Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Utilizing artificial intelligence and electroencephalography to assess expertise on a simulated neurosurgical task

Sharif Natheir[a,*], Sommer Christie[a], Recai Yilmaz[a], Alexander Winkler-Schwartz[a], Khalid Bajunaid[b], Abdulrahman J. Sabbagh[c,d], Penny Werthner[e], Jawad Fares[f], Hamed Azarnoush[g], Rolando Del Maestro[a]

[a] Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology & Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, Quebec, Canada
[b] Department of Surgery, College of Medicine, University of Jeddah, Jeddah, Saudi Arabia
[c] Division of Neurosurgery, Department of Surgery, College of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia
[d] Clinical Skills and Simulation Center, King Abdulaziz University, Jeddah, Saudi Arabia
[e] University of Calgary, Faculty of Kinesiology, Calgary, Alberta, Canada
[f] Department of Neurological Surgery Feinberg School of Medicine, Northwestern University Chicago, Illinois, USA
[g] Department of Biomedical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

ARTICLE INFO

Keywords:
Electroencephalography
Artificial intelligence
Machine learning
Education
Virtual reality
Neurofeedback

ABSTRACT

Virtual reality surgical simulators have facilitated surgical education by providing a safe training environment. Electroencephalography (EEG) has been employed to assess neuroelectric activity during surgical performance. Machine learning (ML) has been applied to analyze EEG data split into frequency bands. Although EEG is widely used in fields requiring expert performance, it has yet been used to classify surgical expertise. Thus, the goals of this study were to (a) develop an ML model to accurately differentiate skilled and less-skilled performance using EEG data recorded during a simulated surgery, (b) explore the relative importance of each EEG bandwidth to expertise, and (c) analyze differences in EEG band powers between skilled and less-skilled individuals. We hypothesized that EEG recordings during a virtual reality surgery task would accurately predict the expertise level of the participant. Twenty-one participants performed three simulated brain tumor resection procedures on the NeuroVR™ platform (CAE Healthcare, Montreal, Canada) while EEG data was recorded. Participants were divided into 2 groups. The skilled group was composed of five neurosurgeons and five senior neurosurgical residents (PGY4-6), and the less-skilled group was composed of six junior residents (PGY1-3) and five medical students. A total of 13 metrics from EEG frequency bands and ratios (e.g., alpha, theta/beta ratio) were generated. Seven ML model types were trained using EEG activity to differentiate between skilled and less-skilled groups. The artificial neural network achieved the highest testing accuracy of 100% (AUROC = 1.0). Model interpretation via Shapley analysis identified low alpha (8–10 Hz) as the most important metric for classifying expertise. Skilled surgeons displayed higher (p = 0.044) low-alpha than the less-skilled group. Furthermore, skilled surgeons displayed significantly lower TBR (p = 0.048) and significantly higher beta (13–30 Hz, p = 0.049), beta 1 (15–18 Hz, p = 0.014), and beta 2 (19–22 Hz, p = 0.015), thus establishing these metrics as important markers of expertise.

ACGME Core Competencies: Practice-Based Learning and Improvement.

1. Introduction

The subpial resection of human brain tumors adjacent to important cortical structures is a challenging operative procedure and one in which neurosurgical trainees are expected to acquire proficiency [1]. Technical errors in this complex bimanual psychomotor skill include subpial vessel hemorrhage and injury to adjacent normal cortex can result in significant patient morbidity [1,2]. To aid learners in the mastery of this technical skill necessary to safely and efficiently carry out these procedures our group has helped develop [3] and validate virtual reality

[*] Corresponding author.
E-mail address: sharif.natheir@mail.mcgill.ca (S. Natheir).

https://doi.org/10.1016/j.compbiomed.2022.106286
Received 10 December 2021; Received in revised form 18 October 2022; Accepted 25 October 2022
Available online 15 November 2022

simulators [4] along with creating complex and realistic virtual reality tumor resection tasks [5]. Virtual reality surgical simulators employed in neurosurgical education provide a safe training environment and allow for self-guided learning [6]. These learning tools are particularly relevant during times when trainees have less clinical interaction such as during the present COVID-19 pandemic, and especially when combined with a mechanism for performance assessment [7,8].

Electroencephalography (EEG), the use of electrodes to assess neural electrical activity, has been used to continuously assess brain activity during surgical performance [9]. EEG data analysis is conducted by transforming the raw data into a variety of frequency bands (e.g., alpha, theta) that are associated with various cognitive processes such as attention, memory, learning and psychomotor efficiency [10,11]. Theta frequencies, for example, are associated with learning and memory, whereas alpha frequencies are associated with tranquillity [12]. The understanding of how each frequency band contributes to surgical expertise, may allow the development and implementation to neurofeedback training interventions to improve technical skills performance [13].

Large EEG data sets can be analyzed by artificial intelligence to deconstruct the frequency bands important in skilled bimanual performance [14]. Artificial intelligence is the use of computers to mimic human decisions. Machine learning is one branch of artificial intelligence that imitates human behavior without the need for a predefined list of rules to follow. Several machine learning algorithms can be trained to discover patterns within a training dataset and their pattern recognition abilities are tested on a separate testing dataset [15].

There are many different types of machine learning algorithms, which are based on different mathematical analytical methods of the data [16]. Some of the most utilized machine learning algorithms include support vector machines, neural networks, logistic regression, linear discriminant analysis, Random Forest, Naïve Bayes, and K-Nearest Neighbors [17]. Machine learning has been applied in neurosurgical care, to assist in the surgical treatment of epilepsy, brain tumors, Parkinson's disease, and brain injury [18]. These algorithms are beginning to play roles throughout the whole arc of neurosurgical care: from presurgical planning to intraoperative guidance, neurological monitoring, and outcome prediction [18]. Our group has employed a number of machine learning algorithms to assess and train surgical learners [19–22]. Machine learning algorithms can be utilized to classify groups into different levels of surgical expertise with greater granularity and precision than previously demonstrated [21,22].

Machine learning models have traditionally been considered black boxes and deciphering their decision-making process has been difficult. Advances in the field of model interpretability have helped to mitigated this problem [23,24] and for less complex models, it is possible to determine the relative importance of each input metric to the model's final classification [25]. One useful interpretability method is Shapley interpretation, where the features of a machine learning problem are treated as players in a coalitional game from game theory. A specific value called a Shapley value, is assigned to each feature, and represents its contribution to the final classification result. However, while Shapley values produce high quality explanations, their exact computation can be implemented efficiently only in certain (decision tree-based) models, whereas they must be approximated when using other models [25].

The hypothesis tested in this study was that EEG signals recorded during surgical performance on a simulated brain tumor resection task would provide an accurate classification of surgical expertise using machine learning algorithms. The specific objectives were 1) to determine which machine learning algorithm provided the greatest precision in classifying skilled from less-skilled performance on a virtual reality brain tumor resection procedure, 2) to outline which EEG frequency bands were most relevant to this classification, and 3) to gain insight into EEG frequency bands differences in between skilled and less-skilled individuals.

2. Methods

2.1. Study participants

A total of 24 individuals from one institution were enrolled in this study including 6 neurosurgeons, 6 senior neurosurgical residents (post-graduate years 4–6), 6 junior neurosurgical residents (post-graduate years 1–3), and 6 medical students. Data were collected at a single time point and no follow-up data were collected. Collected demographic data included age, gender, handedness, resident training level, and hours of video games and musical instruments played weekly. Participants rated the tumor resection procedure difficulty after each tumor resection on a five-point Likert scale. All participants had previous experience with the NeuroVR™ neurosurgical simulator in a previous study [26]. Since previous research suggests differential EEG patterns between left- and right-handed individuals [27], 2 left-handed participants (1 senior resident and 1 medical student) were excluded. One neurosurgeon's data was not utilized due to excessive noise affecting the EEG recording. See Fig. 1 for an illustration of the inclusion and exclusion of participants. The remaining 21 participants were classified *a priori* as skilled (neurosurgeons and senior residents), or less-skilled (junior residents and medical students) groups based on their patient intraoperative experience with the selected brain tumor procedure. Before starting the study, all participants signed a consent form approved by the McGill University Health Centre Research Ethics Board, Neurosciences-Psychiatry. This study follows Consolidated Standards of Reporting Trials involving Artificial Intelligence (CONSORT-AI) [28] and the best practices for Machine Learning to Assess Surgical Experience (MLASE) reporting guidelines [29].

2.2. NeuroVR™ simulator and simulation scenario

The NeuroVR™ platform (CAE Healthcare, Montreal, Canada), is a high-fidelity virtual reality neurosurgical simulator, providing 3D visual operative experience with haptic feedback (Fig. 2A). The virtual reality simulated brain tumor resection scenario consisted of identical tumors and stiffness with random bleeding points [30]. The color, stiffness, and elliptical structure chosen for each tumor was a simulated glioma-like brain tumor embedded in a simulated cortical surface (Fig. 2B). The task was specifically designed to model patient brain tumor resection procedures. Participants were provided with written and verbal instructions and asked to complete a tumor resection while minimizing bleeding and injury to the surrounding simulated normal tissue.

2.3. Study sequence

Participants were equipped with one active electrode placed on the scalp at Cz in accordance with the International 10–20 system [31] and referenced to linked ears (Fig. 2A). The ProComp Infinity (Thought Technology Ltd., Montreal, Canada) continuously acquired EEG data at a sampling frequency of 256 Hz. Impedance values were kept below 5 kΩ. EEG data were digitally band-pass filtered between 0.3 Hz and 40 Hz. Artefact correction was performed by visually inspecting the raw EEG data and rejecting prototypical artefacts, such as eye blinks and muscle tension.

EEG data collection began with a 2-min eyes-closed, and a 2-min eyes-open baseline recording. Following this baseline, participants resected 6 simulated brain tumors on the NeuroVR™ platform (CAE Healthcare, Montreal, Canada) (Fig. 2B). Participants utilized a simulated surgical aspirator in the dominant hand for tumor resection and a simulated sucker in the non-dominant hand to control bleeding (Fig. 2C) [26]. Participants were affixed to the virtual reality headset for the entire surgical task, preventing their head from moving, thus eliminating the possibility of artefacts associated with the movement of participants. Participants began by resecting tumors 1 and 2 (2 min were allocated per tumor resection). This was followed by a 90-s rest period in

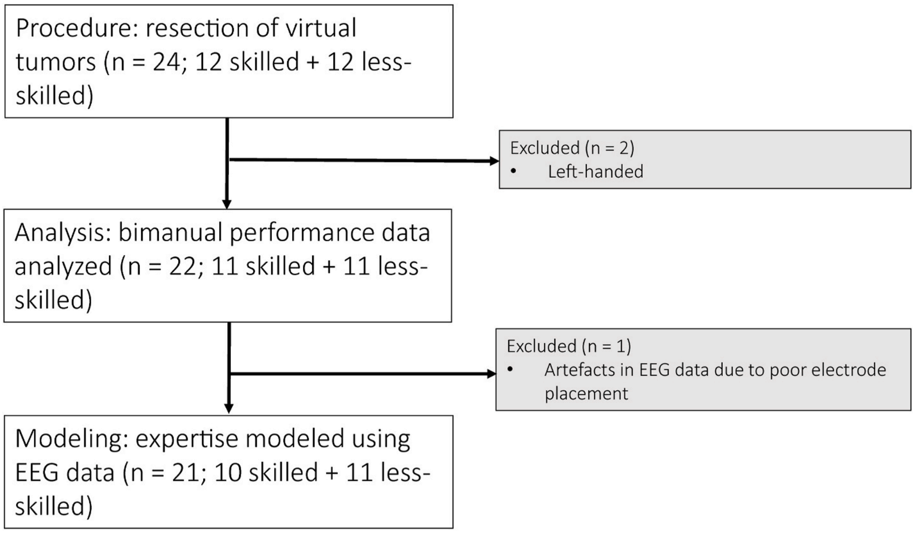
Procedure: resection of virtual tumors (n = 24; 12 skilled + 12 less-skilled)
Excluded (n = 2)
• Left-handed
Analysis: bimanual performance data analyzed (n = 22; 11 skilled + 11 less-skilled)
Excluded (n = 1)
• Artefacts in EEG data due to poor electrode placement
Modeling: expertise modeled using EEG data (n = 21; 10 skilled + 11 less-skilled)

Fig. 1. Inclusion and exclusion of participants
Six neurosurgeons, 6 senior neurosurgical residents (post-graduate year 4–6), 6 junior neurosurgical residents (post-graduate year 1–3), and 6 medical students who expressed an interest in neurosurgery were recruited. Two groups were defined *a priori*. The skilled group included post-graduate year 4 or higher. The less-skilled group included post-graduate years 1–3 and medical students. To ensure homogeneity in handedness, two left-handed trainees were excluded (1 skilled and 1 less-skilled participants). To ensure high fidelity of EEG data, one skilled participant was excluded, for a final sample size of 21 (10 skilled and 11 less-skilled participants).

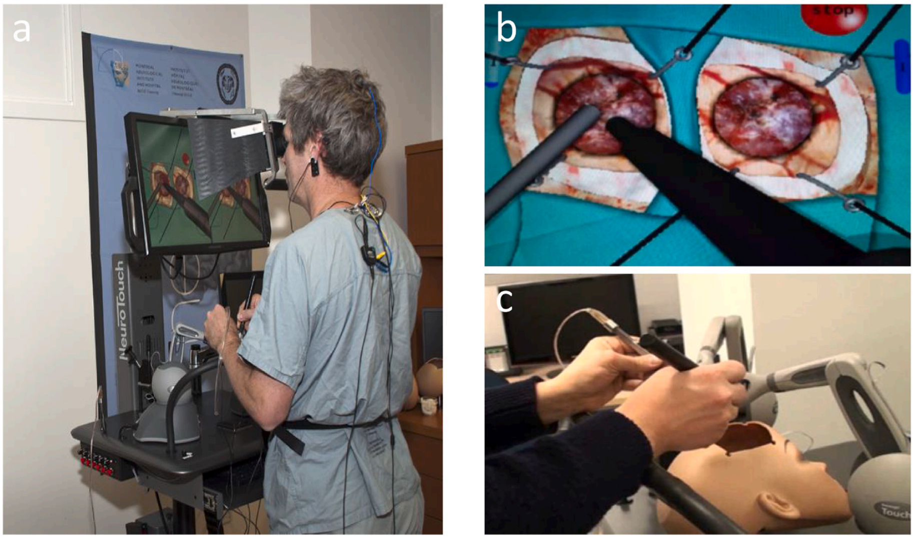

Fig. 2. Virtual neurosurgical experimental setup
(a) A participant performing a simulated brain tumor resection procedure on the NeuroVR™ virtual reality simulation platform whilst equipped with an EEG electrode. Note that the surgical view is perpendicular to the surgical tools. (b) Surgical view demonstrating the simulated surgical aspirator and simulated suction device. (c) Experimental setup with haptic feedback outlining the aspirator held by the dominant hand and sucker in the non-dominant hand.

which participants were instructed to close their eyes and to relax prior to the next task. This sequence was then repeated for scenario two (tumor 3 and tumor 4) and three (tumor 5 and tumor 6). All simulated tumors were identical except for tumor 4, which included uncontrollable intraoperative bleeding resulting in simulated patient cardiac arrest [30]. Due to the acute stress that participants experienced during resection of tumor 4 and impact this may have had on subsequent performance, only data from tumors 1–3 were included in this analysis. Future research will explore differences in expertise under simulated stress. Video 1 is a sample video of the task. Participants completed a post simulated operative questionnaire utilizing a five-point Likert scale to indicate their perception of the difficulty of each tumor resection.

2.4. Feature selection

Fast Fourier Transform (FFT) at 1 Hz resolution was used to separate the raw EEG signal into various power spectra bandwidths using Biograph Infinity software (Thought Technology Ltd., Montreal, Canada). Mean spectral power (μV2) per tumor was then calculated for the 13 relevant bandwidths, which included: delta (2–4 Hz), theta (4–8 Hz), low theta (4–6 Hz), high theta (6–8 Hz), alpha (8–12 Hz), low alpha (8–10 Hz), high alpha (10–12 Hz), beta (13–30 Hz), sensorimotor rhythm (SMR, 12–15 Hz), beta 1 (15–18 Hz), beta 2 (19–22 Hz), beta 3 (23–36 Hz). Thus, the average power was calculated based on the average of artefact-free data for each tumor (up to 2 min), using 1 s FFT window length. Furthermore, the theta/beta ratio (TBR) has been found to be associated with cognitive processing capacity and was thus felt to be an important feature to assess [32]. The TBR is calculated by dividing the square of theta (4–8 Hz) divided by the square of beta (13–21 Hz). All metrics were averaged across each tumor resection per participant. See Table 1 for a detailed analysis of each analyzed feature separated by expertise level.

2.5. Training

Three datapoints were collected per participant, corresponding to the average value of the 13 generated metrics during each tumor resection simulation [30]. Thus, a total of 63 datapoints from 21 participants were available for analysis. Data were randomly divided into training (16 participants, 48 tumors, 76%) and testing datasets (5 participants, 15 tumors, 24%). The testing dataset was composed of 1 neurosurgeon (10 years in practice), 1 senior (post-graduate year 4), 2 junior residents (both post-graduate year 1), and 1 medical student. Statistical comparison of these two datasets revealed no differences in

Table 1
EEG band means across expertise.

EEG Metrics	Skilled (n = 10)	Less-skilled (n = 11)	p-Value
1. Delta (2–4 Hz)	7.92 ± 0.31	8.27 ± 0.47	0.5338
2. Theta (4–8 Hz)	8.21 ± 0.44	7.91 ± 0.45	0.6290
3. Low Theta (4–6 Hz)	6.00 ± 0.24	6.07 ± 0.33	0.8703
4. High Theta (6–8 Hz)	5.60 ± 0.40	5.14 ± 0.34	0.3688
5. Alpha (8–12 Hz)	6.87 ± 0.57	5.77 ± 0.40	0.1183
6. Low Alpha (8–10 Hz)	5.37 ± 0.43	4.33 ± 0.26	0.0443*
7. High Alpha (10–12 Hz)	4.30 ± 0.43	3.81 ± 0.33	0.3671
8. SMR (12–15Hz)	4.14 ± 0.35	3.56 ± 0.17	0.1323
9. Beta (13–30 Hz)	7.94 ± 0.46	6.75 ± 0.36	0.0485*
10. Beta 1 (15–18 Hz)	3.76 ± 0.25	3.02 ± 0.14	0.0141*
11. Beta 2 (19–22 Hz)	3.37 ± 0.18	2.78 ± 0.15	0.0148*
12. Beta 3 (23–36 Hz)	5.96 ± 0.45	5.43 ± 0.38	0.3635
13. TBR Mean (4–8 Hz) [2]/(13–21Hz) [2]	1.93 ± 0.13	2.67 ± 0.33	0.0484*

A comparison between skilled and less-skilled groups on the 13 curated EEG bandwidth metrics selected for this study. Band means were averaged across all three tumor resections per participant. Unpaired two-tailed t-tests were conducted to compare differences between each group, except when the condition of normality was suspect, in which case a Wilcoxon Test was used (Mann-Whitney). Means ± SEs are reported. Significant differences ($p < 0.05$) are denoted by an asterisk. No correction for multiple comparisons has been performed due to a relatively small number of comparisons.

age, years of practice, sex, or proportion of skilled or less-skilled individuals ($p = 0.182$, 0.411, 0.993, and 0.696 respectively). Data was normalized by centering to the mean and scaling component-wise to unit variance, and then shuffled [33].

Seven machine learning algorithms were trained on the training set: Artificial Neural Network (ANN), Naïve Bayes (NB), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF) [34]. These models represent the seven most common algorithms used in the field of artificial intelligence in healthcare [35]. Leave-one-out cross validation was used. This involved the iterative withholding of a participant in the training set, whose membership is predicted by a trained model on all other participant data. This process is repeated until all individuals have been classified. Hyperparameters of each model were manipulated until the training accuracy peaked. A final training was done for each model on the whole training dataset using the optimized hyperparameters. Finally, the trained models were tested on the testing dataset for independent validation.

Since the ANN model provided the highest accuracy, it was selected for interpretation (see Supplementary Fig. 1 for an illustration of the model). A Shapley explainer model was trained to compute the average expected marginal contribution of each EEG metric to each testing participant's tumor resection classification for the model [25]. Shapley values were plotted (Fig. 3). All modelling and interpretability were performed using Python, Tensorflow, and Keras, by code written by the authors.

2.6. Statistical analysis

Pearson's Chi-squared test was used to test differences in proportions, such as gender and expertise differences across the training/testing split and differences in gender across expertise. Unpaired two-tailed T-Tests were used to compare participant age across expertise groups and training/testing split. A Kruskal-Wallis test was used to compare tumor difficulty ratings between expertise groups, due to the scale's discontinuity. Regression analysis was conducted on correlations between age and each of the EEG metrics as well as years in practice and each of the EEG metrics. All findings were assessed at the 0.05 alpha level for significance.

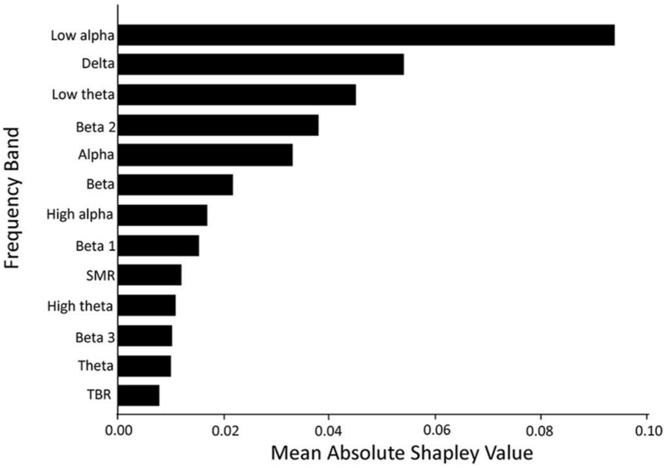

Fig. 3. Shapley interpretability plot
Bar plot illustrating the contribution of each frequency band as calculated by Shapley analysis on the final artificial neural network, in descending order. Shapley values are borrowed from game theory and attempt to quantify the marginal contribution of each player to the final result of a game, with a greater values representing greater contributions [25]. In the case of a machine learning model, a player is an input metric and a final result is the overall model classification. Shapley interpretability thus allows for a model-agnostic interpretation of feature importance. The low alpha band was by far the most important factor in expertise classification.

3. Results

Demographic information is presented in Table 2. Eighteen (87.7%) of the participants were male, with mean age (SD) of the skilled and less-skilled group being 37.2 (8.1) and 26.2 (3.0) which were significantly different ($p < 0.001$). Three participants (12.5%) played musical instruments, whereas 8 (33%) reported playing video games. On the night before participating average sleep time was 6.0 ± 1.5 h, suggesting participants were relatively well rested. No differences in sleep between expertise groups was found ($p = 0.309$). Skilled and less-skilled participants rated the tumor resection procedure difficulty, with mean (SD) of 3.17 (0.83) and 3.70 (0.89) on average on the five-point Likert scale,

Table 2
Participant demographics stratified by expertise level.

	Skilled	Less-skilled	P value
Composition	5 Neurosurgeons 5 Senior Residents	6 Junior Residents 5 Medical Students	
Age ± SD	37.2 ± 8.1	26.2 ± 3.0	0.0005*
Gender, No (%)			
Male	8 (80%)	10 (90.9%)	
Female	2 (20%)	1 (9.1%)	
Years in Medicine (range)	15.45 (8–26)	4.55 (3–7)	0.0001*
Difficulty ratings ± SD			
Tumor 1	3.40 ± 0.93	3.36 ± 0.90	0.8603
Tumor 2	2.90 ± 0.70	3.72 ± 1.06	0.0783
Tumor 3	3.10 ± 0.87	3.72 ± 0.72	0.1392

Demographic data and tumor difficulty ratings (on a five-point Likert scale) of the 10 skilled and 11 less-skilled participants. A two-tailed unpaired T-Test was used to compare age and years in practice differences across expertise groups. A Kruskal-Wallis non-parametric test was used to compare tumor difficulty ratings. Years in practice calculation assumes 4 years of medical school, 6 years of residence training, and 2 years of fellowship, as is standard in neurosurgical education. Significant differences of $p < 0.05$ are denoted by an asterisk. Skilled participants were significantly older ($p = 0.0005$) and more experienced (0.0001) than less-skilled participants. Since expertise categories were based on education level attained and education level was highly correlated to age, these differences are expected. There were no significant differences ($p > 0.05$) in the participants' subjective ratings of each tumor's difficulty. Skilled and less-skilled participants found each tumor moderately difficult.

respectively. Statistical analysis revealed no difference (p = 0.851, 0.067, and 0.110 respectively) in their subjective perception of the difficulty of each tumor resection procedure following operation, with a greater number signifying greater difficulty (Table 1).

The classification accuracies for training and testing are illustrated in Table 3 and the final confusion matrices of the ANN modelling are shown in Fig. 4. The sensitivity and specificity, as well as the F-Measure (the harmonic mean between the precision and the sensitivity), are reported. A receiver operating curve (ROC) was constructed for each testing model to calculate the area under the ROC (AUROC). Due to the slight class imbalance that was present in the testing set (3 less-skilled vs. only 2 skilled), metrics such as F-Measure and AUROCs are more representative of the results relative to accuracy.

The best performing model, ANN, was selected for model interpretation. Shapley value interpretations are plotted in order of magnitude in Fig. 3. The low alpha band was most important in expertise classification and the TBR, a composite of the theta and beta band, was the least important metric.

To explore EEG activity further, we averaged the EEG results across the 3 tumor resections by participant. Then, we compared the skilled and less-skilled groups using unpaired two tailed T-Tests (Table 1). Significantly higher average values of low alpha, overall beta, beta 1, and beta 2 (p = 0.0443, 0.0485, 0.0141, 0.0148, respectively) were found for the skilled compared to the less-skilled group. In addition, a significantly lower TBR was found for the skilled compared to the less-skilled group (p = 0.0484).

4. Discussion

The combination of virtual reality simulation, EEG, and machine learning provides an opportunity to classify surgical expertise. This study demonstrates that an artificial neural network model can predict skilled and less-skilled participant levels of expertise based on EEG recordings with high fidelity during the performance of virtual reality simulated brain tumor resections. It is important to note that previous work has linked educational level with surgical success. For instance, it has been demonstrated that neurosurgeons performed simulated tumor resections with significantly less blood loss than medical students (0.23 vs 0.44) and resected a significant amount of tumor more than medical students (94% vs 47%). Given that resection time was standardized across tumors, this amounted to a faster resection speed amongst neurosurgeons compared to medical students [30]. Thus, classification of surgical expertise according to education level is indicative of performance.

We utilized the Shapley model interpretability technique [25] to conduct an analysis of the metrics that the model identified as important in classification, thus allowing a determination of the relative importance of EEG bands in expertise classification (Fig. 3). A statistical analysis in average EEG bands provided the differences between skilled and less-skilled EEG activity (Table 1).

Since low alpha is associated with calmness and neural efficiency [12], our findings suggest that skilled participants may have acquired abilities resulting in operating with greater composure and purpose than less-skilled participants [36]. Neural efficiency is related to the neural efficiency hypothesis, which states that skilled individuals tend to exhibit lower neural activity during the same cognitive task compared to less-skilled individuals [37]. This result reinforces the concept that skilled surgical performance involves cognitive elements such as enhanced composure and focus [38].

Although beta waves were relatively less important on our Shapley classification, most beta bands (beta 1, beta 2, and overall beta) were significantly different between groups. Skilled participants consistently had higher levels of beta waves (Table 1), suggesting that skilled participants may more consistently operate with greater attention and problem-solving abilities [39]. Skilled participants exhibited significantly lower (p = 0.0484) TBR, consistent with usage of TBR as a means of assessing expertise in several other fields. The TBR is considered a marker of cognitive processing capacity, a quality of importance in bimanual psychomotor surgical performance [32]. Our model did not put a high emphasis on the TBR (lowest Shapley value of the 13 metrics, Fig. 3), which may relate to TBR being a composite measure derived from its interactions with two or more other metrics inputted into the model. In contrast, a model based primarily on TBR may be able to outline this metric as important in expertise classification.

Several of our models were unable to accurately classify expertise, particularly during the testing phase. On examining the testing misclassifications, all the models that were generated accurately classified neurosurgeons and medical students—the extreme ranges of surgical skill levels in this investigation—but failed to accurately classify senior residents (6/32 misclassifications) and junior residents (26/32 misclassifications). In this study residents were assigned to a group based on their year of training. The *a priori* classification system used in this study to place participants into the skilled or less skilled groups may not have been able to accurately outline the actual surgical skills of individuals especially between the third and fourth year of neurosurgical training in which training of subpial resection may be variable. A more comprehensive method to classify trainee expertise level using quantitative assessment across a defined series of operative skills may improve the accuracy of these machine learning classification systems. However, our artificial neural network was able to achieve perfect testing accuracy, demonstrating the robustness of our final model and suggesting that this model has better classification precision and granularity. By calculating Shapley values and plotting them from the most to the least important EEG metric assessed, the Shapley graph allows for the prioritization of metrics for surgical training. In developing a neurofeedback method that builds on our system, to maximize training efficacy, we would recommend focusing the training protocol by iteratively training on the EEG metrics in order of their Shapley values.

Table 3
Modelling results.

Classifier	Training Accuracy	Accuracy	Sensitivity	Specificity	F-Measure	AUROC
Artificial Neural Network	0.979	1.0	1.0	1.0	1.0	1.0
Support Vector Machine	0.958	0.667	1.0	0.8	0.800	0.833
Logistic Regression	0.934	0.556	1.0	0.733	0.714	0.778
K Nearest Neighbors	0.833	0.556	1.0	0.733	0.714	0.778
Linear Discriminant Analysis	0.896	0.733	0.778	0.667	0.738	0.722
Naïve Bayes	0.833	0.778	0.5	0.667	0.571	0.639
Random Forest	0.833	0.667	0.500	0.600	0.667	0.583

The seven most common machine learning model types in healthcare are compared in their ability to distinguish between skilled and less-skilled participants on a virtual reality surgical simulation. Models are ordered by the area under the receiver operating curve (AUROC). Training accuracy, testing accuracy, sensitivity, specificity, F-Measures and AUROCs are reported. All metrics reported other than the training accuracy are derived from the testing set. Algorithm prediction sensitivity and specificity are provided. The F-Measure is the harmonic mean of the precision (true positives over all positives) and the sensitivity. Testing accuracies varied from 67% to 100%, with the artificial neural network (ANN) classifying all participants in the testing set correctly.

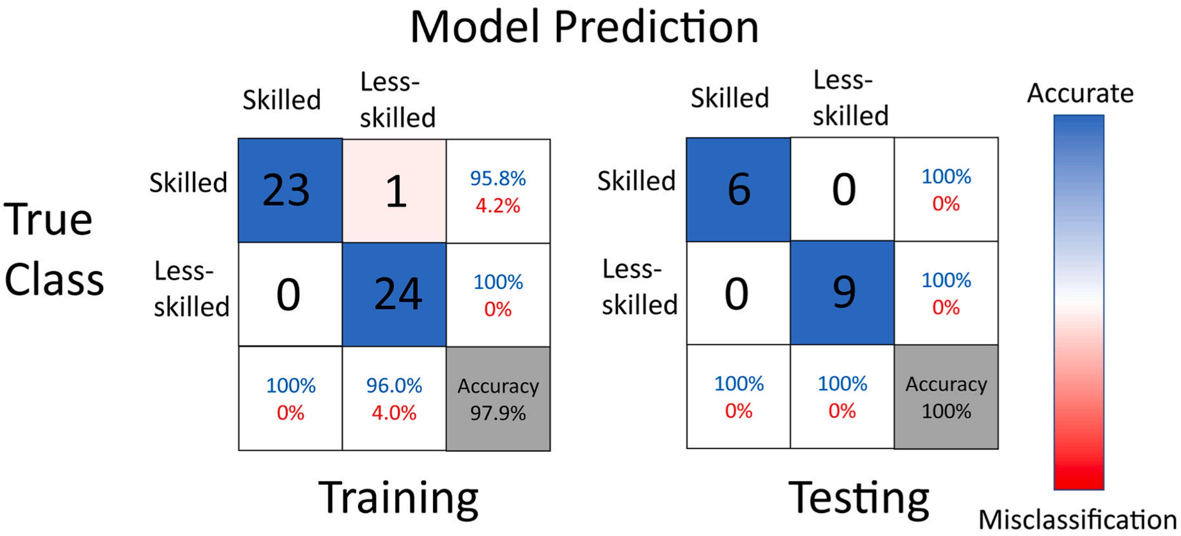

Fig. 4. Confusion matrices of the training and testing results of the artificial neural network
The first confusion matrix illustrates the averaged results from 16 different neural network models, which were trained on tumor resections from 15 participants, leaving one participant for validation in a leave-one-out-cross-validation (LOOCV) fashion (8 skilled and 8 less-skilled in total). Each participant carried out three simulated tumor resections for a total of 48 training procedures. One skilled participant, corresponding to a fourth-year neurosurgical resident, was misclassified as less-skilled during one of their surgical resections, rendering a final training accuracy of 97.9%. A final neural network was trained on all available training data based on the hypertuned parameters arrived at from the LOOCV procedure. The second confusion matrix illustrates the final testing results of this neural network. It achieved 100% accuracy on the 5 testing participants (2 skilled and 3 less-skilled participants).

4.1. Strengths

By assessing several different model types and ordering them based on testing AUROC, we provide evidence that artificial neural networks are the most adept at analyzing averaged EEG data from surgical simulation. Several advantages are intrinsic to EEG monitoring systems. First, since EEG waves may precede action, EEG band frequency data may be utilized to predict future bimanual psychomotor performance and with the application of a feedback system help improve task execution and potentially mitigate potential technical skill errors [40]. Since EEG has a high sampling rate (256 Hz in this study, but potentially much higher), classifications may be possible in real-time, thus allowing for real-time feedback. Although we did not exploit this rapid sampling rate in the present study, by averaging EEG results, we were able to achieve accurate classifications with 2-min tumor resections data, allowing for personalized post-hoc feedback training. It is possible to collect EEG data concurrently and integrate these with other artificial intelligence derived biometrics performance platforms [22] to build a holistic model to both improve our understanding of surgical expertise in a specific surgical setting and suggest modulation of trainee performance to achieve optimal performance [22,41].

4.2. Limitations

There are limitations to this pilot study. Virtual reality simulation allows detailed assessment of bimanual psychomotor technical skills however these systems are unable to recreate the many elements of the dynamic and interactive operating room environment. While spectral analysis is a established technique of quantitating EEG patterns [42], these evaluations provide an incomplete assessment of motor, sensory and cognitive interaction in complex bimanual psychomotor skills involved in surgical procedures. In this study we utilized only one EEG electrode to obtain average spectral band data associated for each of the three individual tumor resections. The advantages of using a single electrode included, less interference with the participants perception of a realistic operative experience, simplicity of EEG scalp application resulting in decreased start-up time and improved cost-effectiveness [43]. Disadvantages included the inability to assess EEG temporal (EEG band variance with time) or spatial analysis, which outlines physiological brain locations underlying the EEG information. However, using one electrode and an ANN machine learning algorithm model we were still able to classify skilled and less-skilled participants with 100% accuracy. Thus, although this level of accuracy may not be possible in all neurosurgical situations given one electrode, it represents a baseline and a first step for evaluating the technical skills of neurosurgical residents. Utilizing multiple electrodes in future studies will provide temporal and spatial data and further our understanding of the relationship between EEG and surgical expertise. EEG data lends itself to timeseries analysis and specialized deep learning algorithms such as the long-short term memory (LSTM) models. Since one of our goals was the implementation of an AI-powered individualized EEG neurofeedback platform to improve learner skill acquisition, the utilization of EEG mean data [44] rather than EEG timeseries information was felt to be easier for trainees to understand and learn. Since EEG [27] and hand ergonomics [45] exhibit differences between left and right-handed individuals, left-handed participants were excluded from this investigation (Fig. 1) preventing our commenting on their EEG patterns during simulated resection. This study involved only a small number of participants from one institution, which limits the generalization of our results. Using larger datasets from multiple institutions, including individuals with quantifiable levels of expertise, would enhance the robustness of models and the precision and granularity of the classification.

It has been shown that higher participant age is associated with changes in EEG patterns, such as increasing beta activity and decreased alpha activity [46,47]. In this study the testing group, which included 5 participants and 15 assessed tumor resections, was composed of a 29-year-old senior and 29- and 30-year-old junior neurosurgical residents. Our ANN model's ability to accurately classify skilled and less-skilled performance despite the overlap in ages suggests that the model was not classifying based on age-related factors.

Although regression analysis of EEG frequency bands during eyes closed and open baselines reveals significant correlation between some band frequencies and participant age, these correlations were rarely as strong as their years in practice counterparts. Moreover, alpha peak frequency (IAF), a robust metric of brain maturation [48], did not significantly correlate with eyes closed or open baselines and age (p = 0.1204 and 0.4004 respectively). Applying our accurate ANN model to the eyes open and closed baselines EEG data yielded classification accuracies of only 40 and 60%, respectively (results not shown). Taken together these results support the conclusion that the ANN model's ability to classify surgical performance in the simulation utilized in this trial is based on this model's ability to use EEG frequency wave rather than age-dependent EEG data.

4.3. Future directions

Studies involving the utilization of more frequent EEG analysis by multiple electrodes will provide more extensive EEG data which will improve our understanding of the relationship between specific temporal and spatial EEG frequency bands and surgical expertise. In particular, electrodes on sensorimotor areas (C3 and C4) may further elucidate bimanual technical expertise. The utilization of specialized deep learning algorithms such as the long-short term memory (LSTM) models for timeseries analysis may result in the development of continuous monitoring of expertise systems which provides personalized feedback and may allow for tutoring and risk detection. The combination of the EEG-dependent ANN model outlined in this study and AI-powered intelligent tutoring platforms, such as the Virtual Operative Assistant (VOA) which utilizes safety and efficiency metrics generated from the support vector machine algorithm [49] for competency evaluation could be assessed in randomized controlled trials [41].

EEG data classification results provides an opportunity for continuous neurofeedback which could provide users with increased self-awareness of their EEG patterns during operative performance. By interpreting which EEG series of metrics the model finds most useful in classifying specific skilled operative performance and alerting learners by neurofeedback methodology, surgical trainees could self-modify their own EEG metrics to approximate these EEG frequency metrics and improve task execution. A proposed neurofeedback system using the algorithms developed in this study is outlined in Fig. 5.

These investigations could help determine which system or combination of systems is more effective in formative surgical training. Artificial intelligent EEG classification systems based on machine and deep learning powered educational platforms could be implemented during human operative procedures, resulting in the development of AI-powered "Smart Operating Rooms". These platforms could offer trainees continuous monitoring of their bimanual psychomotor surgical skills while providing personalized expert-level coaching, error detection, and mitigation of patient risk.

5. Conclusion

Machine learning algorithms successfully differentiated EEG activity between skilled and less-skilled groups during a simulated bimanual surgical task. Our methodology aids in the understanding the components of EEG which contribute to bimanual technical expertise. This system may enhance the ability of surgical educators to develop more quantitative, formative, and summative assessment paradigms to deal with future challenging pedagogic requirements. Machine learning-powered EEG classification systems offer objective, and generalizable continuous monitoring which can be adapted to the evaluation and training of all procedural-based bimanual technical skills interventions.

Funding sources

This work was supported by the Franco Di Giovanni Foundation, Mitacs-Own the Podium collaboration, a Fellowship of Excellence from McGill University, the Royal College of Physicians and Surgeons of Canada, and the Montreal Neurological Institute and Hospital. Dr. Winkler-Schwartz received a Doctoral Training Grant for Applicants with a Professional Degree the Fonds de recherche du Québec – Santé (fund # 261422) and holds the Robert Maudsley Fellowship for Studies in Medical Education by the Royal College of Physicians and Surgeons of Canada. The National Research Council of Canada provided a prototype of the NeuroTouch™/NeuroVR™ simulator that was used in this study.

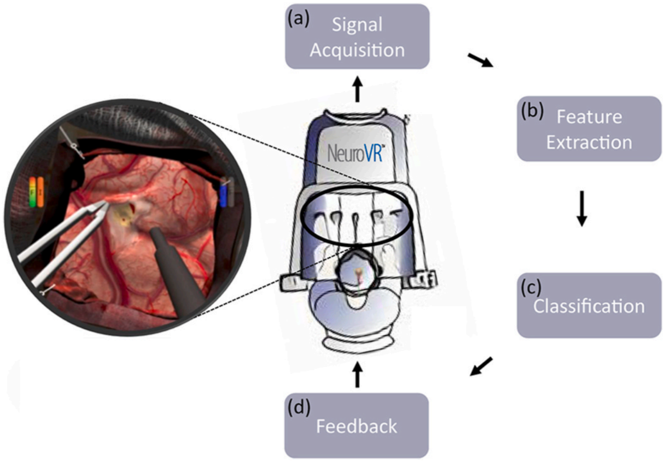

Fig. 5. Proposed neurofeedback training protocol in surgical simulation training
(a) Learners perform a surgical procedure(s) on a virtual reality simulator while equipped with an EEG electrode(s), allowing for raw EEG signal capture. **(b)** The raw signal is processed and specific metrics such as EEG waves bands are extracted. **(c)** The extracted metrics are fed into a machine learning model, which objectively classifies the individual's performance as skilled or less-skilled. **(d)** The expertise classification, along with the resultant explanation of why it was assigned as such, is displayed to the trainee. Training is iteratively done in the order of the model interpretability rankings.

Declaration of competin interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We particularly would like to thank all the neurosurgeons, neurosurgical residents, and medical students who participated in this study. The authors also thank Dr. Jose Andres Correa, Department of Mathematics and Statistics, McGill University, for expert input on statistical analysis. The authors also thank the National Research Council of Canada, Boucherville, Quebec, Canada for their assistance in developing the scenarios used in this study. We acknowledge the National Neurosciences Institute, King Fahad Medical City, Riyadh, Saudi Arabia, for funding help in the creation of the complex realistic brain tumor model.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.106286.

References

[1] D. Silbergeld, A. Hebb, T. Yang, The sub-pial resection technique for intrinsic tumor surgery, Surg. Neurol. Int. 2 (2011) 180.
[2] J.J. Stulberg, et al., Association between surgeon technical skills and patient outcomes, JAMA Surg 155 (2020) 960–968.
[3] S. Delorme, D. Laroche, R. Diraddo, F. Del Maestro, R. NeuroTouch, A physics-based virtual simulator for cranial microneurosurgery training, Neurosurgery 71 (2012) 32–42.
[4] N. Ledwos, et al., Virtual reality anterior cervical discectomy and fusion simulation on the novel sim-ortho platform: validation studies, Oper. Neurosurg. 20 (2021) 74–82.

[5] A.J. Sabbagh, et al., Roadmap for developing complex virtual reality simulation scenarios: subpial neurosurgical tumor resection model, World Neurosurg 139 (2020) e220–e229.

[6] V.N. Palter, T.P. Grantcharov, Individualized deliberate practice on a virtual reality simulator improves technical performance of surgical novices in the operating room: a randomized controlled trial, Ann. Surg. 259 (2014) 443–448.

[7] N. Mirchi, N. Ledwos, R.F. Del Maestro, Intelligent tutoring systems: Re-envisioning surgical education in response to COVID-19, Can. J. Neurol. Sci./J. Can. des Sci. Neurol (2020) 1–3, 00.

[8] K. Lam, et al., Machine learning for technical skill assessment in surgery: a systematic review, npj Digit. Med. 5 (2022).

[9] T. Bocci, et al., How does a surgeon's brain buzz? An EEG coherence study on the interaction between humans and robot, Behav. Brain Funct. 9 (2013).

[10] W. Klimesch, EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis, Brain Res. Rev. (1999) 29 169–195.

[11] H. Marzbani, H.R. Marateb, M. Mansourian, Methodological note: neurofeedback: A comprehensive review on system design, methodology and clinical applications, Basic Clin. Neurosci. 7 (2016) 143–158.

[12] C. Babiloni, et al., Golf putt outcomes are predicted by sensorimotor cerebral EEG rhythms, J. Physiol. 586 (2008) 131–139.

[13] S. Christie, M. Bertollo, P. Werthner, The effect of an integrated neurofeedback and biofeedback training intervention on ice hockey shooting performance, J. Sport Exerc. Psychol. 42 (2020) 34–47.

[14] A.S. Al-Fahoum, A.A. Al-Fraihat, Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains, ISRN Neurosci. 2014 (2014) 1–7.

[15] J.A.M. Sidey-Gibbons, C.J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC Med. Res. Methodol. 19 (2019).

[16] T. Oladipupo, Types of machine learning algorithms, in: New Advances in Machine Learning, InTech, 2010, https://doi.org/10.5772/9385.

[17] F. Jiang, et al., Artificial intelligence in healthcare: past, present and future, Stroke and Vascular Neurology (2017) 2 230–243.

[18] J.T. Senders, et al., An introduction and overview of machine learning in neurosurgical care, Acta Neurochir. (Wien). 160 (2018) 29–38.

[19] S. Alkadri, et al., Utilizing a multilayer perceptron artificial neural network to assess a virtual reality surgical procedure, Comput. Biol. Med. 136 (2021), 104770.

[20] V. Bissonnette, et al., Artificial intelligence distinguishes surgical training levels in a virtual reality spinal task, J. Bone Jt. Surg. - Am. 101 (2019).

[21] N. Mirchi, et al., Artificial neural networks to assess virtual reality anterior cervical discectomy performance, Oper. Neurosurg. 19 (2020) 65–75.

[22] 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 (2019), e198363.

[23] R. Elshawi, M.H. Al-Mallah, S. Sakr, On the interpretability of machine learning-based model for predicting hypertension, BMC Med. Inf. Decis. Making 19 (2019).

[24] J. Adebayo, L. Kagal, Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models, 2016.

[25] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process..Syst. vols 2017-Decem (2017) 4768–4777.

[26] G. Alzhrani, et al., Proficiency performance benchmarks for removal of simulated brain tumors using a virtual reality simulator neurotouch, J. Surg. Educ. 72 (2015) 685–696.

[27] K.A. Provins, P. Cunliffe, The relationship between E.E.G. Activity and handedness, Cortex 8 (1972) 136–146.

[28] X. Liu, S. Cruz Rivera, D. Moher, M.J. Calvert, A.K. Denniston, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension, Nat. Med. (2020) 26 1364–1374.

[29] 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 (2019) 1681–1690.

[30] K. Bajunaid, et al., Impact of acute stress on psychomotor bimanual performance during a simulated tumor resection task, J. Neurosurg. 126 (2017) 71–80.

[31] H.H. Jasper, Report of the committee on methods of clinical examination in electroencephalography, Electroencephalogr. Clin. Neurophysiol. 10 (1958) (1957).

[32] A.R. Clarke, R.J. Barry, D. Karamacoska, S.J. Johnstone, The EEG theta/beta ratio: a marker of arousal or cognitive processing capacity? Appl. Psychophysiol. Biofeedback 2019 *442* (44) (2019) 123–129.

[33] X.H. Cao, I. Stojkovic, Z. Obradovic, A robust data scaling algorithm to improve classification accuracies in biomedical data, BMC Bioinf. 17 (2016) 359.

[34] J. Yuan, Y.M. Li, C.L. Liu, X.F. Zha, Leave-one-out cross-validation based model selection for manifold regularization, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6063, Springer, Berlin, Heidelberg, 2010. LNCS 457–464.

[35] Y. Jiang, et al., A Brief Review of Neural Networks Based Learning and Control and Their Applications for Robots, 2017, https://doi.org/10.1155/2017/1895897.

[36] S. Christie, Individual alpha peak frequency in ice hockey shooting performance, Front. Psychol. 8 (2017) 762.

[37] C. Del Percio, et al., 'Neural efficiency' of athletes' brain for upright standing: a high-resolution EEG study, Brain Res. Bull. 79 (2009) 193–200.

[38] V. Pandey, et al., Technical skills continue to improve beyond surgical training, J. Vasc. Surg. 43 (2006) 539–545.

[39] M. Roohi-Azizi, L. Azimi, S. Heysieattalab, M. Aamidfar, Changes of the brain's bioelectrical activity in cognition, consciousness, and some mental disorders, Med. J. Islam. Repub. Iran 31 (2017) 307–312.

[40] I. Fried, P. Haggard, B.J. He, A. Schurger, Volition and action in the human brain: processes, pathologies, and reasons, J. Neurosci. 37 (2017) 10842–10847.

[41] A.M. Fazlollahi, et al., Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students, JAMA Netw. Open 5 (2022), e2149008.

[42] O. Dressler, G. Schneider, G. Stockmanns, E.F. Kochs, Awareness and the EEG power spectrum: analysis of frequencies, Br. J. Anaesth. 93 (2004) 806–809.

[43] J.M. Morales, J.F. Ruiz-Rabelo, C. Diaz-Piedra, L.L. Di Stasi, Detecting mental workload in surgical teams using a wearable single-channel electroencephalographic device, J. Surg. Educ. 76 (2019) 1107–1115.

[44] P. Nagabushanam, S. Thomas George, S. Radha, EEG signal classification using LSTM and improved neural network algorithms, Soft Comput. 24 (2020) 9981–10003.

[45] R. Sawaya, et al., Virtual reality tumor resection: the force pyramid approach, Oper. Neurosurg. 14 (2018) 686–696.

[46] X. Zhong, J.J. Chen, Variations in the Frequency and Amplitude of Resting-State EEG and fMRI Signals in Normal Adults: the Effects of Age and Sex, 2020, https://doi.org/10.1101/2020.10.02.323840 bioRxiv 2020.10.02.323840.

[47] B. Feige, S. Scaal, M. Hornyak, H. Gann, D. Riemann, Sleep electroencephalographic spectral power after withdrawal from alcohol in alcohol-dependent patients, Alcohol Clin. Exp. Res. 31 (2007) 19–27.

[48] J.C. Edgar, et al., Abnormal maturation of the resting-state peak alpha frequency in children with autism spectrum disorder, Hum. Brain Mapp. 40 (2019) 3288–3298.

[49] N. Mirchi, et al., The Virtual Operative Assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine, PLoS One 15 (2020), e0229596.