Medical & Biological Engineering & Computing (2020) 58:1357–1367
https://doi.org/10.1007/s11517-020-02155-3

ORIGINAL ARTICLE

Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task

Samaneh Siyar[1,2] · Hamed Azarnoush[1,2] · Saeid Rashidi[3] · Alexander Winkler-Schwartz[2] · Vincent Bissonnette[2] · Nirros Ponnudurai[2] · Rolando F. Del Maestro[2]

Received: 5 April 2019 / Accepted: 12 March 2020 / Published online: 11 April 2020

Abstract
This study outlines the first investigation of application of machine learning to distinguish "skilled" and "novice" psychomotor performance during a virtual reality (VR) brain tumor resection task. Tumor resection task participants included 23 neurosurgeons and senior neurosurgery residents as the "skilled" group and 92 junior neurosurgery residents and medical students as the "novice" group. The task involved removing a series of virtual brain tumors without causing injury to surrounding tissue. Originally, 150 features were extracted followed by statistical and forward feature selection. The selected features were provided to 4 classifiers, namely, K-Nearest Neighbors, Parzen Window, Support Vector Machine, and Fuzzy K-Nearest Neighbors. Sets of 5 to 30 selected features were provided to the classifiers. A working point of 15 premium features resulted in accuracy values as high as 90% using the Supprt Vector Machine. The obtained results highlight the potentials of machine learning, applied to VR simulation data, to help realign the traditional apprenticeship educational paradigm to a more objective model, based on proven performance standards.

Keywords Virtual reality simulation · Machine learning · Classifiers · Neurosurgery skill education and assessment · Tumor resection

1 Introduction

Virtual reality (VR) simulators have been proposed as tools to understand, assess, and train neurosurgery residents [1–5]. An important element of simulator performance is the capacity of simulators to distinguish operator expertise. Most studies on operator performance have utilized "metrics." [6–16] Metrics could be defined as standards of reference by which performance, efficiency, and progress can be assessed. Individual metrics can be used to assess aspects of operator performance. Tool acceleration [17], applied forces [9, 18–22], bimanual dexterity [15, 22–24], and effect of stress [24] have all been studied. An operator's performance metrics can be compared with previously defined proficiency benchmarks. The operator could then be placed into 1 of 2 or more groups with specific levels of psychomotor expertise [25, 26]. These articles have applied statistical analysis to various metrics. Statistical analysis has usually been used to determine the quality of individual metrics. Using metrics individually, as practiced so far, to differentiate performance in a unidimensional feature space may not provide adequate distinction. Neurosurgical tasks are complicated, involving multiple cognitive processes and psychomotor skills, and larger sets of metrics or features may be required in combination to differentiate groups in a multidimensional space.

Machine learning algorithms have the capacity to use extensive datasets involving numbers of features to separate groups [27–31]. Machine learning applications have been reviewed in neurosurgery [29]. Supervised learning has been employed in neurosurgical diagnosis, presurgical planning, and outcome prediction. Machine learning has been used to characterize performance during otolaryngology and dental VR procedures [32–37]. In these studies, participants ranged from 1 to 7 skilled (experts) and 5 to 40 novice (less skilled).

✉ Hamed Azarnoush
hamed.azarnoush@mail.mcgill.ca

[1] Department of Biomedical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

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

[3] Science and Research Branch, Islamic Azad University, Tehran, Iran

The focus of this study was on using machine learning to investigate the construct validity of the simulator, which is the ability of the simulator to differentiate between expert and novice performances. We sought to answer the question: to what extent and with what accuracy can the simulators distinguish between experts and novices using classifiers? This could be done with black box classifiers and large number of features that are not necessarily interpretable to humans. Considering the defined scope, we did not choose classifiers based on interpretability. Rather, we chose simple and widely used classifiers and excluded neural network–based classifiers because of data limitation.

To our knowledge, machine learning has not been used in the context of VR simulation of neurosurgery and neurosurgery education. More specifically, it has not been utilized to differentiate "skilled" and "novice" neurosurgical psychomotor performance using a VR simulator. Neurosurgery resident training could be modernized accordingly, once clinical relevance of applying VR simulators together with machine learning has been established. Human behavior is complex, and it could be understood using complex methods. A motivating factor for the current study is that it could pave the way towards application of powerful deep learning and feature learning methods. Upon availability of large datasets, this goal could be realized. Another important issue is that studies like the current study are exploratory studies, i.e., exploring metrics and understanding them, rather than confirmatory studies, i.e., confirming hypotheses. Feature extraction, selection, and machine learning classifiers are helpful tools to achieve these exploratory objectives.

In this study, a virtual reality tumor resection task was used. The task was defined as removing the virtual tumor without removing the surrounding healthy tissue. The participants were divided into 2 groups, namely, the "skilled" group comprising staff neurosurgeons and senior neurosurgery residents, and the "novice" group comprising junior neurosurgery residents and medical students. NeuroVR (CAE Healthcare, Montreal, Canada) was used to simulate the tumor resection scenarios. In simulation scenarios, a simulated ultrasonic aspirator was used as the tool to remove tumors. The aspirator was activated using a foot pedal. The simulator recorded raw data, such as tool tip coordinates, tool tip orientation, haptic forces, and foot pedal status versus time. At the end of the procedure, the simulator also provided the volume of tumor removed as well as the volume of the surrounding healthy tissue removed.

After the trial, the data was postprocessed for all groups to extract features that could relate operation dexterity, e.g., speed, and operation safety, e.g., maximum force applied. These features were used in machine learning algorithms to investigate to what degree the 2 groups could be differentiated by classifiers. Many classifiers could have been chosen. In this research, considering that no reference study existed, the goal was to obtain an estimate on the order of performance measures of various classifiers for such application rather than finding the best classifier. In this preliminary study, we used 4 classifiers as examples, namely, K-Nearest Neighbors, Parzen Window, Support Vector Machine, and Fuzzy K-Nearest Neighbors to distinguish skill levels.

2 Methods

2.1 Subjects

For the VR tumor resection task, 115 individuals including 16 board certified practicing neurosurgeons from 3 institutions and 7 senior residents (PGY 4–6) from one university made up the skilled group ($n = 23$). Eight junior residents (PGY 1–3) and 84 medical students made up the novice group ($n = 92$). No participant had had previous experience with the simulator utilized, and participants signed an approved Research Ethics Board consent.

2.2 NeuroVR

The NeuroTouch, now known as NeuroVR (CAE Healthcare, Montreal, Canada), VR simulation platform was used [5]. Tumor resections were performed using the simulated ultrasonic aspirator held in the dominant hand as shown in Fig. 1a.

2.3 Simulation scenarios

Figure 1 b outlines the 6 scenarios used in this study. Each one of Scenarios 1–3 included 3 tumors with the same visual properties but different tactile properties ("Soft": Young's modulus of 3 kPa, "Medium": Young's modulus of 9 kPa, and "Hard": Young's modulus of 15 kPa). Each one of Scenarios 4–6 included 3 tumors with the same tactile properties but different visual properties (black, glioma-like, and white). Therefore, all tumors in Scenarios 1–3 appeared for a second time in a different order in Scenarios 4–6, and a total of 18 tumors were resected. The background that simulated surrounding healthy tissue had the same tactile property as the soft tumor and the same visual property as the white tumor. In each scenario, the participant was instructed to remove the top tumor first, followed by the left and right tumor. Figure 1 c shows the 3D geometry of the tumors from a side view. A 3-min period was allowed for each tumors removal with a 1-min rest time given between tumor resections to decrease fatigue. The trial involved 54 min of active tumor resection, 71 min in total. To develop procedure familiarity, operators resected a practice scenario but this data was not used. Participants were unaware of study purpose or metrics utilized and were instructed to resect each tumor with minimal removal of the background tissue.

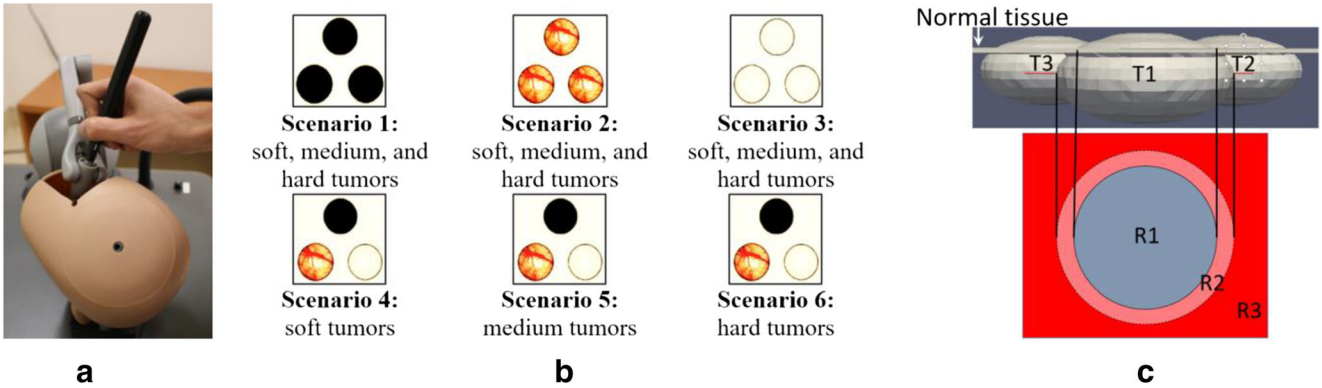

Fig. 1 The hand position of the operator holding the simulated ultrasonic aspirator (**a**), the 6 simulated tumor scenarios with tumor color and sequence (**b**), and lateral view of the brain tumor geometry and ellipsoidal shape utilized in each scenario demonstrating the three identical tumors, tumor buried underneath simulated "normal' tissue and the R1, R2, and R3 regions studied (**c**)

2.4 Process steps

The processing steps, including feature extraction, statistical feature selection, forward feature selection, and classification, are shown in Fig. 2. In feature extraction step, various features were obtained from the raw data, e.g. tool tip coordinates, provided by the simulator. In statistical feature selection step, the features that differentiated the skilled and novice groups with statistical significance were selected. Next, forward feature selection algorithm was used to select premium features to be provided to the classifiers.

2.5 Feature extraction

The simulator recorded signals including tool tip coordinates, tool tip orientation angles, contact force between virtual tool and virtual tissue, and foot pedal state versus time. The list of all signal features is included in Table 1. Based on our investigation, the raw data was clean and was not affected by noise.

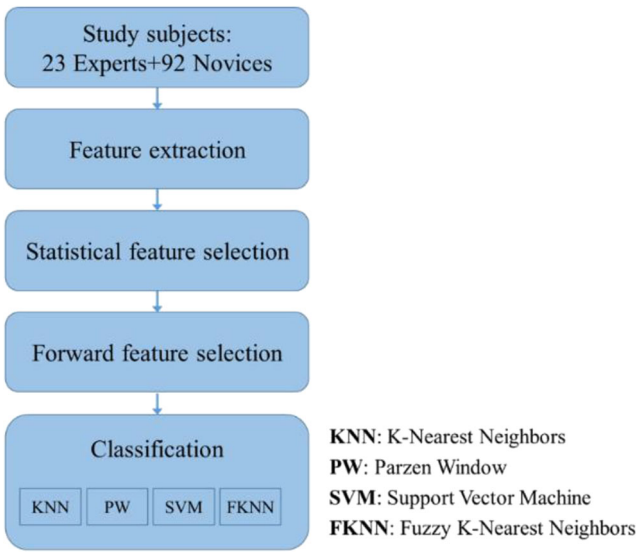

Fig. 2 Process steps

Preprocessing included using B-splines to smooth the signals for differentiation. Different parametric features could be extracted from signal features with the goal to differentiate the skilled and novice groups [38].

2.5.1 Motion-based features

To obtain motion-based parametric features, first, signals such as velocity (first derivative), acceleration (second derivative), and jerk (third derivative) for position and angle signals were obtained. Then, based on signal features, parametric temporal and spatial features were extracted.

Velocity Velocity was computed as the first derivative of motion profile and then speed was considered as the magnitude of the velocity profile. Features based on the speed values included mean speed, maximum speed, number of local maximum in the velocity vector and movement arrest period ratio [39].

Acceleration Acceleration was computed as the second derivative of the motion profile. Features based on the acceleration signal included mean acceleration, maximum acceleration, and the integral of the acceleration vector (IAV) [40], as given by the following:

$$\mathrm{IAV} = \int_0^T \sqrt{\left(\frac{d^2x}{dt^2}\right)^2 + \left(\frac{d^2y}{dt^2}\right)^2 + \left(\frac{d^2z}{dt^2}\right)^2}\, dt \qquad (1)$$

where x, y, and z are Cartesian coordinates and T is the task completion time.

Jerk Jerk is defined as the third derivative of motion profile applied for motor skill assessment. A normalized three dimensional jerk [40] metric is used in this study, given by the

Table 1 List of signal features

$x(t)$:position in the x-direction	$j_z(t) = \frac{da_z(t)}{dt}$: jerk in the z-direction
$y(t)$:position in the y-direction	$j_f(t) = \frac{da_f(t)}{dt}$: third derivative of force signal
$z(t)$:position in the z-direction	roll(t): rotation around the front-to-back axis
$f(t)$: force	$v_{\text{roll}}(t) = \frac{d\text{roll}(t)}{dt}$: first derivative of roll signal
$v_x(t) = \frac{dx(t)}{dt}$: velocity in the x-direction	$a_{\text{roll}}(t) = \frac{dv_{\text{roll}}(t)}{dt}$: second derivative of roll signal
$v_y(t) = \frac{dy(t)}{dt}$: velocity in the y-direction	$j_{\text{roll}}(t) = \frac{da_{\text{roll}}(t)}{dt}$: third derivative of roll signal
$v_z(t) = \frac{dz(t)}{dt}$: velocity in the z-direction	pitch(t): rotation around the side-to-side axis
$v_f(t) = \frac{df(t)}{dt}$: first derivative of the force signal	$v_{\text{pitch}}(t) = \frac{d\text{pitch}(t)}{dt}$: first derivative of pitch signal
$V(t) = \sqrt{\frac{dx}{dt}^2 + \frac{dy}{dt}^2 + \frac{dz}{dt}^2}$: magnitude of velocity	$a_{\text{pitch}}(t) = \frac{dv_{\text{pitch}}(t)}{dt}$: second derivative of pitch signal
$a_x(t) = \frac{dv_x(t)}{dt}$: acceleration in the x-direction	$j_{\text{pitch}}(t) = \frac{da_{\text{pitch}}(t)}{dt}$: third derivative of pitch signal
$a_y(t) = \frac{dv_y(t)}{dt}$: acceleration in the y-direction	yaw(t): rotation around the vertical axis
$a_z(t) = \frac{dv_z(t)}{dt}$: acceleration in the z-direction	$v_{\text{yaw}}(t) = \frac{d\text{yaw}(t)}{dt}$: first derivative of yaw signal
$a_f(t) = \frac{dv_f(t)}{dt}$: second derivative of force signal	$a_{\text{yaw}}(t) = \frac{dv_{\text{yaw}}(t)}{dt}$: second derivative of yaw signal
$j_x(t) = \frac{da_x(t)}{dt}$: jerk in the x-direction	$j_{\text{yaw}}(t) = \frac{da_{\text{yaw}}(t)}{dt}$: third derivative of yaw signal
$j_y(t) = \frac{da_y(t)}{dt}$: jerk in the y-direction	

following:

$$\text{Jerk}_{\text{norm}} = \sqrt{\frac{T^5}{2A_m^2}\int_0^T \left(\frac{d^3x}{dt^3}\right)^2 + \left(\frac{d^3y}{dt^3}\right)^2 + \left(\frac{d^3z}{dt^3}\right)^2 dt} \quad (2)$$

where T is the task completion time and A_m is the amplitude of the motion.

2.5.2 Force-based features

During neurosurgery, tool tip forces on tissue cannot be measured. Not being able to measure forces has limited our understanding of the forces that the human brain is exposed to by this instrument. The VR simulator used had the ability to simulate tool-tissue forces. This data has been utilized to create force pyramids and force heat maps to assess psychomotor function, automaticity, and force fingerprints for VR tumor resections [9, 18, 19, 21]. Force-based features extracted in this study comprise force derivatives, integral of the force, the range of the applied forces, and the interquartile range, i.e., the first quartile subtracted from the third quartile. In addition to the abovementioned force-based features, parametric features including temporal and spatial features were also extracted from the force signal and its derivatives. We also used the 2 features proposed previously to indicate consistency [41], given by following:

$$df_{\text{metric}} = \sqrt{\frac{T}{2f_{iqr}^2}\int_0^T \left(\frac{df}{dt}\right)^2 dt} \quad (3)$$

$$d^2f_{\text{metric}} = \sqrt{\frac{T^3}{2f_{iqr}^2}\int_0^T \left(\frac{d^2f}{dt^2}\right)^2 dt} \quad (4)$$

and one feature to indicate the smoothness of the force application [41], given by the following:

$$d^3f_{\text{metric}} = \sqrt{\frac{T^5}{2f_{\text{iqr}}^2}\int_0^T \left(\frac{d^3f}{dt^3}\right)^2 dt}. \quad (5)$$

where T is the task completion time and f_{iqr} is the interquartile range of the force profile.

We initially extracted 150 parametric features, many of which were eliminated in the subsequent feature selection process.

2.6 Feature normalization

Since the parametric feature values are not in the same order of size for comparison and to train classifiers, the obtained features were normalized. Considering that min-max and z-score normalization have been shown in the literature to be sensitive to outliers [42], we used an exponential normalization using:

$$Z_i = e^{-\frac{x_i}{\max(x)}} \quad (6)$$

where Z_i is the normalized value and x_i is a data point $(x_1, x_2, \ldots, x_n)$.

2.7 Feature selection

Feature selection followed feature extraction to reduce computational complexity while maintaining classifier performance [43]. Irrelevant features were identified and only useful ones were provided to classifiers [44]. Feature selection was carried out in 2 steps; considering that we started with a large number ($n = 150$) of features, in the first step of feature selection, we used statistical t test as a filter-based approach, which was fast with better computational complexity in order to select features that defined statistical differentiation. In the second step, we used forward feature selection as a wrapper-based approach in order to select the premium features that improved classifier performance [45].

2.7.1 Statistical feature selection

For each feature, a t test was applied, following the D'Agostino & Pearson test method to confirm normality of distributions, and the resultant p values were compared for all features as a measure of the usefulness of each individual feature to separate novice and skilled groups. Among the extracted preliminary features, 68 features were able to differentiate the 2 groups with a statistically significant difference were $p < 0.05$ as provided in Table 2.

2.7.2 Forward feature selection

To find the most relevant features, forward feature selection, backward feature selection, and the genetic algorithm were applied. Based on our assessment, the error obtained from forward feature selection was 2–20% smaller than that obtained from backward feature selection and 2–12% smaller than that obtained from the genetic algorithm for different scenarios. Therefore, we continued the process with the results obtained from forward feature selection only. This algorithm starts with an empty set and adds features one by one outlining the best feature set of particular size [44]. This algorithm was applied to rank the best 5, 10, 15, 20, 25, and 30 features from the 68 features previously selected utilizing the t test. In the feature selection, minimum of estimated Mahalanobis distances were used, and feature selection was done independently and irrespective of the subsequent classifiers to be used in the subsequent stage.

2.8 Classification

The goal of this study is to provide an estimate of classifier performance measures for such application. Four classifiers, namely, K-Nearest Neighbors (KNN), Parzen Window (PW), Support Vector Machine (SVM), and Fuzzy K-Nearest Neighbors (FKNN), were applied to classify skilled and

Table 2 List of 68 selected parametric features that provide the best classification. The best 30 features are marked by one asterisk (*) and the best 15 features by two asterisks (**)

1	$\sum t(j_x \le 0)/T$	T: task completion time
2	$\sum N(f > 0.1)$	
3*	$\mathrm{std}(f)/\mathrm{std}(v_x)$	std: standard deviation
4	$(\max(v_x) - \min(v_x)) * (\max(v_y) - \min(v_y)) * (\max(v_z) - \min(v_z))$	
5	$\mathrm{iqr}(f)$	
6	$\sqrt{\frac{T^3}{2(\mathrm{iqr}(f))^2}\sum a_f{}^2}$	
7	$(\sum_{i=1}^{N-1}\lvert f_{i+1} - f_i\rvert)/T$	
8**	$(\sum_{i=1}^{N-1}\lvert v_{i+1} - v_i\rvert)/T(\mathrm{std}(f))$	std: standard deviation
9**	$(\sum t(\max(a_x)))/T$	
10*	$N_{\mathrm{zero-cross}}(v_x)$	
11*	$(\sum t(\min(a_y)))/T$	
12	$(\sum t(\max(a_z)))/T$	
13	$N_{\mathrm{extermum}}(\mathrm{Pitch})$	
14*	$(\sum t(\min(a_f)))/T$	
15	$\mathrm{mean}(V)$	
16*	$\mathrm{std}(f)/\mathrm{std}(v_z)$	std: standard deviation
17**	$\sum f_{\mathrm{low\ frequency}}/\sum f_{\mathrm{high\ frequency}}$	
18	$(\sum t(\max(z)))/T$	
19	$N_{\mathrm{extermum}}(v_f)$	
20**	$N_{\mathrm{minimum}}(a_x)$	
21	$N_{\mathrm{minimum}}(a_y)$	
22	$N_{\max}(x) + N_{\max}(y) + N_{\max}(z)$	
23**	$N_{\mathrm{extermum}}(x)$	
24	$N_{\mathrm{extermum}}(z)$	
25	$(\sum_{i=1}^{N-1}\lvert \mathrm{pitch}_{i+1} - \mathrm{pitch}_i\rvert)/T$	
26	$(\sum_{i=1}^{N-1}\lvert v_{\mathrm{roll}i+1} - v_{\mathrm{roll}i}\rvert)/T$	
27*	$\mathrm{mean}(\mathrm{roll}) * T/(\max(v_{\mathrm{roll}}) - \min(v_{\mathrm{roll}}))$	
28	$N_{\min}(\mathrm{yaw}) + N_{\min}(\mathrm{pitch}) + N_{\min}(\mathrm{roll})$	
29	$N_{\mathrm{extermum}}(\mathrm{pitch})$	
30	$N_{\mathrm{extermum}}(v_{\mathrm{yaw}})$	
31**	$N_{\mathrm{extermum}}(v_{\mathrm{roll}})$	
32*	$\sum t(\max(\mathrm{pitch})) - \sum t(\min(\mathrm{pitch})/T$	
33	Frequency of pedal activation	
34	$\sum_{R3} f$	$R3$ as defined in Fig. 1c
35*	$N_{extermum}(f)$	
36*	$\sum_{R4} f$	$R_4 : R_1 \cup R_2$ as defined in Fig. 1c
37	$\max(f) - \min(f)$	
38	$\mathrm{std}(f)$	std: standard deviation
39	$\mathrm{iqr}(x) * \mathrm{iqr}(y) * \mathrm{iqr}(z)$	
40**	$\sqrt{\frac{T}{2(\mathrm{iqr}(f))^2}\sum v_f{}^2}$	
41	$\sum_i \sqrt{a_{xi}^2 + a_{yi}^2 + a_{zi}^2}$	
42*	$(\sum_{i=1}^{N-1}\lvert a_{i+1} - a_i\rvert)/T$	
43	$\int_{t_1}^{t_2}\lvert f\rvert$	t_1 : start point of signal peak t_2 : end point of signal peak
44*	$(\sum t(\min(a_x)))/T$	
45**	$(\sum t(\max(a_y)))/T$	
46	$N_{\mathrm{zero-cross}}(v_y)$	
47**	$(\sum t(\min(a_z)))/T$	
48*	$(\sum t(\max(a_f)))/T$	

Table 2 (continued)

49	$\max(V)$
50**	$\sqrt{\frac{T^5}{2(\mathrm{iqr}(f))^2}\sum {j_f}^2}$
51**	std(f)/std(v_y) std: standard deviation
52*	$N_{\mathrm{minimum}}(\mathrm{x})$
53**	$N_{\mathrm{minimum}}v_x)$
54**	$(\sum t(\max(f)))/\sum t(min(f)))$
55	$N_{\mathrm{extermum}}(a_f)$
56	$\sum t(v_f \geq 0)/\sum t(v_f) \leq 0$
57	$N_{\min}(x) + N_{\min}(y) + N_{\min}(z)$
58	$N_{\mathrm{extermum}}(\mathrm{y})$
59*	$\sum_{i=1}^{N-1} \lvert \mathrm{yaw}_{i+1} - \mathrm{yaw}_i \rvert$
60	$(\sum_{i=1}^{N-1} \lvert v_{\mathrm{pitch}_{i+1}} - v_{\mathrm{pitch}_i} \rvert)/T$
61	$\mathrm{mean}(\mathrm{pitch}) * T/\max(v_{\mathrm{pitch}}) - \min(v_{\mathrm{pitch}})$
62**	$N_{\max}(\mathrm{yaw}) + N_{\max}(\mathrm{pitch}) + N_{\max}(\mathrm{roll})$
63	$N_{\mathrm{extermum}}(\mathrm{yaw})$
64**	$N_{\mathrm{extermum}}(\mathrm{roll})$
65*	$N_{\mathrm{extermum}}(v_{\mathrm{pitch}})$
66	$\sum_{i=1}^{N-1} \lvert j_{\mathrm{pitch}_{i+1}} - j_{\mathrm{pitch}_i} \rvert / \mathrm{mean} \lvert j_{\mathrm{pitch}} \rvert$
67	$\sum t(\max(\mathrm{yaw})) - \sum t(\min(\mathrm{yaw})/T$
68	$\sum_{R1} f \quad R1$ as defined in Fig. 1c

novice groups [46, 47]. The SVM was used with a linear kernel. These classifiers were chosen as examples because they are simple and effective with proven performance. Classifiers based on neural networks were not used since they require large datasets. Classifiers were implemented in MATLAB.

A nested cross-validation procedure was used to tune the hyperparameters. An outer loop was run for 10 iterations, each time randomly selecting the outer train set and the outer test set. The splitting was done across subjects. In each iteration, k-fold cross validation with $k = 10$ was used within the outer train set, i.e., dividing this set to 10 subsample sets, 9 inner train sets and 1 inner test set. The hyperparameter was tuned in this inner 10-iteration loop providing the best result for the inner test set. The tuned values were used on the isolated outer test set yielding the accuracy values, reported as the results of this article.

For the KNN and FKNN classifiers, the range 1–10, with increments of 1, was investigated as the value of K and $K = 7$ provided the best results. For the SVM, the range 1–10, with increments of 1, was investigated as the value of box constraint c, and $c = 3$ provided the best results.

3 Results

3.1 Influence of the number of premium features

Various splits were investigated, namely, 50%-50%, 60%-40%, and 70%-30%, for the outer train and outer test sets, respectively, and 60%-40% split was chosen. Performance of classifiers was assessed based on different numbers of selected premium features from 5 to 30. Figure 3 demonstrates the obtained values of accuracy based on number of premium features. SVM demonstrated the best overall performance. The results indicate that overall classifier performance was improved when the number of premium features was increased to 15. For higher feature numbers, accuracy either decreased or did not significantly increase.

3.2 Performance at selected working point

Figure 4 provides a comparison of classifier performance for all scenarios for the working point of 15 best features and outlines that SVM classifier has the best overall performance with accuracy values ranging from 86 to 90%. On average, classifier accuracy values range between 83 and 85%. In Table 2, the best 30 features are marked with one asterisk (*) and the best 15 features with two asterisks (**).

4 Discussion

4.1 Differentiating skilled and novice performance

The scenarios utilized in this study involved aspirator skills used in human tumor resections, part of the surgical armamentarium of neurosurgeons and senior residents, but not yet acquired by all junior residents and medical students. It seemed reasonable to define a skilled and novice group based on the required skill set [9, 26].

We applied 4 different classifiers to the dataset involving these participants. Our results demonstrate that the 4 classifiers achieved a comparable performance, which was expected, but since this was a preliminary study, we included several classifiers in our analysis. The obtained accuracy values were as high as 90%, obtained by SVM, indicating the potentials of classifiers in differentiating participants doing VR procedures. We see in Fig. 4 that classifier accuracy values are within an 11% range, which signifies good agreement between them. In addition, based on Fig. 4, average classifier accuracy values range between 83 and 85%, meaning that classifier performance is not sensitive to the scenario considered.

Derivative and force features contain information about how smoothly an individual operates. Table 3 presents the number and index (referring to Table 2) of position, force, and derivative features that were selected among the best 30 and best 15 features. Signals recorded from the simulator included x, y, z, roll, pitch, yaw, and force. We can see that while force was 1 out of 7 signals recorded, the force features constitute 12 out of 30 best features and 6 out of 15 best features, which underlines the importance of force features. Referring to Fig. 4, the accuracy of classifiers is within an 11% range for different scenarios with the same set of features (minimum for Scenario 1 by KNN and maximum for Scenario 4 by SVM). In future studies, scenarios could be defined targeted towards

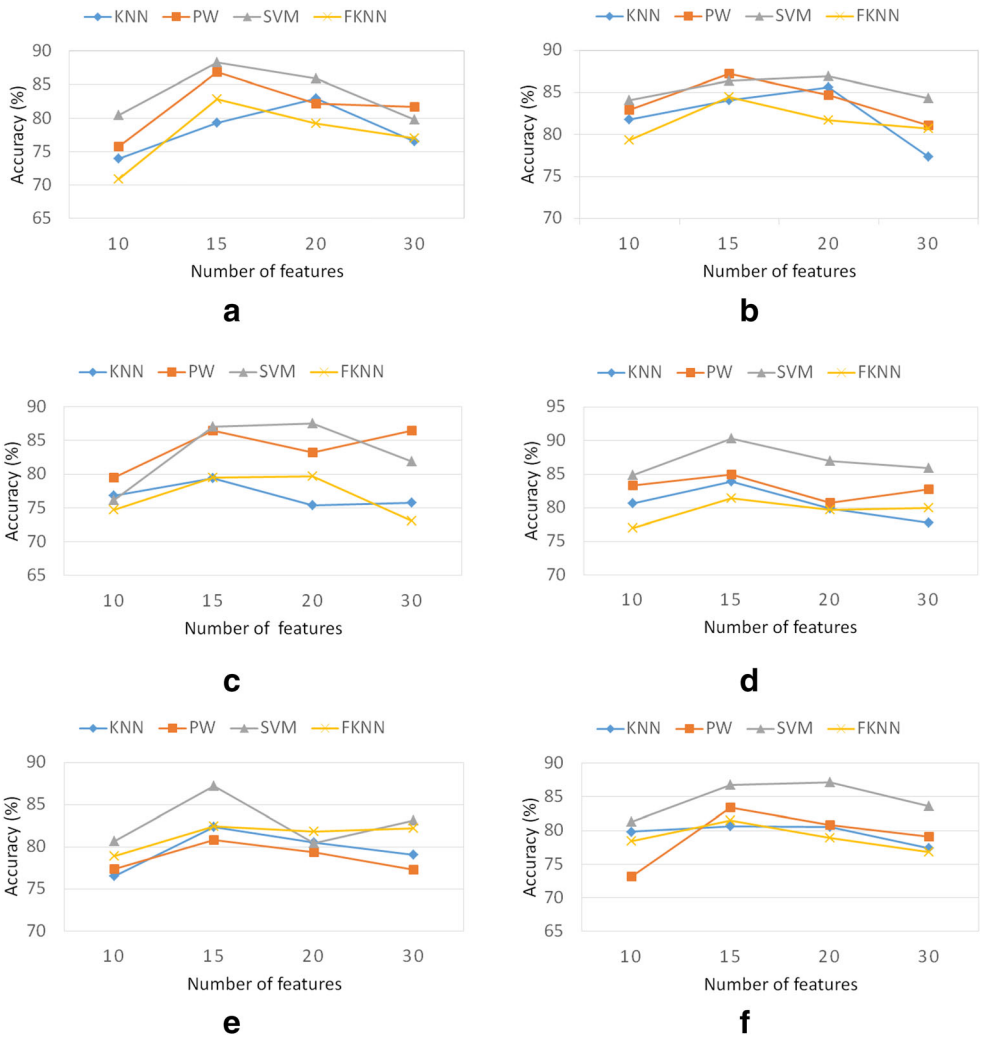

Fig. 3 Classification results for different number of selected premium features and with a 60%–40% train to test set size ratio, Scenario 1 (**a**), Scenario 2 (**b**), Scenario 3 (**c**), Scenario 4 (**d**), Scenario 5 (**e**), Scenario 6 (**f**) with resultant accuracy values (%) for each scenario and each classifier employed: K-Nearest Neighbors (KNN), Parzen Window (PW), Support Vector Machine (SVM), and Fuzzy K-Nearest Neighbors (FKNN)

investigating performance obtained for various sets of features.

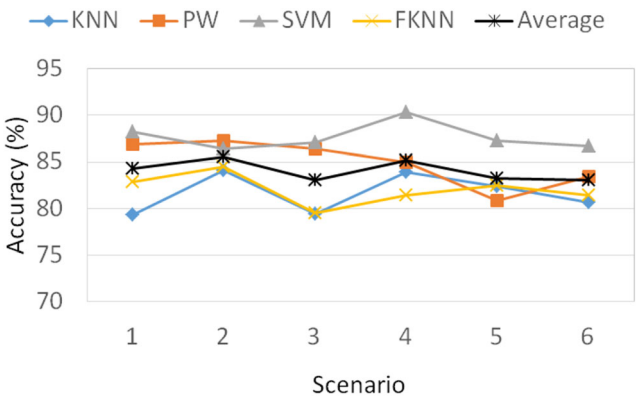

Fig. 4 Classification results for selected working point (15 best features and 60% train set size) with resultant average accuracy (%) for each scenario and each classifier employed: K-Nearest Neighbors (KNN), Parzen Window (PW), Support Vector Machine (SVM), and Fuzzy K-Nearest Neighbors (FKNN)

4.2 Misclassification

Table 4 presents the range of individuals misclassified by the SVM. Using this classifier, 7–9 out of 10 skilled individuals and 31–34 out of 36 novices in the test set were correctly classified. Some neurosurgeons in this study had cerebrovascular, spinal, and functional specialization with little exposure to tumor resection, which may be one reason for misclassification. Some junior residents may have been misclassified since they had obtained the required surgical skills. Studies involving more complex scenarios, larger resident numbers, and better understanding of which factors and/or combination of metrics to use to better differentiate groups are needed. The potential of machine learning classifiers applied to VR procedures in surgical disciplines is that the new features identified will result in new "metrics," which can then be evaluated in other systems. These results may not only help us understand the psychomotor skills needed to increase surgical skills but also aid in resident assessment and training and improve patient outcomes.

Table 3 Breakdown of selected premium features by number and index (referring to Table Table 2)

	Position	Force	Derivatives		
			Position	Force	Position and force
Best 30 features (number and indices)	6 features (23, 32, 52, 59, 62, 64)	4 features (17, 35, 36, 54)	12 features (9, 10, 11, 20, 27, 31, 42, 44, 45, 47, 53, 65)	4 features (14, 40, 48, 50)	4 features (3, 8, 16, 51)
Best 15 features (number and indices)	3 features (23, 62, 64)	2 features (17, 54)	6 features (9, 20, 31, 45, 47, 53)	2 features (40, 50)	2 features (8, 51)

4.3 Agreement between classifiers

Cohen's Kappa measure was used to investigate the agreement between pairs of classifiers. Table 5 defines the level of agreement for different ranges of Kappa.

Table 6 includes the obtained Kappa values for different scenarios. As could be seen the Kappa values, overall lie in the substantial agreement range. The agreement for FKNN-KNN pair is higher than that of other pairs for each scenario. The reason could be the similarity of the method used by these two classifiers.

4.4 Strengths and limitations of the study

The importance of these results lies in their potential educational application to aid in neurosurgical resident training and further defining the psychomotor skill sets of expert surgeons [7]. Machine learning and artificial intelligence as applied to VR surgical studies should be seen as useful adjuncts and not a replacement for standard residency training. By relying on 68 features, these classifiers can automatically capture multiple aspects of psychomotor performance and segregate participants into "skilled" or "novice" group. However, this should be seen as an initial step of a formative educational process, prompting instructors to further assess and coach a resident's performance to a desired level.

A number of strengths and limitations could be highlighted related to subject recruitment, features, classifiers, and the simulator platform used in this study:

1) Comparing similar previous studies [32–37], having a large number of surgeons participating in this study could be considered a strength. However, this is generally a limitation and many institutes may not have access to many surgeons. Defining large populations of residents and neurosurgeons with equivalent experience in surgery and in VR simulation is challenging. Sixteen practicing board-certified neurosurgeons from 3 institutions with different areas of expertise participated in this study, which is felt to be representative of a general neurosurgical population. We only enrolled residents and medical students from one institution that limits extension of these results. The authors believe that increasing study participants from multiple institutions may further our ability to improve classifier performance to distinguish neurosurgical skill level at various stages of resident training. In our study, we were not able to recruit as many individuals in the skilled group in comparison with the novice group.
2) In the statistical and forward feature selection stages, features were selected irrespective of classifiers, which could be considered a strength. On the other hand, many of the features included in this investigation have not been assessed in more complex scenarios. Therefore, it is not known if the same features would also be applicable to those scenarios. Whether these features are the most appropriate or others, such as the force pyramid or automaticity, would be more useful needs to be assessed [18–21].
3) In many studies, the performance of only 1 classifier, e.g., SVM, has been analyzed. In this study, we analyzed the performance of 4 example classifiers. However, this analysis is not exhaustive. For example, classifiers based on neural networks have been excluded. Availability of larger datasets could make performance analysis of such classifiers possible. In this study, we focused on time domain

Table 4 The range of numbers of individuals correctly and incorrectly classified by the SVM ($c = 3$, and 15 features selected) in the 6 different scenarios

	Classified as skilled	Classified as novice
Skilled, $N = 10$	7–9	1–3
Novice, $N = 36$	2–5	31–34
$N = 46$		

Table 5 Classifier agreement based on Kappa values

Kappa < 0: No agreement
Kappa between 0.00 and 0.20: Slight agreement
Kappa between 0.21 and 0.40: Fair agreement
Kappa between 0.41 and 0.60: Moderate agreement
Kappa between 0.61 and 0.80: Substantial agreement
Kappa between 0.81 and 1.00: Almost perfect agreement

Table 6 Kappa values for different classifier pairs

	Classifier pair	Kappa
Black	SVM-KNN	0.6168 ± 0.12
	SVM-FKNN	0.6431 ± 0.09
	SVM-PW	0.6761 ± 0.13
	PW-KNN	0.6185 ± 0.17
	PW-FKNN	0.6153 ± 0.12
	FKNN-KNN	0.7145 ± 0.13
Glioma	SVM-KNN	0.5913 ± 0.16
	SVM-FKNN	0.5797 ± 0.15
	SVM-PW	0.6141 ± 0.12
	PW-KNN	0.5830 ± 0.12
	PW-FKNN	0.6572 ± 0.22
	FKNN-KNN	0.7790 ± 0.09
White	SVM-KNN	0.6290 ± 0.10
	SVM-FKNN	0.6336 ± 0.26
	SVM-PW	0.6408 ± 0.17
	PW-KNN	0.6769 ± 0.09
	PW-FKNN	0.7280 ± 0.13
	FKNN-KNN	0.7149 ± 0.10
Soft	SVM-KNN	0.5177 ± 0.08
	SVM-FKNN	0.5778 ± 0.10
	SVM-PW	0.6155 ± 0.05
	PW-KNN	0.5183 ± 0.08
	PW-FKNN	0.5998 ± 0.11
	FKNN-KNN	0.7901 ± 0.19
Medium	SVM-KNN	0.6022 ± 0.10
	SVM-FKNN	0.5781 ± 0.07
	SVM-PW	0.5807 ± 0.11
	PW-KNN	0.6385 ± 0.22
	PW-FKNN	0.6600 ± 0.16
	FKNN-KNN	0.7737 ± 0.22
Hard	SVM-KNN	0.6158 ± 0.13
	SVM-FKNN	0.6613 ± 0.11
	SVM-PW	0.6090 ± 0.18
	PW-KNN	0.5939 ± 0.16
	PW-FKNN	0.6104 ± 0.09
	FKNN-KNN	0.7597 ± 0.13

features. While, some of these features, such as those based on derivatives, implicitly contain information about frequency of the signal, the frequency domain features could be investigated further in future studies.

4) A simulated aspirator was utilized in the dominant hand, which is not representative of the bimanual psychomotor skills and multiple instruments employed during real tumor resections. Previous studies have demonstrated differences in ergonomics between right- and left-handed operators, and this issue was not addressed in this investigation and deserves further study [18].
5) The different visual and haptic complexities of simulated tumors utilized and task duration may not adequately discriminate operator performance. More complex and realistic tumor scenarios with simulated bleeding involving use of bimanual instruments are being studied using classifiers, which may be more useful.
6) Virtual reality simulation of neurosurgery is a developing field still in its infancy. The focus of the current exploratory study was on the investigation of the separability of expert and novice performance in a VR simulator using machine learning, which has not been reported earlier. Furthermore, such studies could provide an analysis of what features could be more important in the context of expert performance as identified by asterisks in Table 2 and summarized in Table 3. The analyses generated by such exploratory studies could provide hypotheses, which could be proven or disproven in the future by more focused confirmatory studies. We believe more machine learning exploratory studies are needed on other scenarios that may involve other factors such as bimanual dexterity, bleeding, and more complex tumor geometries.

Improving the choice of features and classifiers goes hand in hand with improving the simulator design and performance. Studies such as the current study could shed light on useful features and classifiers and could guide simulator development in the right direction. From performance assessment point of view, the classifiers, even when they are not interpretable, could be useful in distinguishing skill level. From training point of view, a future stage of this research could be to make better sense of models and features obtained. This understanding would be important in teaching residents the required surgical skills and establishing a formative assessment strategy. From interpretability point of view, future machine learning research could adopt different paths, e.g., choosing interpretable models or interpreting black box models [48]. Improving the simulators on the other hand could make scenarios more realistic for neurosurgeons and creating a better platform for them to show their skills, and as a result, classifier performance would be improved.

5 Conclusion

The goal of this article is to highlight the potentials of machine learning in VR simulators in the context of neurosurgical resident training and helping further define the psychomotor skill set of the neurosurgeons. This study could construct a platform for more advanced machine learning algorithms, e.g., based on deep learning, and justify the necessity of obtaining

larger datasets from a large number of institutes. In the longer term, this may help realign the present apprenticeship educational paradigm with a more objective model, based on proven performance standards.

Acknowledgments We thank all the neurosurgeons, residents, and medical students from the Montreal Neurological Institute and Hospital and other institutions who participated in this study. We would also like to thank Robert DiRaddo, Group Leader, Simulation, Life Sciences Division, National Research Council Canada at Boucherville and his team, including Denis Laroche, Valérie Pazos, Nusrat Choudhury, and Linda Pecora, for their support in the development of the scenarios used in these studies and all the members of the Simulation, Life Sciences Division, National Research Council Canada.

Funding information This work was supported by the Di Giovanni Foundation, the Montreal English School Board, the B-Strong Foundation, the Colannini Foundation, the Montreal Neurological Institute and Hospital, and the McGill Department of Orthopedics.

Compliance with ethical standards

Disclaimer Samaneh Siyar is a Visiting Scholar in the Neurosurgical Simulation Research and Training Centre. Dr. H. Azarnoush previously held the Postdoctoral Neuro-Oncology Fellowship from the Montreal Neurological Institute and Hospital and is a Visiting Professor in the Neurosurgical Simulation Research and Training Centre. Dr. Winkler-Schwartz holds a Robert Maudsley Fellowship from the Royal College of Physicians and Surgeons of Canada and Nirros Ponnudurai is supported by a Heffez Family Bursary. Dr. Del Maestro is the William Feindel Emeritus Professor in Neuro-Oncology at McGill University.

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval Research Ethics Board at McGill University approved this study and Informed consent was obtained for experimentation with human subjects.

References

1. Kockro RA, Serra L, Tseng-Tsai Y, Chan C, Yih-Yian S, Gim-Guan C et al (2000) Planning and simulation of neurosurgery in a virtual reality environment. Neurosurgery. 46(1):118–137
2. Bernardo A, Preul MC, Zabramski JM, Spetzler RF (2003) A three-dimensional interactive virtual dissection model to simulate transpetrous surgical avenues. Neurosurgery. 52(3):499–505 **discussion 504–505**
3. Radetzky A, Rudolph M (2001) Simulating tumour removal in neurosurgery. Int J Med Inform 64(2–3):461–472
4. Lemole GM Jr, Banerjee PP, Luciano C, Neckrysh S, Charbel FT (2007) Virtual reality in neurosurgical education: part-task ventriculostomy simulation with dynamic visual and haptic feedback. Neurosurgery. 61(1):142–149
5. Delorme S, Laroche D, DiRaddo R, Del Maestro RF (2012) NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training. *Neurosurgery* 71(suppl_1):ons32–ons42
6. Choudhury N, Gelinas-Phaneuf N, Delorme S, Del Maestro R (2013) Fundamentals of neurosurgery: virtual reality tasks for training and evaluation of technical skills. World Neurosurg 80(5):e9–e19
7. Gelinas-Phaneuf N, Del Maestro RF (2013) Surgical expertise in neurosurgery: integrating theory into practice. *Neurosurgery* 73(suppl_1):S30–S38
8. Gelinas-Phaneuf N, Choudhury N, Al-Habib AR, Cabral A, Nadeau E, Mora V et al (2014) Assessing performance in brain tumor resection using a novel virtual reality simulator. Int J Comput Assist Radiol Surg 9(1):1–9
9. Azarnoush H, Alzhrani G, Winkler-Schwartz A, Alotaibi F, Gelinas-Phaneuf N, Pazos V, Choudhury N, Fares J, DiRaddo R, del Maestro R (2015) Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. Int J Comput Assist Radiol Surg 10(5):603–618
10. Cline BC, Badejo AO, Rivest II, Scanlon JR, Taylor WC, Gerling GJ (2008) Human performance metrics for a virtual reality simulator to train chest tube insertion. IEEE SIEDS :168–173
11. Kazemi H, Rappel JK, Poston T, Hai Lim B, Burdet E, Leong TC (2010) Assessing suturing techniques using a virtual reality surgical simulator. Microsurgery. 30(6):479–486
12. Trejos AL, Patel RV, Malthaner RA, Schlachta CM (2014) Development of force-based metrics for skills assessment in minimally invasive surgery. Surg Endosc 28(7):2106–2119
13. Kovac ERA, Azhar A, Quirouet J, Delisle, Anidjar M (2012) Construct validity of the lapSim virtual reality laparoscopic simulator within a urology residency program. CUAJ 6(4):253
14. Alotaibi FE, Al Zhrani G, Bajunaid K, Winkler-Schwartz A, Azarnoush H et al (2015) Assessing neurosurgical psychomotor performance: role of virtual reality simulators, current and future potential. SOJ Neurol 2(1):1–7
15. Alotaibi FE, AlZhrani GA, Mullah MA, Sabbagh AJ, Azarnoush H, Winkler-Schwartz A et al (2015) Assessing bimanual performance in brain tumor resection with NeuroTouch, a virtual reality simulator. Oper Neurosurg 11(1):89–98
16. Alotaibi FE, AlZhrani GA, Sabbagh AJ, Azarnoush H, Winkler-Schwartz A, Del Maestro RF (2015) Neurosurgical assessment of metrics including judgment and dexterity using the virtual reality simulator NeuroTouch (NAJD Metrics). Surg Innov 22(6):636–642
17. Jensen Ang WJ, Hopkins ME, Partridge R, Hennessey I, Brennan PM, Fouyas I, Hughes MA (2013) Validating the use of smartphone-based accelerometers for performance assessment in a simulated neurosurgical task. Oper Neurosurg 10(1):57–65
18. Azarnoush H, Siar S, Sawaya R, Zhrani GA, Winkler-Schwartz A, Alotaibi FE, Bugdadi A, Bajunaid K, Marwa I, Sabbagh AJ, del Maestro R (2017) The force pyramid: a spatial analysis of force application during virtual reality brain tumor resection. J Neurosurg 127(1):171–181
19. Sawaya R, Bugdadi A, Azarnoush H, Winkler-Schwartz A, Alotaibi FE, Bajunaid K, AlZhrani GA, Alsideiri G, Sabbagh AJ, Del Maestro RF (2017) Virtual reality tumor resection: the force pyramid approach. Operative Neurosurgery. 14(6):686–696
20. Bugdadi A, Sawaya R, Olwi D, AlZahrani G, Azarnoush H, Sabbagh A et al (2018) Automaticity of force application during simulated brain tumor resection: testing the Fitts and Posner model. J Surg Educ 75(1):104–115
21. Sawaya R, Alsidieri G, Bugdadi A, Winkler-Schwartz A, Azarnoush A, Bajunaid K, AJ JS, Del Maestro R (2018) Development of a performance model for virtual reality tumor resections. J Neurosurg 1(aop):1–9
22. Winkler-Schwartz A, Bajunaid K, Mullah MA, Marwa I, Alotaibi FE, Fares J et al (2016) Bimanual psychomotor performance in

neurosurgical resident applicants assessed using NeuroTouch, a virtual reality simulator. J Surg Educ 73(6):942–953
23. Holloway T, Lorsch Z, Chary M, Sobotka S, Moore MM, Costa AB, del Maestro R, Bederson J (2015) Operator experience determines performance in a simulated computer-based brain tumor resection task. Int J Comput Assist Radiol Surg 10(11):1853–1862
24. Bajunaid K, Mullah MA, Winkler-Schwartz A, Alotaibi FE, Fares J, Baggiani M et al (2017) Impact of acute stress on psychomotor bimanual performance during a simulated tumor resection task. J Neurosurg 126(1):71–80
25. Alzhrani G, Del Maestro RF (2014) A validation study of NeuroTouch in neurosurgical training. LAP LAMBERT Academic Publishing, Saarbrücken
26. Alzhrani G, Alotaibi F, Azarnoush H, Winkler-Schwartz A, Sabbagh A, Bajunaid K et al (2015) Proficiency performance benchmarks for removal of simulated brain tumors using a virtual reality simulator NeuroTouch. Journal of Surgical Education 72(4): 685–696
27. Samuel AL (1988) Some studies in machine learning using the game of checkers. In: Computer games I. Springer, New York, pp 366–400
28. Obermeyer Z, Emanuel EJ (2016) Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 375(13): 1216
29. Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML et al (2017) Natural and artificial intelligence in neurosurgery: a systematic review. Neurosurgery 83(2):181–192
30. Azimi P, Mohammadi HR, Benzel EC, Shahzadi S, Azhari S, Montazeri A (2015) Artificial neural networks in neurosurgery. J Neurol Neurosurg Psychiatry 86(3):251–256
31. Watson RA (2014) Use of a machine learning algorithm to classify expertise: analysis of hand motion patterns during a simulated surgical task. Acad Med 89(8):1163–1167
32. Rhienmora P, Haddawy P, Khanal P, Suebnukarn S, Dailey MN (2010) A virtual reality simulator for teaching and evaluating dental procedures. Methods Inf Med 49(04):396–405
33. Kerwin T, Wiet G, Stredney D, Shen HW (2012) Automatic scoring of virtual mastoidectomies using expert examples. Int J Comput Assist Radiol Surg 7(1):1–11
34. Ma X, Wijewickrema S, Zhou S, Zhou Y, Mhammedi Z, O'Leary S, et al. Adversarial generation of real-time feedback with neural networks for simulation-based training. arXiv preprint:1703.01460. 2017 Mar 4
35. Ma X, Wijewickrema S, Zhou Y, Zhou S, O'Leary S, Bailey J (2017) Providing effective real-time feedback in simulation-based surgical training. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, pp 566–574
36. Wijewickrema S, Ma X, Piromchai P, Piromchai P, Briggs R, James BJ et al (2018) Providing automated real-time technical feedback for virtual reality based surgical training: is the simpler the better? In: *International Conference on Artificial Intelligence in Education*. Springer, pp 584–598, Cham
37. Sewell C, Morris D, Blevins NH, Dutta S, Agrawal S, Federico Barbagli F et al (2008) Providing metrics and performance feedback in a surgical simulator. Comput Aided Surg 13(2):63–81
38. Rashidi S, Fallah A, Towhidkhah F (2013) Authentication based on pole-zero models of signature velocity. J Medical Signals Sens 3(4): 195–208
39. Rohrer B, Fasoli S, Krebs H, Hughes R, Volpe B, Frontera W, Stein J, Hogan N (2002 Sep 15) Movement smoothness changes during stroke recovery. J Neurosci 22(18):8297–8304
40. Cavallo F, Megali G, Sinigaglia S, Tonet O, Dario P (2006) A biomedical analysis of a surgeon's gesture in a laparoscopic virtual scenario. Stud Health Technol Inf 119:79–84
41. Al T, Patel RV, Naish MD, Malthaner RA, Schlachta CM (2013) The application of force sensing to skills assessment in minimally invasive surgery. In: *2013 IEEE international conference on robotics and automation*, pp 4370–4375
42. Jain A, Nandakumar K, Ross A (2005) Score normalization in multimodal biometric systems. Pattern Recogn 38(12):2270–2285
43. Deng K (1998) Omega: on-line memory-based general purpose system classifier. PhD diss. Carnegie Mellon University, Pittsburgh
44. Ladha L, Deepa T (2011) Feature selection methods and algorithms. IJCSE 3(5):1787–1797
45. Kumari B, Swarnkar T (2011) Filter versus wrapper feature subset selection in large dimensionality micro array: a review. Int J Comput 2(3):1048–1041
46. Kung SY (2014) Kernel Methods and Machine Learning. Cambridge University Press, Cambridge, p 34
47. Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbor algorithm. IEEE T SYST MAN CYB (4):580–585
48. Lipton, Z.C., 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*

Samaneh Siyar received her MSc from Amirkabir University of Technology and is collaborating with McGill University. Her research interests include using machine learning to assess skill in VR simulators.

Hamed Azarnoush completed his doctoral and postdoctoral studies at McGill University. His research interests include virtual reality and computer vision. He is now an assistant professor at Amirkabir University of Technology.

Saeid Rashidi is an assistant professor at Science and Research Branch, Islamic Azad University. His research interests include biomedical signal processing, biometrics, motor control, modeling, and chaos.

Dr. Alexander Winkler-Schwartz is a neurosurgery resident at McGill. He holds the Robert Maudsley Fellowship for his work on simulation and AI to understand surgical expertise.

Vincent Bissonnette graduated medical school from the University of Sherbrooke in 2018. He is currently completing a Master of Science at McGill University in the Department of Experimental Surgery.

Nirros Ponnudurai has a bachelor's degree in mechanical engineering with a specialization in mechatronics. He is currently completing a joint program in Medicine and Management at McGill University.

Dr. Rolando Del Maestro is the William Feindel Professor Emeritus in Neuro-Oncology and the Director of Neurosurgical Simulation and Artificial Intelligence Learning Centre. Research focus: Educational roles of surgical simulation.