Computers in Biology and Medicine 179 (2024) 108809

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

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 [a,c], Rolando F. Del Maestro [c], Mark Driscoll [a,b,*]

[a] Musculoskeletal Biomechanics Research Lab, Department of Mechanical Engineering, McGill University, Macdonald Engineering Building, 815 Sherbrooke St W, Montreal, H3A 2K7, QC, Canada
[b] Orthopaedic Research Lab, Montreal General Hospital, 1650 Cedar Ave (LS1.409), Montreal, H3G 1A4, Quebec, Canada
[c] Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology & Neurosurgery, Montreal Neurological Institute, McGill University, 2200 Leo Pariseau, Suite, 2210, Montreal, H2X 4B3, Quebec, Canada

ARTICLE INFO

Keywords:
Multilayered artificial neural network
Transfer learning
Data augmentation
Feature importance
Virtual reality
Surgical simulation
Surgical education
Performance metric
Surgical expertise

ABSTRACT

Background: Virtual and augmented reality surgical simulators, integrated with machine learning, are becoming essential for training psychomotor skills, and analyzing surgical performance. Despite the promise of methods like the Connection Weights Algorithm, the small sample sizes (small number of participants (N)) typical of these trials challenge the generalizability and robustness of models. Approaches like data augmentation and transfer learning from models trained on similar surgical tasks address these limitations.
Objective: 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 artificial neural network algorithms within a novel simulator platform, using data from the simulated tasks to generate 276 performance metrics across motion, safety, and efficiency. Innovatively, it applies transfer learning from a pre-trained ANN model developed for a similar spinal simulator, enhancing the training process, and addressing the challenge of small datasets.
Setting: Musculoskeletal Biomechanics Research Lab; Neurosurgical Simulation and Artificial Intelligence Learning Centre, McGill University, Montreal, Canada.
Participants: Twenty-seven participants 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 presents a novel blueprint for addressing limited datasets in surgical simulations through the strategic use of transfer learning and data augmentation. It also evaluates and reinforces the application of the Connection Weights Algorithm from our previous publication. Together, these methodologies not only enhance the precision of performance classification but also advance the validation of surgical training platforms.

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].

* Corresponding author. Macdonald Engineering Building, RM 153, 817 Sherbrooke St W, Montreal, Quebec, H3A 2K7, Canada.
E-mail address: mark.driscoll@mcgill.ca (M. Driscoll).

https://doi.org/10.1016/j.compbiomed.2024.108809
Received 27 March 2024; Received in revised form 10 June 2024; Accepted 24 June 2024

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.

2. Material and methods

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, (Fig. 1a). 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 Fig. 1b.

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

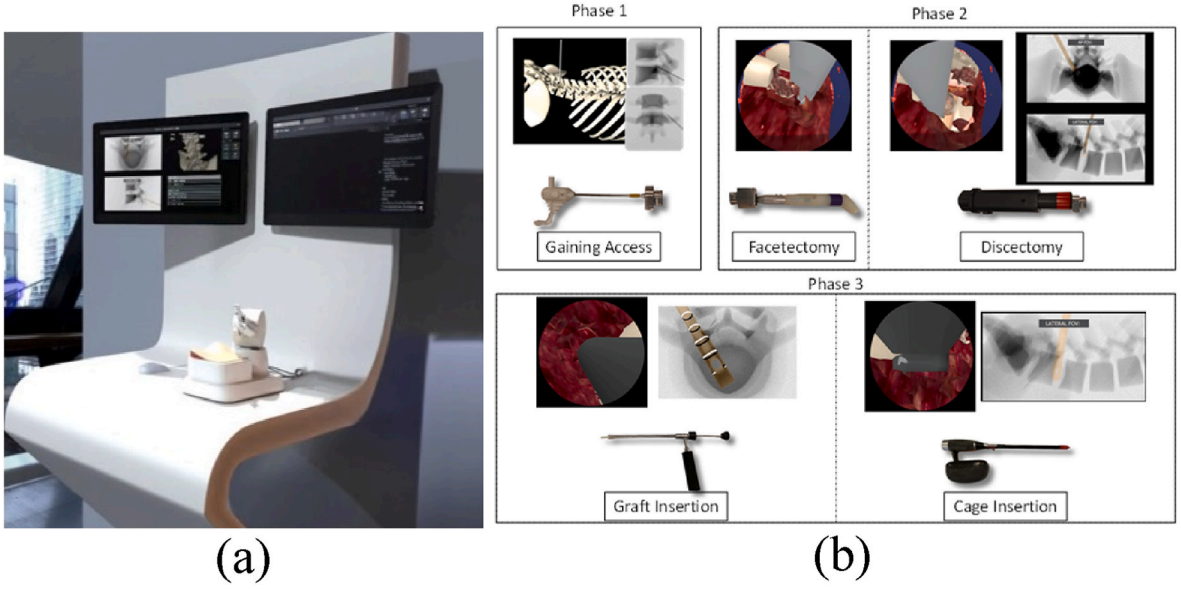

Fig. 1. (a) Simulator layout. Right screen indicates the instruction of the surgery process. The haptic device and benchtop model are in the middle. Left screen indicates the four camera views that demonstrate the surgical area. (b) The three phases of the simulated surgery: Phase 1 includes gaining 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.

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.

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 1 and Table 2 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 min 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].

2.3. Machine learning analysis

A systematic approach was used in integrating a MLP ANN in classifying the virtual surgical performance. As illustrated in Fig. 2, 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].

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

	Junior Residents (n = 9)	Senior Residents (n = 6)	Post-Residents (n = 12)
Male	8	5	11
Female	1	1	1
Level of Training	Surgical Specialty		
	Neurosurgery	Orthopaedic Surgery	
PGY 1-3	4	5	
PGY 4-6	4	2	
Fellows	2	2	
Consultants	3	5	

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:			
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)
TLIF comfort level with a consultant in the room	3.0 (1.0–4.0)	4.0 (2.0–5.0)	4.5 (2.0–5.0)
TLIF comfort level solo	1.0 (1.0–2.0)	2.0 (1.0–4.0)	3.0 (1.0–5.0)

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]. In simulation, 73 variables were recorded including tool position, time, angles, forces, volumes removed, and contacts with anatomical structures. This data was processed to generate metrics assessing participant performance. For example, tool position and time data were used to calculate velocities, forces and contacts assessed removal effectiveness of structures, and combined position and contacts measured interaction path lengths. 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.

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 Fig. 2, 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 3). 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).

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

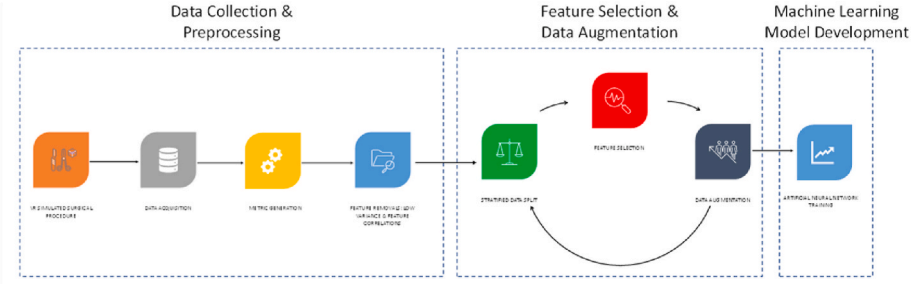

Fig. 2. The study methodology consists of three main steps: Data Collection & Preprocessing, Feature Selection & Data Augmentation, and Machine Learning Model Development.

Table 3
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
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 %

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. This value was chosen to mimic realistic variations expected in surgical settings, such as those due to hand tremors, tool handling, or dexterity control. By introducing a jitter that represents only 1 % of the standard deviation in the normalized data, the model effectively incorporates subtle yet significant variations that enhance its robustness and generalizability without compromising the integrity of the data. 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 5 and Table 6 respectively.

Table 5
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

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.

An MLP architecture consists of multiple interconnected hidden neurons within multiple layers as presented in Fig. 3. 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 7 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,

Table 6
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	$F_{SAP\,BurrTool_{mean}}$

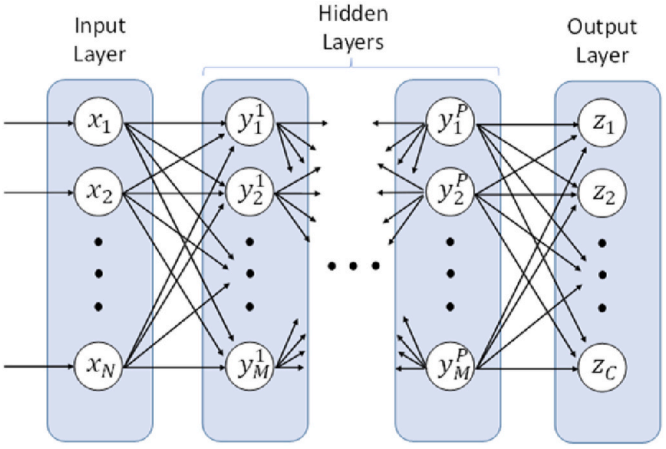

Fig. 3. A general MLP diagram showing the input layer, the hidden layers and the interconnected hidden units, and the output layer.

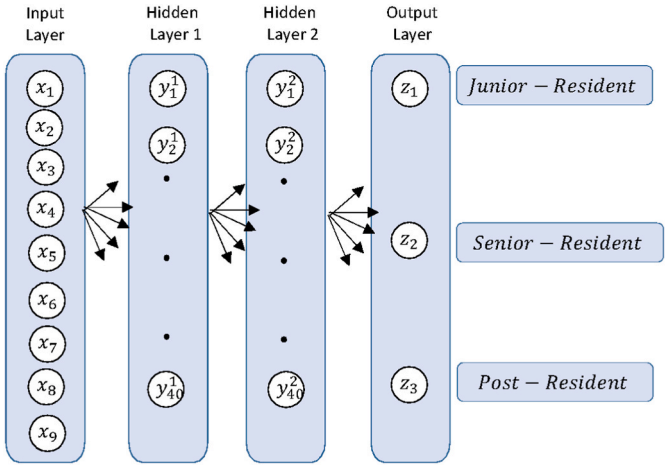

Fig. 4. Pre-trained model architecture.

Table 7
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

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.

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 8 and Fig. 4. 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 a important drawback of this method when dealing with ANNs with few layers applied on small datasets, as in the current application.

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.

Table 8
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**

Table 9 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. Table 10 details the hyperparameters of the two top-performing models, including architecture and optimization parameters.

Fig. 5 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 (Fig. 5).

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 Fig. 6 and Equation (1). 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 Fig. 7 and Equation (2), 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.

$$CWP_{x,z} = \sum_{m=1}^{M} w_{xm} q_{mz} \tag{1}$$

$$CWP_{x,z} = \sum_{m=1}^{M} \sum_{n=1}^{N} w_{xn} v_{nm} q_{mz} \tag{2}$$

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

Table 9
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	**1**Ŧ	**0.0005**	**0.6**	**100 %**	**0.01**
	6	**2**Ŧ	**0.001**	**0.6**	**83.33 %**	**0.04**
	20	**3**Ŧ	**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 10
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.

Fig. 7 and Equation (2), 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:

$$CWP_{x,z} = \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} \cdots \sum_{i_L=1}^{N_L} w_{xi_1}^{(0)} w_{i_1 i_2}^{(1)} \cdots w_{i_{L-1} i_L}^{(L-1)} w_{i_L z}^{(L)} \quad (3)$$

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

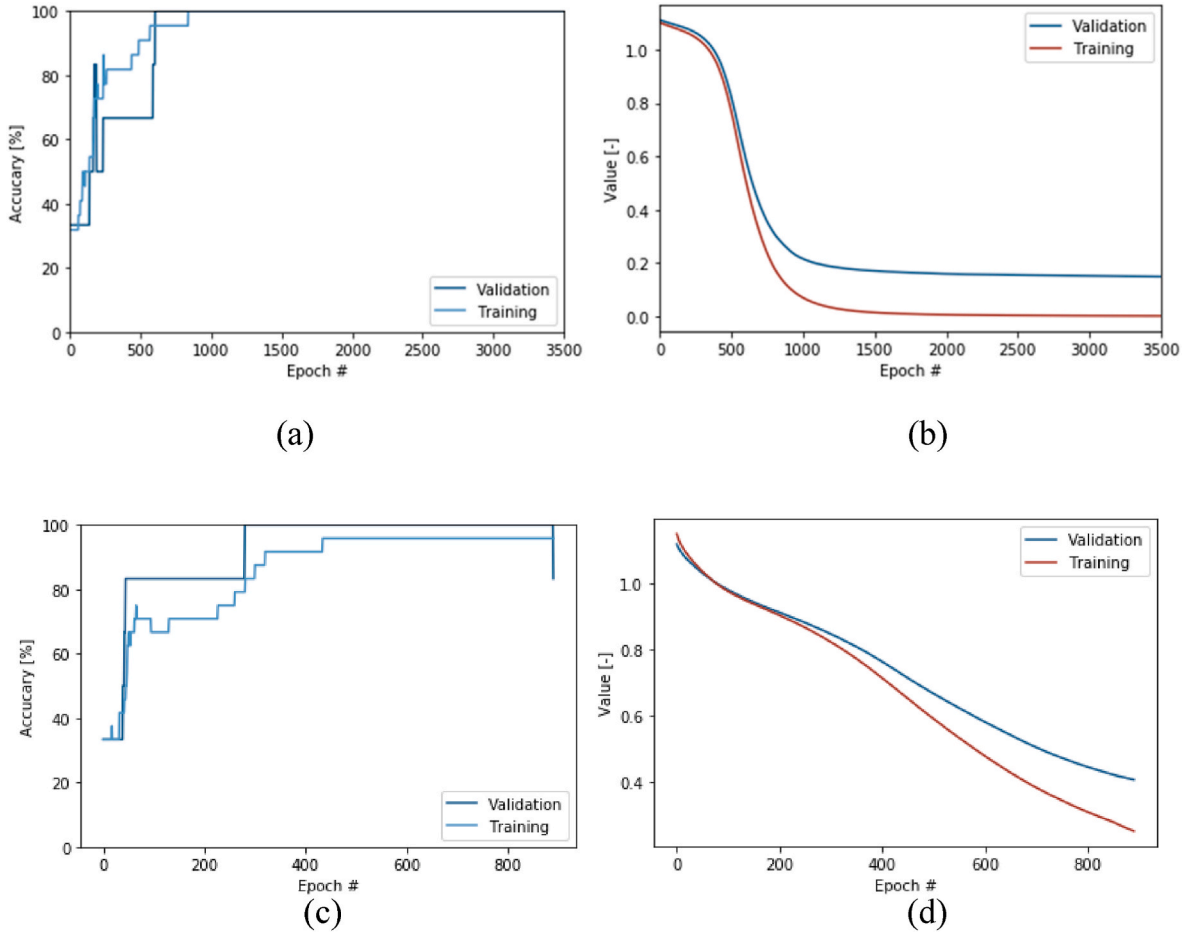

Fig. 5. The performance of the models at each training epoch: (a) the accuracy of the optimal stand-alone model on the training and validation sets at each training epoch; (b) the value of the loss function of optimal stand-alone model on the training and validation sets at each training epoch; (c) the 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.

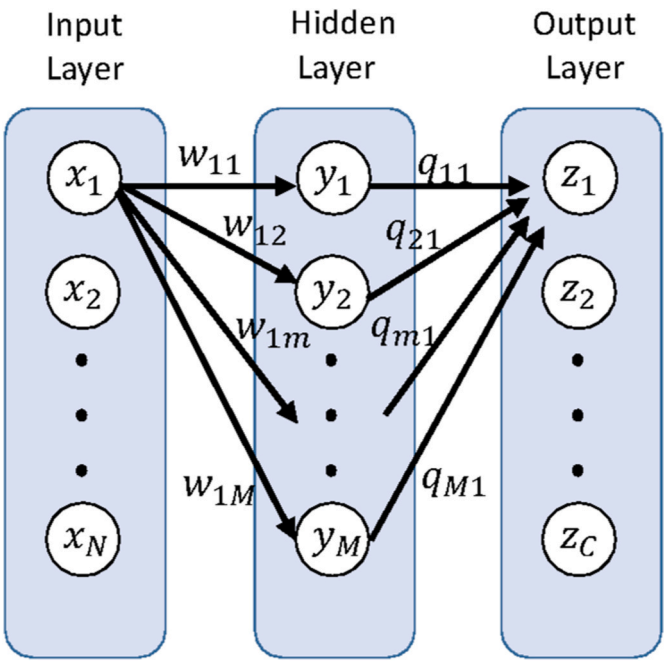

Fig. 6. Schematic of a one hidden layer network demonstrating the weights that connect the first input node to the first output node.

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.

3. Results

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 6). 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 presented in Figs. 8 and 9.

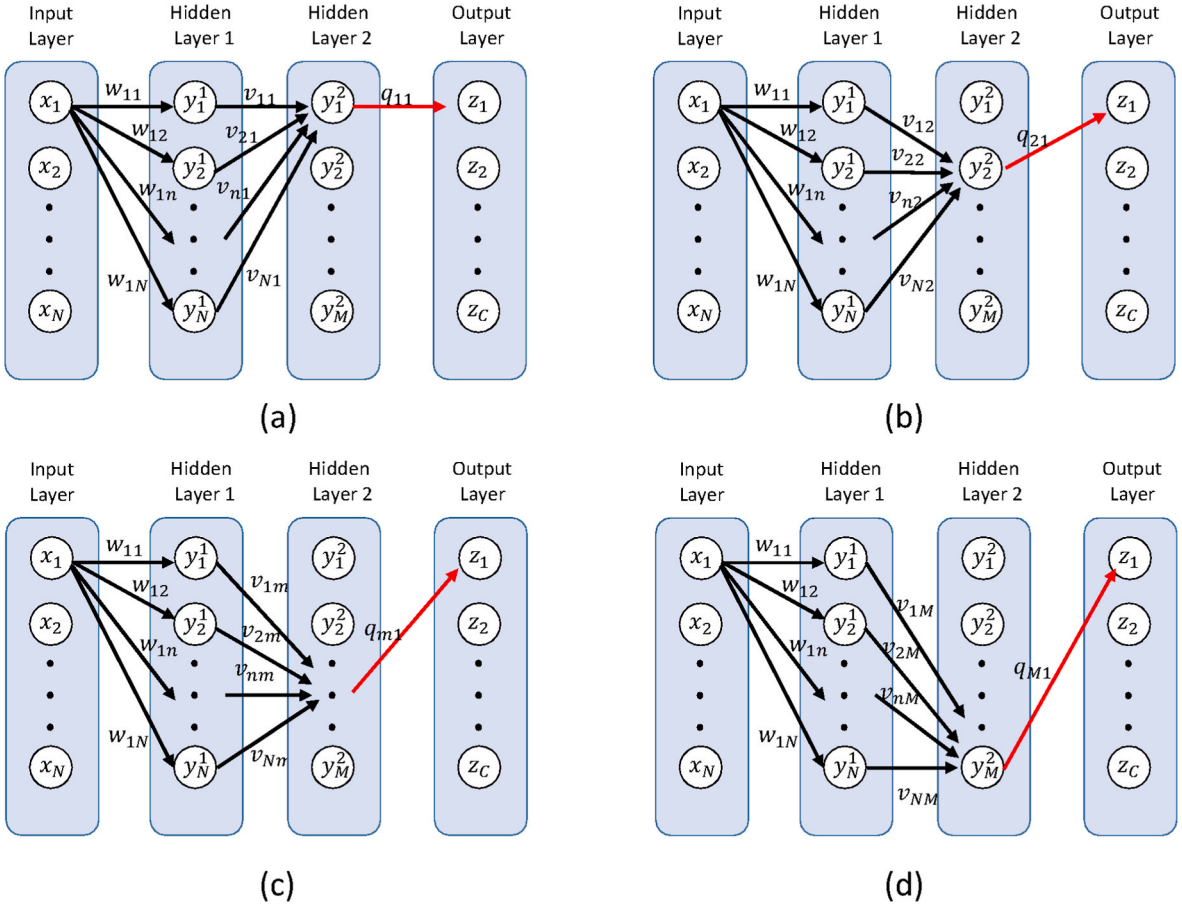

Fig. 7. Schematic of a two hidden layer network demonstrating the weights that connect the first input node to the first output node. To simplify the illustration, the connection weights are broken into multiple schematics (a–d) by varying the last hidden layer m from 1 to M.

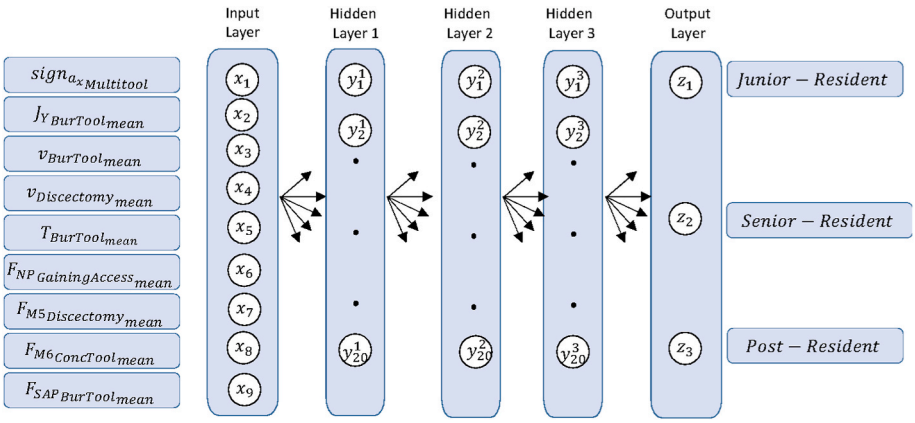

Fig. 8. Model architecture of the final stand-alone MLP ANN model developed from scratch demonstrating the input surgical metrics, the number of hidden units and layers, as well as the output variables.

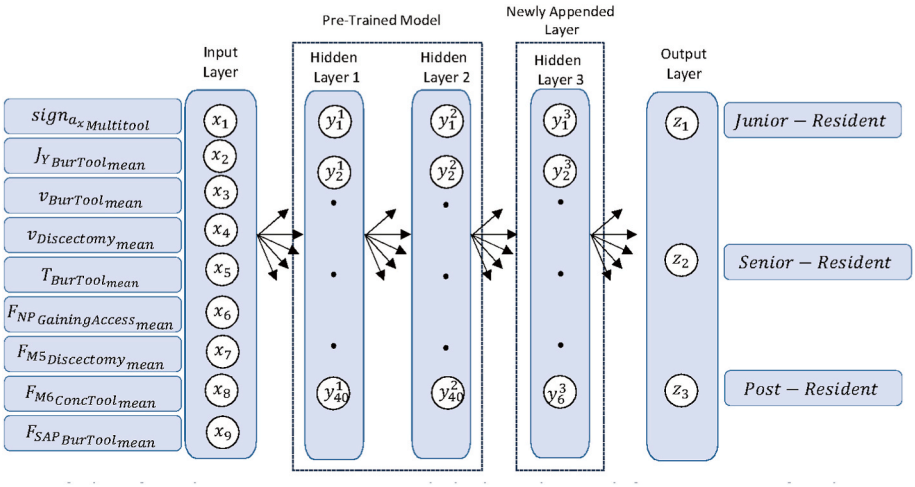

Fig. 9. Model architecture of the final MLP ANN model developed from transfer learning demonstrating the input surgical metrics, the number of hidden units and layers, as well as the output variables.

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 11, with performance visualized in confusion matrices (Figs. 10 and 11). 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 11
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**

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 12, Table 13, and Table 14 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 A1–A10 in Appendix provide detailed CWP values, feature relative importance, and permutation feature importance for the training and testing sets for each surgical class group. Fig. 12 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.

4. Discussion

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

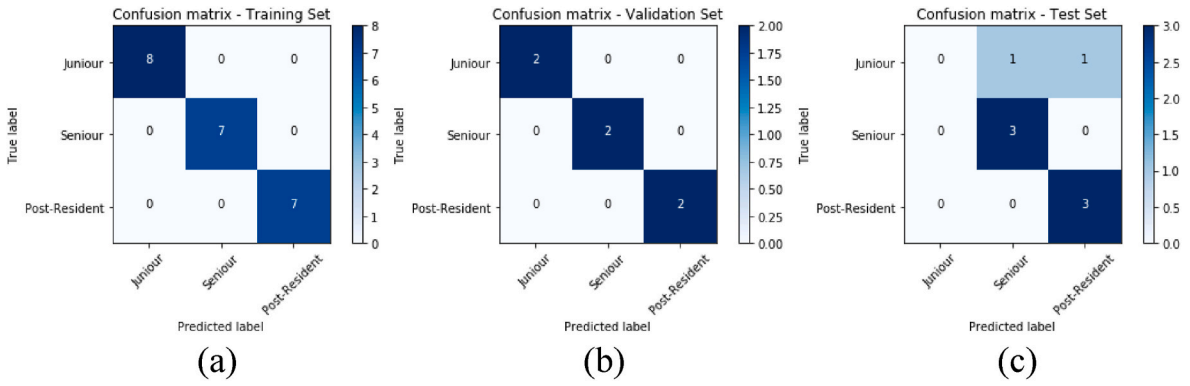

Fig. 10. Confusion matrices highlighting the performance of the stand alone MLP ANN model trained from scratch on the: (a) training set, (b) validation set, and (c) testing set.

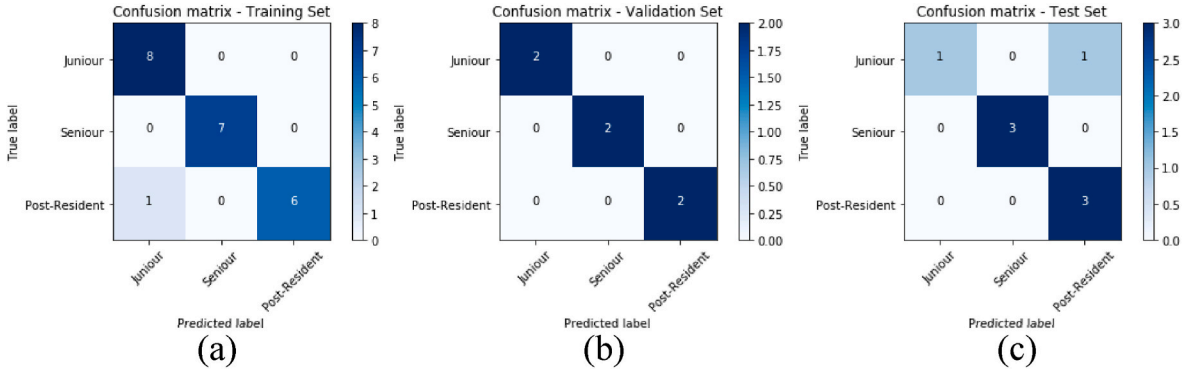

Fig. 11. Confusion matrices highlighting the performance of the MLP ANN model with transfer learning on the: (a) training set, (b) validation set, and (c) testing set.

Table 12
Surgical Performance Metrics Ranking for each model: CWPs & Permutation Importance for Junior Residents.

Rank	Stand-Alone MLP ANN Model			Transfer Learning MLP ANN Model		
	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	$F_{M5\,Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{SAP\,BurTool_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{SAP\,BurTool_{mean}}$	$F_{SAP\,BurTool_{mean}}$
2	$v_{Discectomy_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$
3	$F_{SAP\,BurTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$v_{Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$
4	$v_{BurTool_{mean}}$	$F_{SAP\,BurTool_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{SAP\,BurTool_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{NP\,GainingAccess_{mean}}$
5	$J_{Y\,BurTool_{mean}}$	$T_{BurTool_{mean}}$	$T_{BurTool_{mean}}$	$sign_{\alpha_x\,Multitool}$	$J_{Y\,BurTool_{mean}}$	$T_{BurTool_{mean}}$
6	$sign_{\alpha_x\,Multitool}$	$v_{Discectomy_{mean}}$	$v_{Discectomy_{mean}}$	$J_{Y\,BurTool_{mean}}$	$T_{BurTool_{mean}}$	$v_{Discectomy_{mean}}$
7	$T_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$	$v_{BurTool_{mean}}$
8	$F_{NP\,GainingAccess_{mean}}$	$J_{Y\,BurTool_{mean}}$	$J_{Y\,BurTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$v_{Discectomy_{mean}}$	$J_{Y\,BurTool_{mean}}$
9	$F_{M6\,ConcTool_{mean}}$	$sign_{\alpha_x\,Multitool}$	$sign_{\alpha_x\,Multitool}$	$T_{BurTool_{mean}}$	$sign_{\alpha_x\,Multitool}$	$sign_{\alpha_x\,Multitool}$

Table 13
Surgical Performance Metrics Ranking for each model: CWPs & Permutation Importance for Senior-Residents.

Rank	Stand-Alone MLP ANN Model			Transfer Learning MLP ANN Model		
	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	$F_{M5\,Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{SAP\,BurTool_{mean}}$	$sign_{\alpha_x\,Multitool}$	$F_{SAP\,BurTool_{mean}}$	$F_{SAP\,BurTool_{mean}}$
2	$F_{SAP\,BurTool_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$J_{Y\,BurTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$
3	$J_{Y\,BurTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{M6\,ConcTool_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$F_{M5\,Discectomy_{mean}}$
4	$F_{NP\,GainingAccess_{mean}}$	$F_{SAP\,BurTool_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{SAP\,BurTool_{mean}}$	$F_{NP\,GainingAccess_{mean}}$	$F_{NP\,GainingAccess_{mean}}$
5	$F_{M6\,ConcTool_{mean}}$	$T_{BurTool_{mean}}$	$T_{BurTool_{mean}}$	$F_{M5\,Discectomy_{mean}}$	$J_{Y\,BurTool_{mean}}$	$T_{BurTool_{mean}}$
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	$sign_{\alpha_x\,Multitool}$	$J_{Y\,BurTool_{mean}}$	$J_{Y\,BurTool_{mean}}$	$T_{BurTool_{mean}}$	$v_{Discectomy_{mean}}$	$J_{Y\,BurTool_{mean}}$
9	$v_{Discectomy_{mean}}$	$sign_{\alpha_x\,Multitool}$	$sign_{\alpha_x\,Multitool}$	$F_{M6\,ConcTool_{mean}}$	$sign_{\alpha_x\,Multitool}$	$sign_{\alpha_x\,Multitool}$

Table 14
Surgical Performance Metrics Ranking for each model: CWPs & Permutation Importance for Post-Residents.

Rank	Stand-Alone MLP ANN Model			Transfer Learning MLP ANN Model		
	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	$v_{Discectomy\,mean}$	$F_{M5\,Discectomy\,mean}$	$F_{SAP\,BurTool\,mean}$	$J_{Y\,BurTool\,mean}$	$F_{SAP\,BurTool\,mean}$	$F_{SAP\,BurTool\,mean}$
2	$F_{NP\,GainingAccess\,mean}$	$F_{NP\,GainingAccess\,mean}$	$F_{M5\,Discectomy\,mean}$	$v_{Discectomy\,mean}$	$F_{M6\,ConcTool\,mean}$	$F_{M6\,ConcTool\,mean}$
3	$v_{BurTool\,mean}$	$F_{M6\,ConcTool\,mean}$	$F_{M6\,ConcTool\,mean}$	$F_{NP\,GainingAccess\,mean}$	$F_{M5\,Discectomy\,mean}$	$F_{M5\,Discectomy\,mean}$
4	$J_{Y\,BurTool\,mean}$	$F_{SAP\,BurTool\,mean}$	$F_{NP\,GainingAccess\,mean}$	$F_{M5\,Discectomy\,mean}$	$F_{NP\,GainingAccess\,mean}$	$F_{NP\,GainingAccess\,mean}$
5	$sign_{a_x\,Multitool}$	$T_{BurTool_{mean}}$	$T_{BurTool_{mean}}$	$sign_{a_x\,Multitool}$	$J_{Y\,BurTool\,mean}$	$T_{BurTool_{mean}}$
6	$F_{M6\,ConcTool\,mean}$	$v_{Discectomy\,mean}$	$v_{Discectomy\,mean}$	$v_{BurTool\,mean}$	$T_{BurTool_{mean}}$	$v_{Discectomy\,mean}$
7	$T_{BurTool_{mean}}$	$v_{BurTool\,mean}$	$v_{BurTool\,mean}$	$T_{BurTool_{mean}}$	$v_{BurTool\,mean}$	$v_{BurTool\,mean}$
8	$F_{SAP\,BurTool\,mean}$	$J_{Y\,BurTool\,mean}$	$J_{Y\,BurTool\,mean}$	$F_{SAP\,BurTool\,mean}$	$v_{Discectomy\,mean}$	$J_{Y\,BurTool\,mean}$
9	$F_{M5\,Discectomy\,mean}$	$sign_{a_x\,Multitool}$	$sign_{a_x\,Multitool}$	$F_{M6\,ConcTool\,mean}$	$sign_{a_x\,Multitool}$	$sign_{a_x\,Multitool}$

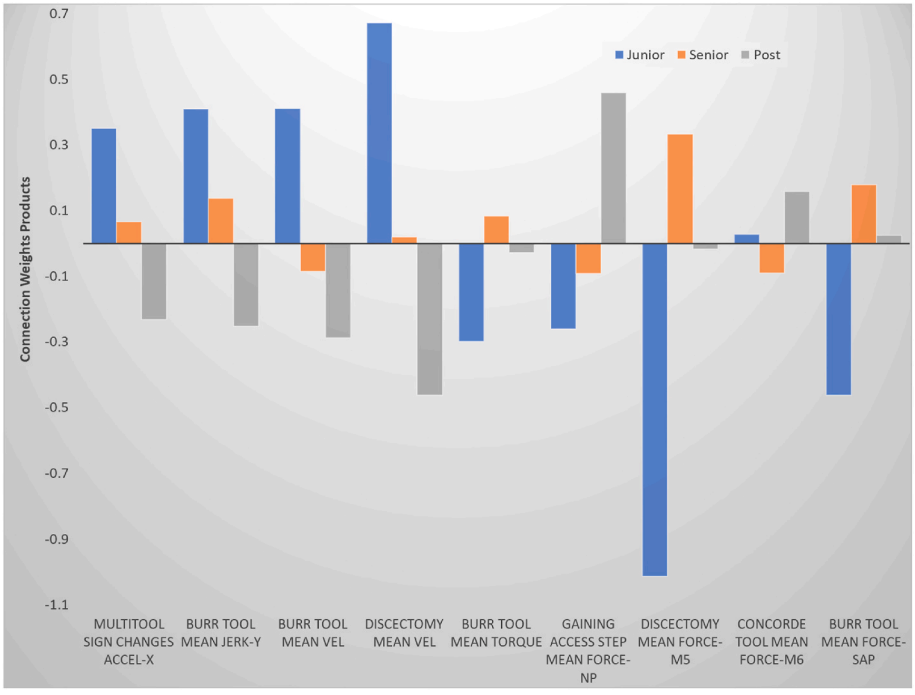

Fig. 12. Learning patterns of the Connection Weights Products for each input feature on the Stand-Alone MLP ANN.

"facetectomy and discectomy" steps of the OLLIF simulation, this study identified nine critical features for neural network development. Using the methodology shown in Fig. 2, 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 %–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 A1 – A3), it was observed that the two misclassified junior-resident individuals exhibited performance traits that resembled senior and post-residents in key overlapping features (Table 15). 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 15). 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

Table 15
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 %)

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.

4.2. Insights and surgical performance patterns revealed by the ANNs

Table 12 to Table 14 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 A1–A6).

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 A7–A10).

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 12 to Table 14 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 4.1, 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 15). 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 standalone and transfer learning models.

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.

Fig. 12 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.

Fig. 12 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.

4.2.3. 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.

5. Limitations

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. Unlike broader applications in fields such as bioinformatics and computational biology, where large datasets typically enable the effective training of deep learning models, our research had to innovate within the constraints of smaller data volumes [31–35]. This situation mirrors challenges in other specialized fields where data scarcity can hinder model performance and applicability. In response, 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.

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.

5.3. Surgical performance metrics

While the current simulator effectively captures and quantifies psychomotor skills, it currently lacks the capability to assess qualitative metrics such as professionalism, communication, and teamwork. These "other skills" are integral to holistic surgical training but pose significant challenges for measurement within simulation environments, primarily due to their subjective nature and the complexity of their assessment. Future enhancements to the simulator could include the integration of technologies that assess these qualitative metrics. For example, incorporating video and audio analysis tools could enable the evaluation of communication skills and team dynamics during simulated procedures. This would provide a more comprehensive training tool that aligns better with the holistic training approaches advocated by leading surgical education bodies. Such developments would require collaborative efforts between engineers, AI specialists, and clinical educators to ensure that the new metrics are not only measurable but also relevant and valuable for attaining surgical expertise beyond technical proficiency.

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, and other domains with similar data constraints. Additionally, the study expanded on the authors' previous work by comparing the new approach 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. This methodology provides a foundation for enhancing surgical training and may be adapted to improve real-time decision-making in live surgical environments.

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.

CRediT authorship contribution statement

Sami Alkadri: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Rolando F. Del Maestro:** Writing – review & editing, Validation, Supervision, Resources. **Mark Driscoll:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

"None declared".

Appendix

Table A1
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 A2
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 A3
Ranked 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 A4
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 A5
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 A6
Ranked 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 A7
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 A8
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 A9
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 A10
Permutation Feature Importance applied on the testing set with 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 %

References

[1] M. Goldenberg, J.Y. Lee, Surgical education, simulation, and simulators-updating the concept of validity, Curr. Urol. Rep. 19 (7) (May 17 2018) 52, https://doi.org/10.1007/s11934-018-0799-7, in eng.

[2] M. Alaker, G.R. Wynn, T. Arulampalam, Virtual reality training in laparoscopic surgery: a systematic review & meta-analysis, Int. J. Surg. 29 (2016/05/01/2016) 85–94, 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," Med. Biol. Eng. Comput.

[4] K. El-Monajjed, 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 (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. (2020) 1, 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, in: Undergraduate Honours Program - Mechanical Engineering, Mechanical Engineering, 2018. McGill University.

[6] N. Ledwos, N. Mirchi, V. Bissonnette, A. Winkler-Schwartz, R. Yilmaz, R.F.J.O. N. Del Maestro, Virtual reality anterior cervical discectomy and fusion simulation on the novel sim-ortho platform: validation studies, Operat. Neurosurg. 20 (1) (2020) 74–82.

[7] H. Azarnoush, et al., Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection, Int. J. Comput. Assist. Radiol. Surg. 10 (5) (May 2015) 603–618, https://doi.org/10.1007/s11548-014-1091-z, in eng.

[8] N. Mirchi, et al., Artificial neural networks to assess virtual reality anterior cervical discectomy performance, Operat. Neurosurg. 19 (1) (2019) 65–75, https://doi.org/10.1093/ons/opz359.

[9] R. Sawaya, et al., Development of a performance model for virtual reality tumor resections, J. Neurosurg. 131 (1) (2018) 192, https://doi.org/10.3171/2018.2.Jns172327, in English.

[10] N. Mirchi, V. Bissonnette, R. Yilmaz, N. Ledwos, A. Winkler-Schwartz, R.F. Del Maestro, The Virtual Operative Assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine, PLoS One 15 (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 Netw. Open 2 (8) (2019), https://doi.org/10.1001/jamanetworkopen.2019.8363.

[12] N.M. Nasrabadi, Pattern recognition and machine learning, J. Electron. Imag. 16 (4) (2007) 049901.

[13] J. Heaton, S. McElwee, J. Fraley, 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, J. Appl. Sci. Res. 9 (11) (2013) 5692–5700.

[15] J.D. Olden, D.A. Jackson, Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks, Ecol. Model. 154 (1) (2002/08/15/2002) 135–150, 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, Comput. Biol. Med. 136 (2021) 104770, https://doi.org/10.1016/j.compbiomed.2021.104770, 2021/09/01/.

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

[18] N. Tajbakhsh, et al., Convolutional neural networks for medical image analysis: full training or fine tuning? 35 (5) (2016) 1299–1312.

[19] A. Winkler-Schwartz, et al., Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation, J. Surg. Educ. 76 (6) (Nov-Dec 2019) 1681–1690, https://doi.org/10.1016/j.jsurg.2019.05.015 (in eng).

[20] E. Von Elm, D.G. Altman, M. Egger, S.J. Pocock, P.C. Gøtzsche, J.P.J.T. L. Vandenbroucke, The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies 370 (9596) (2007) 1453–1457.

[21] A. Paszke, et al., Pytorch: an Imperative Style, High-Performance Deep Learning Library, 2019 *arXiv preprint*.

[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, F. Dominici, All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (177) (2019) 1–81.

[24] J. Chan, et al., A systematic review of virtual reality for the assessment of technical skills in neurosurgery 51 (2) (2021) E15.

[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, Y. Park, Kambin's triangle approach of lumbar transforaminal epidural injection with spinal stenosis 35 (6) (2011) 833–843.

[28] P. Kambin, L.J.C.O. Zhou, R. Research, Arthroscopic discectomy of the lumbar spine 337 (1997) 49–57.

[29] R. Yilmaz, et al., Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation, npj Digit. Med. 5 (1) (2022/04/26 2022) 54, https://doi.org/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 Netw. Open 5 (2) (2022) e2149008, https://doi.org/10.1001/jamanetworkopen.2021.49008.

[31] L. Zheng, et al., AnnoPRO: a strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path

encoding, Genome Biol. 25 (1) (Feb 1 2024) 41, https://doi.org/10.1186/s13059-024-03166-1, in eng.
[32] Y. Wang, et al., A task-specific encoding algorithm for RNAs and RNA-associated interactions based on convolutional autoencoder, Nucleic Acids Res. 51 (21) (Nov 27 2023) e110, https://doi.org/10.1093/nar/gkad929, in eng.
[33] M. Mou, et al., A Transformer-Based Ensemble Framework for the Prediction of Protein-Protein Interaction Sites, 6, Research (Wash D C), 2023, p. 240, https://doi.org/10.34133/research.0240, in eng.
[34] J. Hong, et al., Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Briefings Bioinf. 21 (4) (Jul 15 2020) 1437–1447, https://doi.org/10.1093/bib/bbz081, in eng.
[35] J. Hong, et al., Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery, Briefings Bioinf. 21 (5) (Sep 25 2020) 1825–1836, https://doi.org/10.1093/bib/bbz120, in eng.