simulation: an historic relationship," pp. 3-13, 2020.

[32] S. Delorme, D. Laroche, R. DiRaddo, and R. F. Del Maestro, "NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training," *Operative Neurosurgery,* vol. 71, no. suppl_1, 2012.

[33] A. Gallagher, E. Ritter, R. J. S. e. Satava, and o. i. techniques, "Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training," vol. 17, pp. 1525-1529, 2003.

[34] D. A. Cook and R. J. A. i. s. Hatala, "Validation of educational assessments: a primer for simulation and beyond," vol. 1, no. 1, pp. 1-12, 2016.

[35] S. M. J. M. e. Downing, "Validity: on the meaningful interpretation of assessment data," vol. 37, no. 9, pp. 830-837, 2003.

[36] M. Slater, S. J. P. T. Wilbur, and V. Environments, "A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments," vol. 6, no. 6, pp. 603-616, 1997.

[37] P. Kourtesis, S. Collina, L. A. Doumas, and S. E. J. F. i. H. N. MacPherson, "Technological competence is a pre-condition for effective implementation of virtual reality head mounted displays in human neuroscience: a technological review and meta-analysis," vol. 13, p. 342, 2019.

[38] F. J. Carter *et al.*, "Consensus guidelines for validation of virtual reality surgical simulators," *Surgical Endoscopy And Other Interventional Techniques,* vol. 19, no. 12, pp. 1523-1532, 2005/12/01 2005.

[39] O. Søvik *et al.*, "Virtual reality simulation training in stroke thrombectomy centers with limited patient volume—Simulator performance and patient outcome," p. 15910199231198275, 2023.

[40] S. S. Van Nortwick, T. S. Lendvay, A. R. Jensen, A. S. Wright, K. D. Horvath, and S. Kim, "Methodologies for establishing validity in surgical simulation studies," *Surgery,* vol. 147, no. 5, pp. 622-630, 2010/05/01/ 2010.

[41] O. Søvik *et al.*, "Virtual reality simulation training in stroke thrombectomy centers with limited patient volume—Simulator performance and patient outcome," vol. 0, no. 0, p. 15910199231198275.

[42] R. Crossley *et al.*, "Validation studies of virtual reality simulation performance metrics for mechanical thrombectomy in ischemic stroke," vol. 11, no. 8, pp. 775-780, 2019.

[43] T. Liebig *et al.*, "Metric-based virtual reality simulation: a paradigm shift in training for mechanical thrombectomy in acute stroke," vol. 49, no. 7, pp. e239-e242, 2018.

[44] S. Chawla, S. Devi, P. Calvachi, W. B. Gormley, and R. Rueda-Esteban, "Evaluation of simulation models in neurosurgical training according to face, content, and construct validity: a systematic review," *Acta Neurochirurgica,* vol. 164, no. 4, pp. 947-966, 2022/04/01 2022.

[45] B. Stew, S. S.-T. Kao, N. Dharmawardana, and E. H. Ooi, "A systematic review of validated sinus surgery simulators," vol. 43, no. 3, pp. 812-822, 2018.

[46] C. Huang, H. Cheng, Y. Bureau, H. M. Ladak, and S. K. Agrawal, "Automated Metrics in a Virtual-Reality Myringotomy Simulator: Development and Construct Validity," (in eng), *Otol Neurotol,* vol. 39, no. 7, 2018.

[47] R. M. Kwasnicki, R. Aggarwal, T. M. Lewis, S. Purkayastha, A. Darzi, and P. A. Paraskeva, "A Comparison of Skill Acquisition and Transfer in Single Incision and Multi-port Laparoscopic Surgery," *Journal of Surgical Education,* vol. 70, no. 2, pp. 172-179, 2013/03/01/ 2013.

[48] N. Mirchi *et al.*, "Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance," *Operative Neurosurgery,* vol. 19, no. 1, pp. 65-75, 2019.

[49] N. Mirchi, V. Bissonnette, R. Yilmaz, N. Ledwos, A. Winkler-Schwartz, and R. F. Del Maestro, "The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine," *PLOS ONE,* vol. 15, no. 2, 2020.

[50] W. Z. Ray, A. Ganju, J. S. Harrop, and D. J. Hoh, "Developing an Anterior Cervical Diskectomy and Fusion Simulator for Neurosurgical Resident Training," *Neurosurgery,* vol. 73, no. suppl_1, pp. S100-S106, 2013.

[51] A. Reich *et al.*, "Artificial Neural Network Approach to Competency-Based Training " 2020.

[52] R. Sawaya *et al.*, "Development of a performance model for virtual reality tumor resections," *Journal of Neurosurgery,* vol. 131, no. 1, pp. 192-200, 2018.

[53] A. Winkler-Schwartz *et al.*, "Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation," (in eng), *J Surg Educ,* vol. 76, no. 6, pp. 1681-1690, Nov-Dec 2019.

[54] J. D. Olden, M. K. Joy, and R. G. Death, "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data," *Ecological modelling,* vol. 178, no. 3-4, pp. 389-397, 2004.

[55] S. J. Russell and P. Norvig, *Artificial intelligence a modern approach*. London, 2010.

[56] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of electronic imaging,* vol. 16, no. 4, p. 049901, 2007.

[57] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[58] M. E. Celebi and K. Aydin, *Unsupervised learning algorithms*. Springer, 2016.

[59] X. Zhu, A. B. J. S. l. o. a. i. Goldberg, and m. learning, "Introduction to semi-supervised learning," vol. 3, no. 1, pp. 1-130, 2009.

[60] C. J. S. l. o. a. i. Szepesvári and m. learning, "Algorithms for reinforcement learning," vol. 4, no. 1, pp. 1-103, 2010.

[61] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[62] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[63] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," vol. 35, no. 5, pp. 1299-1312, 2016.

[64] J. Heaton, S. McElwee, J. Fraley, and J. Cannady, "Early stabilizing feature importance for TensorFlow deep neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 4618-4624.

[65] L. J. M. l. Breiman, "Random forests," vol. 45, no. 1, pp. 5-32, 2001.

[66] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," *Journal of Machine Learning Research,* vol. 20, no. 177, pp. 1-81, 2019.

[67] J. D. Olden and D. A. Jackson, "Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks," *Ecological Modelling,* vol. 154, no. 1, pp. 135-150, 2002/08/15/ 2002.

[68] O. Ibrahim, "A comparison of methods for assessing the relative importance of input variables in artificial neural networks," *Journal of Applied Sciences Research,* vol. 9, no. 11, pp. 5692-5700, 2013.

[69] S. Alkadri *et al.*, "Utilizing a multilayer perceptron artificial neural network to assess a virtual reality surgical procedure," *Computers in Biology and Medicine*, vol. 136, p. 104770, 2021/09/01/ 2021.

[70] A. Winkler-Schwartz *et al.*, "Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation," *JAMA Network Open,* vol. 2, no. 8, 2019.

[71] V. Bissonnette *et al.*, "Artificial intelligence distinguishes surgical training levels in a virtual reality spinal task," vol. 101, no. 23, p. e127, 2019.

[72] N. Mirchi *et al.*, "Artificial neural networks to assess virtual reality anterior cervical discectomy performance," vol. 19, no. 1, pp. 65-75, 2020.

[73] J. Chan *et al.*, "A systematic review of virtual reality for the assessment of technical skills in neurosurgery," vol. 51, no. 2, p. E15, 2021.

[74] E. Bilgic *et al.*, "Exploring the roles of artificial intelligence in surgical education: A scoping review," 2021.

[75] A. Winkler-Schwartz *et al.*, "Bimanual Psychomotor Performance in Neurosurgical Resident Applicants Assessed Using NeuroTouch, a Virtual Reality Simulator," *Journal of Surgical Education,* vol. 73, no. 6, pp. 942-953, 2016/11/01/ 2016.

[76] R. Yilmaz *et al.*, "Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation," *npj Digital Medicine,* vol. 5, no. 1, p. 54, 2022/04/26 2022.

[77] A. M. Fazlollahi *et al.*, "Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial," *JAMA Network Open,* vol. 5, no. 2, pp. e2149008-e2149008, 2022.

[78] A. J. Talia, M. L. Wong, H. C. Lau, and A. H. Kaye, "Comparison of the different surgical approaches for lumbar interbody fusion," *Journal of Clinical Neuroscience,* vol. 22, no. 2, pp. 243-251, 2015/02/01/ 2015.

[79] J. Quillo-Olvera, D. Quillo-Olvera, J. Quillo-Reséndiz, and M. Barrera-Arreola, "Unilateral Biportal Endoscopic Transforaminal Lumbar Interbody Fusion: Technique, Variants, and Navigation," in *Unilateral Biportal Endoscopy of the Spine: An Atlas of Surgical Techniques*, J. Quillo-Olvera, D. Quillo-Olvera, J. Quillo-Reséndiz, and M. Mayer, Eds. Cham: Springer International Publishing, 2022, pp. 389-421.

[80] R. Morgenstern, C. Morgenstern, and A. T. Yeung, "The Learning Curve in Foraminal Endoscopic Discectomy: Experience Needed to Achieve a 90% Success Rate," vol. 1, no. 3, pp. 100-107, 2007.

[81] M. Driscoll. (2019). *Musculoskeletal Biomechanics Research Lab*. Available: https://mcgill.ca/mbr/research/applied-research

[82] M. M. J. J. o. s. d. Panjabi, "The stabilizing system of the spine. Part I. Function, dysfunction, adaptation, and enhancement," vol. 5, pp. 383-383, 1992.

[83] I. Jasiuk, "Micromechanics of bone modeled as a composite material," in *Micromechanics and Nanomechanics of Composite Solids*: Springer, 2018, pp. 281-306.

[84] E. M. M. Van Lieshout, G. H. Van Kralingen, Y. El-Massoudi, H. Weinans, and P. Patka, "Microstructure and biomechanical characteristics of bone substitutes for trauma and orthopaedic surgery," *BMC Musculoskeletal Disorders,* vol. 12, no. 1, p. 34, 2011/02/02 2011.

[85] E. Hamed, Y. Lee, and I. Jasiuk, "Multiscale modeling of elastic properties of cortical bone," *Acta Mechanica,* journal article vol. 213, no. 1, pp. 131-154, August 01 2010.

[86] E. Hamed and I. Jasiuk, "Elastic modeling of bone at nanostructural level," *Materials Science and Engineering: R: Reports,* vol. 73, no. 3, pp. 27-49, 2012/03/22/ 2012.

[87] K. Piekarski, "Analysis of bone as a composite material," *International Journal of Engineering Science,* vol. 11, no. 6, pp. 557-565, 1973/06/01/ 1973.

[88] N. Newell, J. P. Little, A. Christou, M. A. Adams, C. J. Adam, and S. D. Masouros, "Biomechanics of the human intervertebral disc: A review of testing techniques and results," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 69, pp. 420-434, 2017/05/01/ 2017.

[89] H. Schmidt, F. Galbusera, A. Rohlmann, and A. J. J. o. b. Shirazi-Adl, "What have we learned from finite element model studies of lumbar intervertebral discs in the past four decades?," vol. 46, no. 14, pp. 2342-2355, 2013.

[90] W. K. Durfee and K. I. Palmer, "Estimation of force-activation, force-length, and force-velocity properties in isolated, electrically stimulated muscle," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 3, pp. 205-216, 1994.

[91] K. El-Monajjed and M. Driscoll, "Analysis of Surgical Forces Required to Gain Access using a Probe for Minimally Invasive Spine Surgery via Cadaveric-based Experiments towards use in Training Simulators," *IEEE Transactions on Biomedical Engineering,* pp. 1-1, 2020.

[92] M. Hong, J. W. Rozenblit, and A. J. Hamilton, "Simulation-based surgical training systems in laparoscopic surgery: a current review," *Virtual Reality,* vol. 25, no. 2, pp. 491-510, 2021/06/01 2021.

[93] J. Zhang, Y. Zhong, and C. Gu, "Deformable Models for Surgical Simulation: A Survey," *IEEE Reviews in Biomedical Engineering,* vol. 11, pp. 143-164, 2018.

[94] K. El-Monajjed and M. J. J. o. C. S. Driscoll, "Haptic integration of data-driven forces required to gain access using a probe for minimally invasive spine surgery via cadaveric-based experiments towards use in surgical simulators," vol. 60, p. 101569, 2022.

[95] K. El-Monajjed, "Implementation of a virtual reality module for gaining surgical access via planned oblique lateral lumbar interbody fusion," 2021.

[96] V. Favier, G. Subsol, M. Duraes, G. Captier, and P. Gallet, "Haptic Fidelity: The Game Changer in Surgical Simulators for the Next Decade?," (in English), Opinion vol. 11, 2021-August-11 2021.

[97] N. Choudhury, N. Gélinas-Phaneuf, S. Delorme, and R. Del Maestro, "Fundamentals of neurosurgery: virtual reality tasks for training and evaluation of technical skills," *World Neurosurgery*, vol. 80, no. 5, 2013.

[98] T. Cotter, R. Mongrain, and M. Driscoll, "Design synthesis of a robotic uniaxial torque device for orthopedic haptic simulation," *Journal of Medical Devices,* vol. 16, no. 3, p. 031008, 2022.

[99] B. Stott, M. J. M. Driscoll, B. Engineering, and Computing, "Face and content validity of analog surgical instruments on a novel physics-driven minimally invasive spinal fusion surgical simulator," vol. 60, no. 10, pp. 2771-2778, 2022.

[100] R. J. Rueda Esteban, J. S. López-McCormick, A. S. Rodríguez-Bermeo, M. Andrade, J. D. Hernández Restrepo, and E. M. Targarona Soler, "Face, Content, and Construct Validity Evaluation of Simulation Models in General Surgery Laparoscopic Training and Education: A Systematic Review," vol. 30, no. 2, pp. 251-260, 2023.

[101] F. J. Seagull and D. Rooney, "Filling a void: Developing a standard subjective assessment tool for surgical simulation through focused review of current practices," *Surgery*, vol. 156, pp. 718-22, 09/01 2014.

[102] N. J. Borgersen *et al.*, "Gathering validity evidence for surgical simulation: a systematic review," vol. 267, no. 6, pp. 1063-1068, 2018.

[103] A. S. Rao, A. L. R. Michael, and J. Timothy, "Surgical Technique of Anterior Cervical Discectomy and Fusion (ACDF)," in *Practical Procedures in Elective Orthopedic Surgery: Upper Extremity and Spine*, P. V. Giannoudis, Ed. London: Springer London, 2012, pp. 189-193.

[104] J. W. Park, H. S. Nam, S. K. Cho, H. J. Jung, B. J. Lee, and Y. J. A. o. r. m. Park, "Kambin's triangle approach of lumbar transforaminal epidural injection with spinal stenosis," vol. 35, no. 6, pp. 833-843, 2011.

[105] M. W. Beck, "NeuralNetTools: Visualization and Analysis Tools for Neural Networks," *Journal of statistical software,* vol. 85, no. 11, p. 20, 2018-07-30 2018.

[106] B. Stott *et al.*, "A Critical Comparison of Comparators Used to Demonstrate Credibility of Physics-Based Numerical Spine Models," vol. 51, no. 1, pp. 150-162, 2023.

[107] H. V. Chorney, J. R. Forbes, M. J. C. i. B. Driscoll, and Medicine, "System identification and simulation of soft tissue force feedback in a spine surgical simulator," vol. 164, p. 107267, 2023.