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.

Outline

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.

2 Material and Methods

2.1 The Virtual Reality Simulator & The Simulated Scenario

This study utilized the Sim-Ortho VR simulat or 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.

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 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 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 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

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

Table 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)

2.3 AI Analysis

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

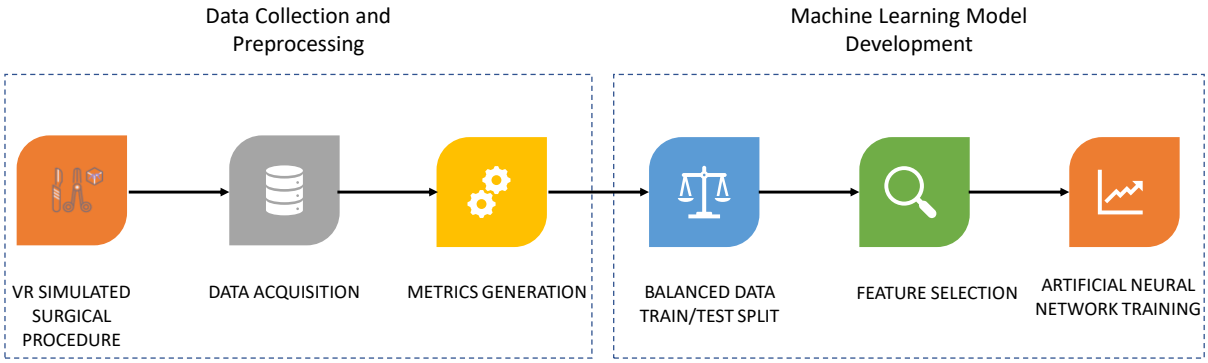


Figure 1 The study methodology consisted of two main steps: Data Collection & Preprocessing and Machine Learning Model Development

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).

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 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 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 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.

Table 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	$ContactLength_{C4}$

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.

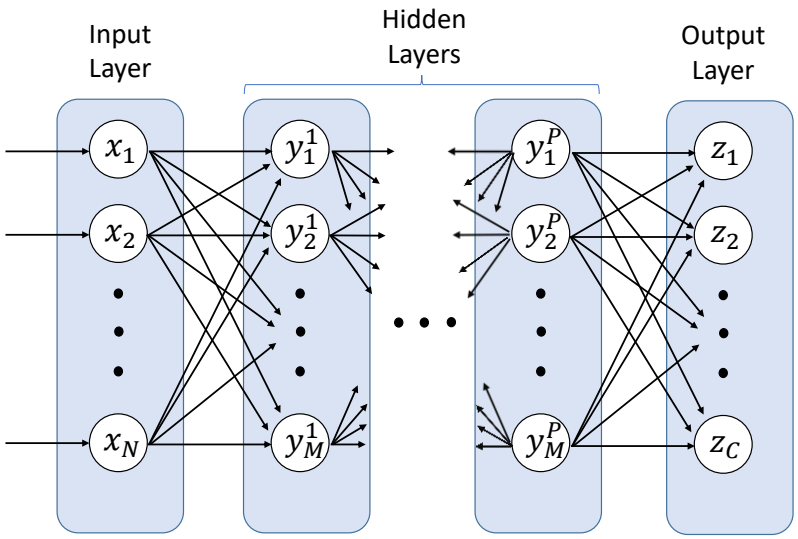


Figure 2 A general MLP diagram showing the input layer, the hidden layers and the interconnected hidden units, and the output layer.

An MLP architecture consists of multiple interconnected hidden neurons within multiple layers as presented in Figure 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 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 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

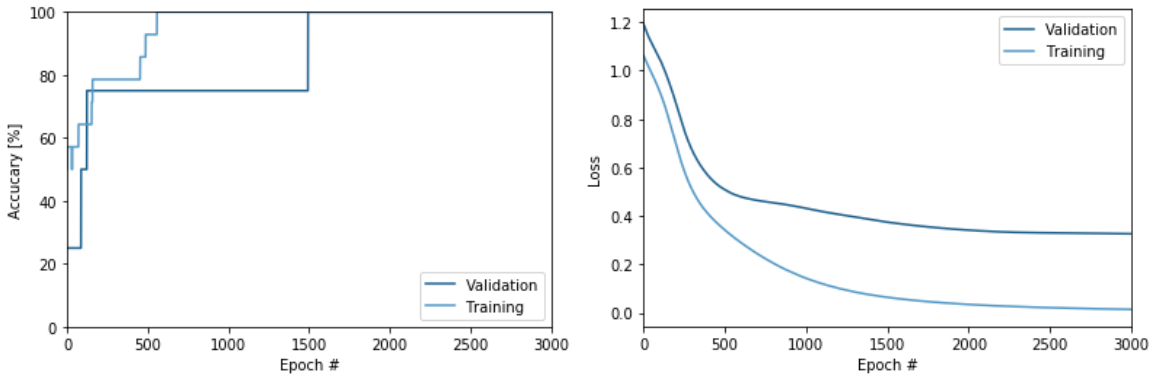
Table 6 presents the best performing models found based on the search criteria in the one-layered, two-layered, and three-layered ANNs. As seen in Table 6, the two-layered network resulted in a better model performance on the validation set. Table 7 shows the chosen model with the best hyperparameters. Figure 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 3).

Table 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 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 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 and Equation 1.

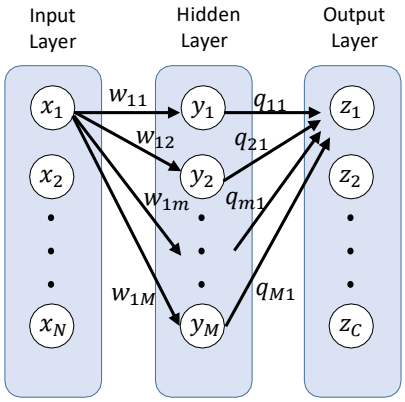


Figure 4 Schematic of a one hidden layer network demonstrating the weights that connect the first input node to the first output node.

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

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 5 and Equation 2:

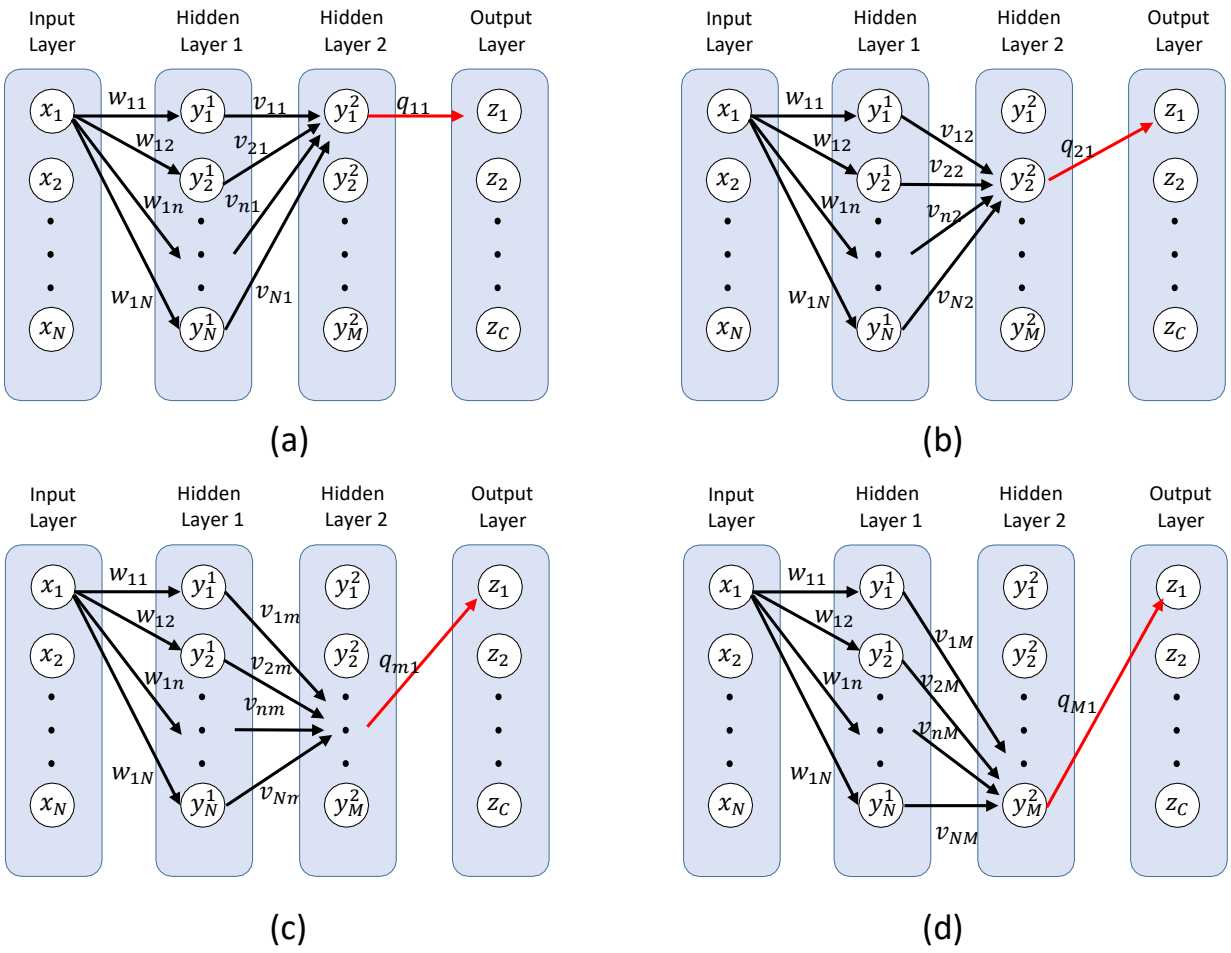


Figure 5 Schematic of a two hidden layer network demonstrating the weights that connect the first input node to the first output node. To simplify the illustration, the connection weights are broken into multiple schematics (a-d) by varying the last hidden layer m from 1 to M.

$$CWP_{x,z} = \sum_{m=1}^{M} \sum_{n=1}^{N} w_{xn} v_{nm} q_{mz} \qquad \text{Eq (2)}$$

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 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.

3 Results

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. 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:

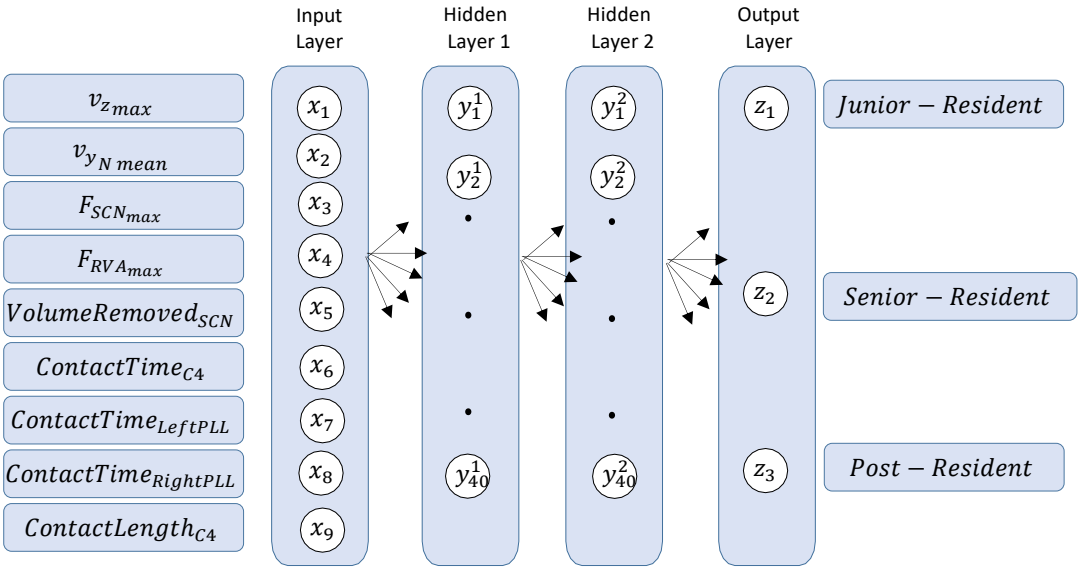


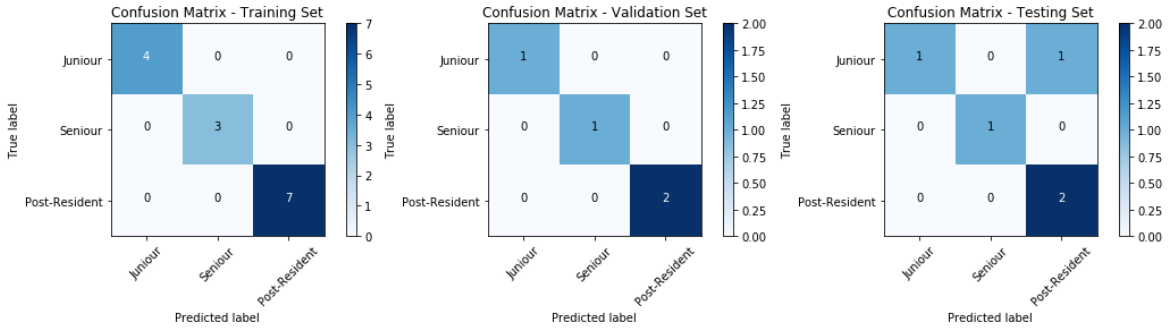
Figure 6 Model architecture of the final developed ANN model demonstrating the input surgical metrics, the number of hidden units and layers, as well as the output variables.

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

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 9, Table 10, and Table 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 12 and Table 13 present the permutation feature importance applied to the training and testing sets, respectively. Figure 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 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 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 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 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 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%

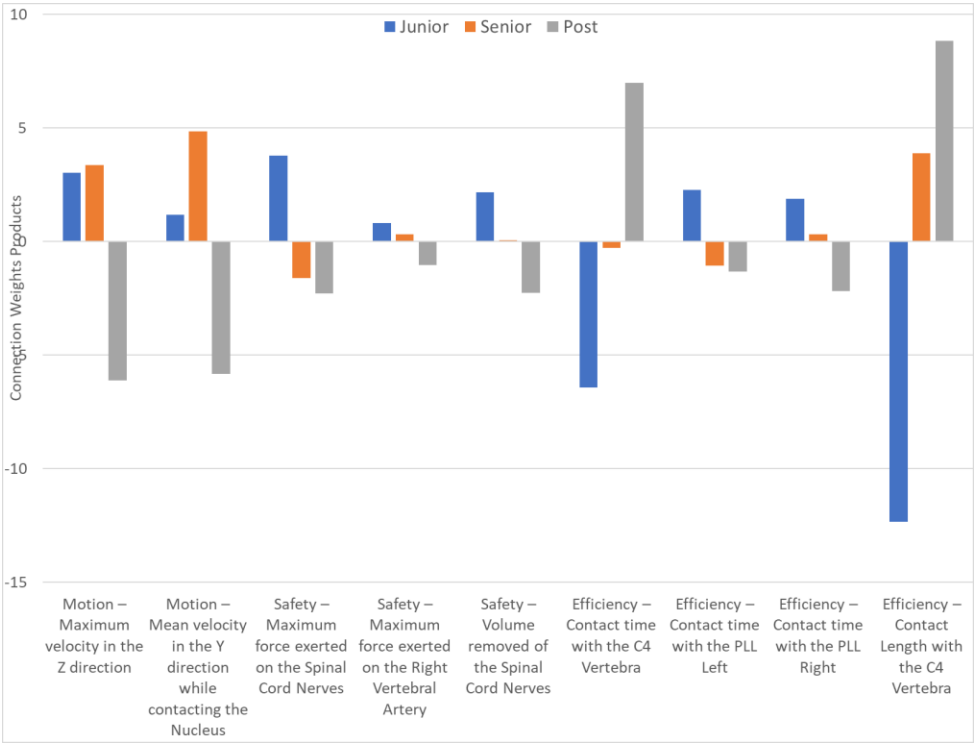


Figure 8 Learning patterns of the Connection Weights Products for each input feature.

4 Discussion

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 9, Table 11, and Table 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 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 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 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.

4.2.1 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 Tables 8-10, 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].

4.2.2 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 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 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.

4.2.3 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 12 and Table 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 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 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.

4.3 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 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 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 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

5 Limitations

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 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.

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.

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.

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. 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. 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.