Utilizing Machine Learning and Electroencephalography to Assess Expertise in Virtual Neurosurgical Performance

Sharif Natheir

Experimental Surgery

McGill University, Montreal

August 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science in Experimental Surgery with a specialization in Surgical Innovation at McGill University.

Contents

ABSTRACT

Background: Neurosurgical education has been facilitated by virtual reality surgical simulators, which provide a safe training environment for trainees. Electroencephalography (EEG) has been used to assess the electrical activity of the brain during surgical performance. When divided into various frequency bands, EEG data lends itself well to analysis by machine learning, a branch of artificial intelligence. Although machine learning classification has traditionally been difficult to interpret, advances in model interpretability techniques, such as those utilizing Shapley values, allow for the determination of the significance of each input metric to the overall model classification. Although EEG has been used widely to explore expertise in sport and aviation simulation training, it has yet to be used to classify surgical expertise.

Objective: The goals of this study were (a) to develop a machine learning model to accurately differentiate skilled and less-skilled individuals using EEG performance data recorded during a simulated surgery task, (b) to explore the relative importance of selected EEG bandwidths to surgical expertise, and (c) to gain insight into differences in EEG bands between skilled and less-skilled individuals.

Hypothesis: EEG recordings during a virtual reality surgery task would accurately predict the expertise level of participants.

Methods: Twenty-one participants performed three simulated brain tumor resection procedures on the NeuroVR™ platform (CAE Healthcare, Montreal, Canada) while EEG data was simultaneously recorded. Participants were divided into 2 groups. The skilled group was composed of five neurosurgeons and five senior neurosurgical residents (post-graduate years 4-6) and the

less-skilled group was composed of six junior residents (post-graduate year 1-3) and five medical students. A total of 13 metrics from EEG frequency bands and ratios (e.g., theta, alpha, beta, theta/beta ratio) were generated. Machine learning models were trained using EEG activity to differentiate between skilled and less-skilled groups. The relative importance of each EEG metric was calculated using Shapley values.

Results: Seven models were trained with the artificial neural network achieving a 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. Beta (13–30 Hz), beta 1 (15–18 Hz), beta 2 (19–22 Hz) and the theta/beta ratio (TBR) were also shown to be important metrics for discerning expertise. Skilled surgeons showed significantly lower TBR ($p = 0.048$) and significantly higher beta, beta 1, and beta 2 ($p = 0.049$, 0.014, 0.015 respectively).

Conclusion: Machine learning algorithms successfully differentiates EEG activity between skilled and less-skilled groups during a simulated bimanual surgical task. Our methodology aids in the understanding of which EEG components contribute to expertise. The findings from this study can be used to help inform future research on surgical skill education and develop intelligent tutoring systems to provide trainees neurofeedback on virtual reality surgical simulation.

Key words: Brain tumor resection, Artificial intelligence, Artificial neural networks, Machine learning, Simulation, Surgical training, Virtual reality, Neurofeedback

RÉSUMÉ

Contexte : L'enseignement de la neurochirurgie a été facilité ces dernières années par des simulateurs chirurgicaux en réalité virtuelle, qui offrent un environnement de formation sûr aux stagiaires. L'électroencéphalographie (EEG) a été utilisée pour évaluer l'activité électrique du cerveau pendant la performance chirurgicale. Lorsqu'elles sont divisées en différentes bandes de fréquences, les données EEG se prêtent bien à l'analyse par apprentissage automatique, une branche de l'intelligence artificielle. Bien que la classification de l'apprentissage automatique soit traditionnellement difficile à interpréter, les progrès des techniques d'interprétabilité des modèles, telles que celles utilisant les valeurs de Shapley, permettent de déterminer l'importance de chaque métrique d'entrée pour la classification globale du modèle. Bien que l'EEG ait été largement utilisé pour explorer l'expertise dans la formation par simulation sportive et aéronautique, il n'a pas encore été utilisé pour classer l'expertise chirurgicale.

Objectif : Les objectifs de cette étude étaient (a) de développer un modèle d'apprentissage automatique pour différencier avec précision les individus qualifiés et moins qualifiés en utilisant les données de performance EEG enregistrées lors d'une tâche de chirurgie simulée, (b) d'explorer l'importance relative des bandes passantes EEG sélectionnées expertise, et (c) pour mieux comprendre les différences dans les bandes EEG entre les individus qualifiés et moins qualifiés.

Hypothèse : Les enregistrements EEG lors d'une tâche de chirurgie en réalité virtuelle prédisent avec précision le niveau d'expertise de l'interprète.

Méthodes : Vingt-et-un participants ont effectué trois résections simulées de neuroblastome sur la plateforme NeuroVR™ (CAE Healthcare, Montréal, Canada) tandis que les données EEG étaient

enregistrées simultanément. Les participants ont été divisés en 2 groupes. Le groupe qualifié était composé de cinq neurochirurgiens et de cinq résidents en neurochirurgie (années d'études supérieures 4-6) et le groupe moins qualifié était composé de six résidents juniors (années d'études supérieures 1-3) et de cinq étudiants en médecine. Un total de 13 métriques de bandes de fréquences et de rapports EEG (par exemple, thêta, alpha, bêta, rapport thêta/bêta) ont été générés. Des modèles d'apprentissage automatique ont été formés à l'aide d'une activité EEG pour différencier les groupes qualifiés et moins qualifiés. L'importance relative de chaque mesure EEG a été calculée à l'aide des valeurs de Shapley.

Résultats : Sept modèles ont été entraînés avec le réseau de neurones artificiels atteignant une précision de test de 100 % (AUROC = 1,0). L'interprétation du modèle via l'analyse de Shapley a identifié un faible alpha (8 à 10 Hz) comme la mesure la plus importante pour classer l'expertise. Les chirurgiens qualifiés ont affiché un alpha faible ($p = 0,044$) plus élevé que le groupe moins qualifié. De plus, le bêta (13-30 Hz), le bêta 1 (15-18 Hz), le bêta 2 (19-22 Hz) et le rapport thêta/bêta (TBR) se sont avérés être des métriques importantes pour discerner l'expertise. Les chirurgiens qualifiés ont montré un TBR significativement plus faible ($p = 0,048$) et des bêta, bêta 1 et bêta 2 significativement plus élevés ($p = 0,049$, 0,014, 0,015 respectivement).

Conclusion : Les algorithmes d'apprentissage automatique différencient avec succès l'activité EEG entre les groupes qualifiés et moins qualifiés au cours d'une tâche chirurgicale bimanuelle simulée. De plus, notre méthodologie aide à comprendre quels composants de l'EEG contribuent à l'expertise. Les résultats de cette étude peuvent être utilisés pour éclairer les futures recherches sur l'enseignement des compétences chirurgicales et développer des systèmes de tutorat

intelligents pour fournir aux stagiaires un neurofeedback sur la simulation chirurgicale en réalité virtuelle.

Key words: Résection d'une tumeur cérébrale, Intelligence artificielle, Réseaux de neurones artificiels, Machine learning, Simulation, Formation chirurgicale, Réalité virtuelle, Neurofeedback

ACKNOWLEDGEMENTS

I would like to extend my sincere gratitude to my supervisor, Dr. Rolando Del Maestro. His flexibility during the COVID-19 Pandemic has been inspiring and his guidance and support throughout this past year made this project possible. Through his hosting of our weekly Friday Forums, I was able to get to know the members of the lab despite working remotely whilst learning about neurosurgical education, Leonardo Da Vinci, the soul, and a myriad of other interesting topics, all whilst benefiting from Dr. Del Maestro's life advice.

I am further grateful for the support of Dr. Sommer Christie throughout this project. Her passion for sports and excellence in this regard have been an inspiration and her tutorage in electroencephalography and related neuroscience concepts was extremely helpful. I am further grateful for the support and friendship of Dr. Recai Yilmaz throughout this project. He has taught me a lot about applied artificial intelligence and I benefited tremendously from the ability to exchange and discuss various ideas together. Furthermore, his kindness is truly contagious.

I would also like to thank all of those who regularly attended our Friday Forums. Through this medium, we turned what may have otherwise been a local board room meeting into an international collaborative platform. Thank you to Ali Fazlollahi, Nykan Mirchi, Nicole Ledwos, Mostafa Siddiqui, Emerson De Fazio, Paolo Alimonti, Kayla Benson, and Drs. Ahmad Alsayegh and Mohamad Bakhaidar, in addition to those already mentioned.

I would also like to thank my co-supervisor, Dr. Carlo Santaguida, and the members of my Research Advisory Committee, Drs. Jason Harley and Jeffery Hall, for their support and flexibility. I would also like to thank my first RAC chair, Dr. Hadil Al-Jallad, for her support at the onset of

this project, and my second chair, Dr. Rahul Gawri, who kindly stepped in and supported me afterwards. Likewise, thank you to Dr. Roy Dudley for his grading of, and feedback on, my study. Thank you as well to Pam Del Maestro, who edited and provided me with feedback on the present thesis. It has been an enormous privilege to present to and learn from such accomplished individuals.

I would also like to acknowledge Dr. David Sinclair for kindly allowing me to use his photo in demonstrating the study procedure in this thesis and manuscript. I would particularly like to thank all the neurosurgeons, senior and junior residents along with the medical students who participated in this study.

Furthermore, thank you to Dr. Jose Andres Correa who made himself available to me, and consistently makes himself available to McGill graduate students in general, as an excellent resource for statistical analysis feedback and suggestions.

I would also like to extend my utmost gratitude to the Franco Di Giovanni Foundation, the Fonds de Recherche du Québec (FRQ) – Santé, the Montreal Neurological Institute and Hospital, the Royal College of Physicians and Surgeons of Canada, and the Mitacs-Own the Podium collaboration, all of whom have supported the Neurosurgical and Artificial Intelligence Learning Centre and its personnel for many years. Thank you to the National Research Council of Canada, who provided us with a prototype of the surgical simulator that was used in this study and for their assistance in developing the scenarios used in this study. I 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. I would also like to thank McGill for the Fellowship of Excellence program that personally supported me through this project.

Finally, I owe my utmost gratitude to my parents, without whose support this thesis would not have been possible. Their perseverance in the face of obstacles and compassion for others have served as archetypes that I have sought to model my life after. They provided a shoulder to lean on during the uncertain period through which this thesis was written, which entirely coincided with the COVID-19 Pandemic. Thank you.

PREFACE AND AUTHOR CONTRIBUTIONS

This thesis is founded on a manuscript that is being submitted to the peer-reviewed Computers in Biology and Medicine journal. The data analyzed in this study was collected previously by the Neurosurgical Simulation and Artificial Intelligence Learning Centre.

The candidate led the study throughout its entirety including data preprocessing and processing, modelling via machine learning, statistical analysis, interpreting the results, generating tables and figures, and writing the manuscript.

Drs. Alexander Winkler-Schwartz, Khalid Bajunaid, Abdulrahman Sabbagh, and Penny Werthner were instrumental in study planning, data collection, and running participants through the trial.

Dr. Sommer Christie removed artefacts from data, provided data exports, provided expertise on electroencephalography and its relation to expertise, and thoroughly edited multiple revisions of the manuscript.

Dr. Recai Yilmaz provided his expertise and guidance on artificial intelligence, machine learning, and model interpretability, and thoroughly edited multiple revisions of the manuscript.

Dr. Rolando Del Maestro supervised the study from beginning to end and was instrumental in its planning and direction. Prior to the study, he was also involved in developing and improving the virtual reality simulator and developing surgical simulation sequences. He also provided continuous guidance for the entire project and thoroughly edited revisions of the manuscript.

ABBREVIATIONS

AI: artificial intelligence

ANN: artificial neural network

AUROC: area under the receiver operating curve

CFC: cross frequency coupling

CONSORT-AI: consolidated standards of reporting trials involving artificial intelligence

CV: computer vision

CWP: connection weight product (Olden's Method)

DL: deep learning

EEG: electroencephalography

ERP: event-related potential

FFT: fast Fourier transform

GIGO: garbage in, garbage out

GRS: global rating scale

HIPAA: Health Insurance Portability and Accountability Act

KNN: K nearest neighbors

LDA: linear discriminant analysis

LIME: local interpretable model-agnostic explanations

LOOCV: leave-one-out cross validation

LR: logistic regression

LSTM: long-short term memory

MAE: mean absolute error

ML: machine learning

MLASE: machine learning to assess surgical experience

MSE: mean squared error

NB: naïve Bayes

NLP: natural language processing

OCASE-T: objective computer-aided skill evaluation - technical

OSATS: objective structured assessment of technical skills

PAC: phase-amplitude coupling

PGY: post-graduate year

ReLU: rectified linear unit

RF: random forest

RL: reinforcement learning

ROC: receiver operating curve

SGD: stochastic gradient descent

SHAP: Shapley additive explanations

SL: supervised learning

SVM: support vector machine

Tanh: hyperbolic tan

TBR: theta/beta ratio

UL: unsupervised learning

VR: virtual reality

BACKGROUND

Surgical Education

History of Surgical Education

Prior to the 19th century, surgery was largely carried out by barber-surgeons and consisted primarily of bloodletting and debridement.[1] It was not until the advent of the 19th and 20th centuries, with the rise of anaesthesia and antisepsis, that surgery in its modern scientific form developed. With this increased complexity arose the need for more stringent surgical education through apprenticeships, which would start at puberty and last 5-7 years.[2] In this way, students learned to conduct surgery by observing surgery in clinical settings and then replicating the procedure. Since no formal curricula or guidelines were developed, students varied greatly in the quality, length of training, and setting of their education, and no equitable system was in place to dictate who to train.[2]

In seeking to rectify the lack of basic clinical sciences in American surgical education relative to his native German system, and by embracing Sir William Osler's concept of clinical clerkships, William Halsted at Johns Hopkins spearheaded the move to a graded responsibility system. The Halstedian model, also called the "see one, do one, teach one" model, systematically assigned increasing responsibility to trainees while not guaranteeing that students will advance to the next stage.[1] This model, however, tended to concentrate on students with a stronger aptitude for surgical skills and at times only 1 of 8 students would eventually become a staff surgeon.[1] In 1939, Edward Churchill at the Massachusetts General Hospital proposed an alternative graduated

teaching system which called for the simultaneous training of multiple trainees rather than only a few individuals.[1] This model is still the primary method of training surgeons in the modern day.[2]

The dawn of the 21st century began to unravel some of the limitations of this training system. The discovery of novel diseases and surgical treatments, development of innovative surgical technologies, and the increasing trends of litigation and patient-centred care and ensuing paperwork[1] have necessitated that surgical residents learn more in less time. Furthermore, noting the danger of overworked trainees on patient safety, the United States introduced a weekly limit of 80 hours for medical residents and limited each shift to a maximum of 30 hours in 2003.[3] Other such limits on residency training time were instituted in Europe in 1993, limiting work to an average of 48 hours, and more recently in Canada, where the National Steering Committee on Resident Duty Hours offered guidelines in 2013 that included a limit on the maximum working hours of 60 – 90 hours and a 24 – 26 hour limit on shift time.[3] While this trend of introducing workplace limits on resident working hours has been generally regarded as praiseworthy in light of its benefits on mental health and burnout rate reduction in residents, it has also been criticized as further decreasing training time for surgeons.[4] This is particularly troubling given the comparatively short time span of a modern residency program relative to a practicing surgeon's career, which in turn is problematic given the modern pace of advancement in surgical techniques and technology.[5]

To address gaps in surgical technical skill and ensure the incorporation of the ever-increasing body of surgical knowledge into practice, surgical organizations offer post-graduate educational programs which render certificates of completion.[1] Also, advancements in technology has permitted advancements in surgical education tools. Surgical simulation, for example, has been

made more accessible due to tissue engineering,[6] additive manufacturing,[7] and virtual reality[8] technology. These technologies have resulted in some surgical training centres supplementing the traditional training model with surgical simulation.[4,9]

Virtual Reality Simulation

There has been a shift in recent years towards more simulated training in supplementing traditional surgical education,[10] driven by increased operating room costs and the desire to limit patient risk.[4] As a result, simulation is not just a technique for trainees but is likewise applicable to experts wishing to acquire new techniques.[11] Also, simulation is particularly important at times when less clinical interaction is possible, and has therefore gained interest during the COVID-19 pandemic.[12]

Simulators are usually classified as high- or low-fidelity, depending on a subjective appraisal of their realism and range of features.[13] Surgical simulation has consisted mostly of the use of lower-fidelity bench-top and 3D-printed models, as well as high-fidelity cadaveric, animal, virtual reality, and robotic models.[4] The non-reusability of animal model systems, the expenditure of animals in this cause, and concern over potential infection of trainees by animal-borne diseases such as bovine encephalopathy, have made virtual reality (VR) simulation an especially fashionable option in the high-fidelity sphere.[14]

Virtual reality simulation provides a safe training environment and allows for self-guided learning using a computer-generated graphical training procedure.[15] There is considerable evidence that training on a virtual reality simulator improves surgical practice.[15] For example, VR simulation has been shown to decrease operative time while improving surgical performance.[16]

Performance-related measures have also been shown to improve with real operating room surgical expertise.[17] Furthermore, there is the future potential for patient-specific VR simulation, wherein a training procedure is customized to a planned case by way of imaging. To this end, a patient is imaged using magnetic resonance imaging (MRI) or a similar technology, and this data is incorporated into the training program. Such an option would supplement a surgeon's mental preparation for a case with physical preparation and may further provide an avenue for patient communication concerning a surgical procedure.[18]

As with the aforementioned methods of surgical simulation, several limitations exist in the virtual reality model. For instance, currently, there is a large cost associated with developing the virtual reality simulator itself and given the high spatial acuity and the need for haptic feedback, such systems are usually dependent on more expensive computational hardware. This translates into a higher cost of acquisition.[19] There is also a large cost associated with developing and validating training procedures which may limit the range of procedures available for practice.[20] Furthermore, not all simulators include a full haptic feedback system, which provides real-time tactical force feedback based on the trainee's actions.[19] This can be problematic given the dependency of surgeons on haptic feedback in guiding their procedures, and the fact that the presence of haptic feedback is thought to reduce surgical error and as a result increase patient safety.[21] Furthermore, some VR simulators lack realism in their graphical output.[19]

As with other simulators, VR simulators are assessed on reliability and validity. Reliability measures the extent to which each simulation is reproducible, and consists of three types: inter-observer, intra-observer, and test-retest reliability.[14] Inter-observer reliability measures reproducibility between trainees, whereas intra-observer reliability measures the reproducibility of

a single trainee's results. Test-retest reliability measures the correlation between multiple instances of a trainee's performance. In contrast, validity measures the simulator's ability to teach what it purports to and is composed of three primary types: face, content, and construct validity.[14] Face validity measures how realistic the simulation is. Content validity measures the extent to which it simulates all aspects of a given surgical procedure. Finally, construct validity measures its ability to classify expertise.[14] A common theme in surgical simulation is a lack of reproducible validation studies of simulators, although work is being done to address this.[14]

The NeuroVR™ simulator (CAE, Montreal, Canada) is a notable counter-example to this last limitation and is a VR platform for neurosurgical simulation that has been relatively well validated.[22] It was developed during a long-term collaboration between the Neurosurgical Simulation and Artificial Intelligence Learning Centre and the National Research Council of Canada starting in 2008, and in consultation with a network of experts on neurosurgery and neurosurgical education. Equipped with endoscopic and stereoscopic views as well as bimanual surgical tools, the NeuroVR™ has been marketed as "the world's most advanced simulator for neurosurgery."[23] It further features a 1280x1024 pixel surgical view and auditory, visual, and full haptic feedback systems.[24] Additionally, the NeuroVR™ allows for the recording and exportation of psychomotor performance metrics, which have in turn been used to validate the device by objectively characterizing the expertise of trainees.[22,25] See **Figure 1** for an illustration of a simulated brain tumor resection on the NeuroVR™ platform.

Objective Assessment of Expertise

To address the aforementioned gaps, a recent trend in surgical education has been a renewed focus on surgical outcomes rather than completion of curricula,[26] coupled with the move

towards a competency-based, as opposed to a time-based, training system.[27] To adequately begin to assess competence, a working definition of competence as it is applied in surgery was required, given that the expectations that patients had of surgeons differ drastically from what is expected of other professions.[5] As such, surgical competence has been defined as the capacity to apply knowledge and technical skills to new and familiar tasks[5] and is founded on five pillars: problem-solving, effective communication, teamwork, ethical integrity, and the technical ability to safely operate.[5,27] Although competence, as opposed to expertise or mastery, may otherwise seem to be a rather underwhelming bar to meet for surgeons aiming to treat life-threatening, this stringent definition ensures that surgeons are able to meet their patients' expectations.[27]

It has been shown that subpar technical skills are associated with worse patient outcomes.[28] Thus, several global rating scales (GRS) have been developed to address the technical pillar of competency.[29] The gold-standard of these systems is the objective structured assessment of technical skills (OSATS-GRS),[30] which consists of both a surgery-specific checklist and a global rating five-point Likert scale.[30,31] These objective assessments have been shown to reliably distinguish between different levels of surgical experience,[31] but are costly to implement, particularly when done as an examination outside of normal operating room duties.[30,31]

This limitation has given rise to research into objective computer-aided skill evaluation (OCASE) of surgical technical skill (OCASE-T).[29] In this approach, surgeons are monitored using cameras,[30,32] physiological sensors,[33] tool motion sensors,[34] tissue manipulation sensors,[35] and electroencephalography (EEG).[36,37] This data is then passed through an automated classification system that is trained on surgical learners of various experience levels in order to automatically assess their expertise.[34] Such approaches have been used to achieve accuracies upward of 98% in

classifying trainees between expert, intermediate, or novice levels.[29] In order for the system to have the desired effect of improving surgical skill, as opposed to only assessing it, it is necessary to provide feedback to the learner concerning the classification and reasons for such a decision.[29]

Electroencephalography

Electroencephalography (EEG) is a technique used to measure the electrical activity of cortical neurons spontaneously or during an event.[38] Differential amplifiers are used to amplify small voltage potential differences to render them detectable.[39] Although EEG does not have the granular capacity to detect single neurons, it can detect the simultaneous summation of post-synaptic potentials of neuron clusters (as opposed to action potentials, which do not summate and as such are thought to only slightly contribute to the total EEG signal).[40] To this end, electrodes are often placed on the scalp and EEG is able to monitor the local electrical activity surrounding each electrode.[39]

Much processing goes into transforming raw EEG signal into meaningful insight. For instance, raw EEG signal is captured in the time domain and consists of waves of various frequencies, amplitudes, and shapes. To permit more meaningful analysis, it is transformed into the frequency-domain using the Fast Fourier Transform (FFT) algorithm. Moreover, raw EEG signal is ordinarily contaminated with irrelevant electrical signals from the body, such as those from eye movements,[41] eyeblinks,[41] muscular activity,[42] and cardiac rhythm.[43] These are each removed using specialized software. The raw EEG is also contaminated with electrical signals from the environment, such as those associated with power outlets at the 50 or 60 Hz level.[44] These artefacts are removed by passing the data through a Notch filter which filters out the particular frequencies.[45] The data is then sorted into frequency bands that have been derived empirically and each associated with meaningful psychological traits (**Table 1**). The full EEG processing workflow employed in this study is illustrated in **Figure 2**.

EEG is commonly used as a clinical test, in diverse applications such as diagnosing diseases, monitoring general seizure activity, and monitoring sleep stages and quality.[46] The primary advantages of EEG include its high sampling rate and relatively low cost. In addition, EEG has a notable advantage over other biosensors in that it is predictive of motor action,[47] which could potentially be utilized to prevent fatal errors in surgery. However, its limitations include poor spatial acuity and difficulty in monitoring subcortical structures.[48]

Band significances

Bandpass filters are commonly used to focus in on a particular range of frequencies for analysis.[40] EEG data, divided into various frequency bands, are associated with various cognitive processes. For instance, certain bands are associated with learning and memory, whereas others are associated with tranquillity.[49] The most commonly studied frequency bands include alpha, beta, theta, delta, and sigma,[40] although the exact frequency ranges and significances for each of these bands is not always agreed upon and may differ between individuals.[40]

Specifically, this study included delta (2-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), and beta (13-30 Hz) bands, as well as the sensorimotor rhythm (SMR, 12-15 Hz). Low (4-6 Hz) and high theta bands (6-8 Hz) and beta 1 (15-18 Hz), 2 (19-22 Hz), and 3 (23-36 Hz) bands were also included. Furthermore, a composite metric known as the theta/beta ratio (TBR) that has been associated with technical expertise was calculated as the square of the theta band divided by the square of the 13-21 Hz beta band.[50] **Table 1** illustrates these EEG frequency bands and their respective significances based on a survey of the literature.[49,51,52,53] A strategic positioning of electrodes around the neural areas associated with the trait assessed has been shown to increase the utility of training procedures based on EEG signals. For example, electrodes focused around

the front-parietal neural network, important in attention, may be overrepresented in a training procedure for attention-deficit/hyperactivity disorder (ADHD).[54]

Furthermore, in addition to individual frequency bands, the relationship and interdependence between different frequency bands has been implicated in neural function in a process known as cross-frequency coupling (CFC). For instance, in a particular type of CFC known as phase-amplitude coupling (PAC), the phase of a lower frequency bands affects the power of a higher frequency bands, synchronizing the amplitude of the faster bands with the phase of slower bands.[55] PAC has been detected in various brain regions, such as the basal ganglia and hippocampus, and associated with various forms of cognitive functioning, including attention, visual perception, and memory processing.[56] These brain regions and functions play a role in surgical performance, and thus it may be important to note the potential effects of CFC in a full EEG analysis of surgical expertise.

Temporal, Spatial, and Spectral Analyses

A complete EEG analysis usually includes a combination of temporal, spatial, and spectral analysis and typically includes about 21 electrodes, with less electrodes resulting in lower spatial acuity.

Temporal analysis examines EEG band variance with time. As a timeseries signal, EEG naturally lends itself to this sort of analysis. The most common type of temporal EEG analysis involves event-related potentials (ERPs), which attempt to quantify the psychological response to a stimulus to capture sensory and cognitive processes.[57] In this analysis, EEG waves are synchronized with physical events in order to objectively map neural response across individuals.

Human ERPs are divided into sensory, which are exhibited immediately following a stimulus, and cognitive, which occur afterwards and are associated with the subject's perception of the stimulus.[57] The shape, frequency, and type of waves are all considered in characterizing response. The main limitation of this analysis is the technical requirement of synchronizing EEG recordings to events via time stamps, which has historically been arduous but has been facilitated in recent years with more advanced software technology. Interestingly, features extracted from this sort of analysis have been shown to be more important than those from the two other analytical methods in a motor imagery classification task.[58] See **Figure 3** for sample temporal EEG data.

Spatial analysis arises from the use of multiple EEG electrodes and encompasses the impact of the relative location of each electrode to the EEG signal. To this end, a system known as the 10-20 System has been developed to standardize the placement of electrodes on subjects' heads.[38] This system relies on four skull landmarks: the nasion (Nz) at the bridge of the nose; the inion (Iz), the bony extrusion at the back of the head; and the two pre-auricular points beside each ear (LPA and RPA).[59] See **Figure 4** for an illustration of the 10-20 system. The advantage of this analysis is the ability to better localize EEG signals to particular cortical areas. However, this sort of analysis has the limitation of requiring multiple electrodes, which results in increased setup time, costs, and general inconvenience due to potential wiring. These limitations are usually well worth the effort and it is very common to use multiple electrodes in both academic and therapeutic practice to enable this form of analysis.

Spectral analysis considers the power spectral density of the EEG signal as a function of frequency.[60] This sort of analysis may consist of averaging out temporal differences in EEG signal across frequency band in a given time span so as to only consider overall trends. Spectral analysis

arose out of the relative lack of computational resources in the 20th century but continues to be the most popular form of EEG analysis in use today. Although it may depend on less data than the aforementioned types of analysis, spectral analysis provides a big picture overview and lends itself well to neurofeedback.[61] Nonetheless, it is important to emphasize the importance of all of these forms of analysis to a more complete understanding of neural activity.

Neurofeedback

Neurofeedback is a technique used to increase a subject's self-awareness of their neuronal activity in order to facilitate signal change, all in an effort to promote optimal context-dependent brain functioning, such as relaxation or concentration. Neurofeedback is usually done with EEG data, however more modern advances have experimented with the use of magnetic resonance imaging (MRI) to this end.[62] Rooted in operant conditioning, neurofeedback calls for EEG patterns to be recorded, analysed and fed-back live in the form of visual and/or audio cues to the participant.[63] This process promotes the learning of correlations between the participant's internal mental state and recorded neural signal, facilitating voluntary control over the precipitation and maintenance of particular states[62] and has been linked to underlying physiological transformations, such as increases in white and grey matter in the brain.[64]

There are several major neurofeedback training protocols, each characterized by the frequency band that they target, and include alpha, beta, alpha/theta, delta, theta, and SMR. Alpha training improves relaxation and is usually used for the treatment of pain, stress, or anxiety.[62] Beta training is usually carried out with the goal of increasing focus, attention, and cognitive processing.[62] Next, alpha/theta is one of the primary neurofeedback types to control stress and is usually done using auditory feedback with closed eyes with the goal of decreasing alpha while

increasing theta. The delta protocol strives to increase comfort, reduce pain, and promote sleep. Also, the theta training promotes relaxation by increasing alpha and theta waves while decreasing high beta, which is associated with concentrated thinking[62]. Finally, SMR training is thought to improve concentration by decreasing theta waves, which are associated with wandering, and increasing SMR, which is associated with calmness and focus[65].

Although initially established as a treatment for certain neurological disorders such as ADHD, anxiety, and Post-Traumatic Stress Disorder,[66] much research has investigated the use of neurofeedback in improving technical expertise such as in athletics[61,67] and pilot[68] simulation training. This is known as peak performance training and is utilized by artists, athletes, musicians, and executives but under-utilized by surgeons, which the author feels is unfortunate, given the high stakes at play in surgery relative to the aforementioned fields. However, this is to be expected since surgical training curricula have traditionally aimed to develop competence in trainees, and although competence is set at a high-standard in surgery, it is not at the level of excellence that is expected of elite athletes or fighter pilots.[69] Surgical simulation training has enabled the possibility of supplementing surgical practice with neurofeedback in order to train surgeons more holistically. Furthermore, this transition may be facilitated by the proliferation of neurofeedback training tools, with many options now being available commercially.[70] For instance, the Muse headband offers a relatively inexpensive option that functions through auditory feedback, playing a particular genre of music corresponding to whether a subject is within the optimal range or not.[70]

Neurofeedback protocols are usually initiated via an EEG to establish baseline levels. Comparing these baseline recordings with a database replete of other individuals with similar demographical characteristics, it is possible to identify bandwidths that are under- or over-

stimulated. A neurofeedback training protocol is then initiated to increase or reduce these bandwidths respectively. However, it is unclear as to the order of training in the case of multiple sub-optimal bands. Thus, the elucidation of the relative contribution of each EEG frequency band to the final training goal may facilitate the integration of neurofeedback in a training protocol.[62] In the present study, a greater understanding of the relative contribution of each EEG frequency band to surgical expertise, may lead to the development of an evidence-based system of neurofeedback training for surgical training during virtual reality simulation.[62] To this end, large EEG datasets can be analysed by artificial intelligence (AI) to deconstruct the frequency bands important in expert surgical performance.[61] A potential neurofeedback training protocol making use of the results of this study is illustrated in **Figure 5**.

Artificial Intelligence

Artificial intelligence (AI) is the use of computers to mimic human behavior. AI traces its computational roots to the 1950s, when Alan Turing, in an attempt to answer the question of whether a machine can think, proposed what became known as his eponymous Turing Test.[71] A machine passes the test and attains true intelligence if it is able to deceive a person into believing that it is human from behind a veil.[72] This test has inspired generations of artificial intelligence practitioners and has led to ground-breaking developments, particularly in language understanding and generation.[73]

Since then, AI has gone through several "AI winters" where it has seen periods of relative inadvertence. However, recent advances in computing capacity, data storage, and computer analytics have elevated AI capabilities and have once again brought this discipline to the corporate and commercial spotlight.[74] With the ability to automate specialized work, reducing costs and accelerating growth in the process, AI has the potential to revolutionize many aspects of human life.[74] Prevalent branches of AI include natural language processing (NLP), robotics, and computer vison (CV), which have respectively spearheaded advancements in chatbots, robot-assisted surgery, and autonomous vehicles. See **Figure 6** for an illustration of the relationship between AI and selected subbranches.

In healthcare, artificial intelligence is projected to assist in facilitating physician-patient interactions by supplementing triage,[75] diagnosis,[76] prognosis and survivorship prediction,[77] and medical or surgical education.[78] Smart scribes are also expected to utilize voice recognition to automate clinical documentation, mitigating legal challenges and liberating physicians for full engagement with patients while decreasing clinical burnout.[79,80]

Machine Learning

The rise of sensors, data pipelines, larger data storage capacities, and more advanced data analytical tools has resulted in a move away from explicitly programmed expert systems to data-based learning (machine learning, ML). This new field has reinvigorated artificial intelligence research and it has resulted in systems which exhibit more generalizability than previously possible, thus addressing a key limitation of AI research. Notably, biomedical data collection and storage in the form of electronic medical records and associated medical test and imaging results,[77] genetic tests,[81] and internet of things (IoT) biosensors[82] have made the applications of this field particularly exciting in healthcare. See **Figure 6** for an illustration of the relationship of ML with other branches of artificial intelligence.

Machine learning is generally divided into four subcategories based on how algorithms are trained: reinforcement learning (RL), supervised learning (SL), unsupervised learning (UL), and semi-supervised learning. Reinforcement learning is the learning of behaviour in order to maximize a quantitative reward signal and minimize a punishment signal. In this system, an agent is fit with sensors and allowed to interact with its environment and learns to successfully navigate the challenges in its environment by seeking to maximize its reward function.[83] In supervised machine learning, an algorithm is trained by being exposed to multiple examples of a series of observations associated with human-labeled classes. In contrast, unsupervised learning aims to address this costly limitation of human-labeled classes by learning patterns from unlabelled data, usually through cluster analysis.[84] Semi-supervised learning combines principles from the two aforementioned approaches by training algorithms using a seed amount of labeled data and a relatively large amount of unlabeled data.[84]

Nonetheless, supervised learning continues to be the most popular subbranch of machine learning.[85] Supervised learning algorithms aim to conduct either classification or regression.[76] Classification is the goal of assigning a discrete category to a particular observation, whereas regression is the estimation of a continuous dependent variable based on an observation.[76] To these ends, several types of ML models exist but the seven most common types in healthcare in order of popularity are support vector machines (SVMs), artificial neural networks (ANNs), logistic regression (LR), linear discriminant analysis (LDA), Random Forest (RF), Naïve Bayes (NB), and K-Nearest Neighbors (KNN).[86] Each model utilizes a slightly different mathematical analytical method to arrive at classifications or regression results.[87] The particular model type that is chosen for a task is empirically determined and dependent on the nature of the data present and the amount of computing resources available during training and deployment.[88]

Linear SVMs work on linearly separable data by selecting a hyperplane that maximizes the minimum margin, which is the minimum distance between a particular class label and the classification cut-off, whereas kernel SVMs work on non-linear data using a predefined kernel function to map data to higher dimensional representations.[89] Artificial neural networks are inspired by their biological counterparts and are discussed extensively in the following section. Logistic regression is used for binary classification and is based on fitting the input data on a sigmoid function. Since it predicts the probability of being in the target class, a cut-off value of 0.5 is usually assigned in order to assign discrete classes.[89] Linear discriminant analysis reduces the dimensionality of the problem by creating a new dimension and projecting the data on that dimension in such a way so as to maximize the separation of the categories. It does so by seeking to maximize the distance between the sample means while reducing the variability in each category.[89] Next, Random Forests work by generating a myriad of decision trees, each based on a

random subsample of the total training data, and then tallying the number of classifications of each class from these trees. The class with the most tallies is designated as the model output.[89,90] Naïve Bayes models generalize the Bayes rule to an arbitrary number of input variables and returns the probability of an output value given the input values. They have subtypes that are specialized to different data distributions, including Bernoulli, multinomial, and Gaussian.[89] Finally, K Nearest Neighbors works by assigning a class depending on the weighted average of the closest K examples in the sample. It has the notable disadvantage of being relatively slow at run-time since it does not train over data in advance, and thus may be impractical in applications where real-time feedback is necessary.[89]

In classification tasks, supervised learning algorithms first go through a training phase wherein they are fed labeled training data, allowing them to adjust their algorithmic parameters in such a way so as to progressively improve their ability at classifying examples. In the case of neural networks, this is done by identifying this step as a minimization problem and defining a loss function which quantifies the distance between the model's current predictions and the true predictions.[91] The loss function is minimized, leading to optimal performance given the current training parameters. Afterwards, a model is tested on previously withheld testing data and generally outputs a list of probabilities associated with each predefined possible class. The class associated with the highest output probability is then taken as the model's prediction for that example.[88] A notable exception exists in binary classification, where only one probability corresponding to a target class is generally output, since the opposite probability is mutually exclusive. The workflow for regression tasks is very similar, with only slight differences in the type of loss function used and the class labels.

As with all artificial intelligence constructs, machine learning currently has a number of major limitations. For example, it is estimated that over 80% of enterprise data is unstructured, meaning that it cannot be represented in tabular form, which poses a challenge given the relative weakness of machine learning in analyzing this sort of data.[92] Notably, this includes audio, emails, and physician notes.[77] Furthermore, data silos make accruing enough data to train larger models difficult, although a recent trend of democratizing data[93] is promising and a field known as federated learning[94] is making it possible to train on data held in remote devices. Furthermore, maintaining confidentiality is integral when collecting training data. This is particularly concerning as malicious techniques have been recently developed that are capable of recreating training data from final models,[95] and is even more pressing in healthcare, where one must stay compliant with Health Insurance Portability and Accountability Act (HIPAA) regulations and maintain the trust of patients.[96] Also, biases in the dataset utilized for training may result in biased predictions, in a phenomenon known as garbage in, garbage out (GIGO). This has been particularly concerning in healthcare, as such biases could lead to life-threatening recommendations.[97] Finally, as a fast-moving field, legislation and ethical research has struggled to keep up with the pace of advancements in machine learning.[98]

Artificial Neural Networks

Artificial neural networks are a special type of machine learning structure inspired by the biological structure of the human brain. Nodes, called neurons, are connected by synapses, known as weights. These weights are often stochastically initialized, which contributes to the fact that multiple identical training trials may yield different final neural networks.[99]

In a traditional sequential neural network, the first layer in a neural network is known as an input layer and corresponds to the data being input, with one neuron being assigned to each input metric. The last layer in a neural network is known as the output layer and corresponds to the type of output, which may vary widely from a probability to a continuous variable to an image. In between these two layers is a number of hidden layers, each with a number of neurons. More modern neural networks, such as LSTMs[100] or Transformers,[101] may exhibit more complex structural characteristics in that they are not strictly sequential.

As metrics propagate through a neural network, they are transformed appropriately. The internal structure of each neuron may vary, but generally, each neuron takes a number of inputs and multiplies them by their respective learned weights. Then, a neuron summates all of its inputs to one value, adds to it a learned bias value, and then applies an activation function to it.[99] The role of activation functions is to introduce nonlinearity into the system, thus increasing the potential complexity of the relationship between the input metrics and the output results, while also controlling for a class of quantitative problems associated with propagation known as vanishing or exploding gradient problems.[102] Several activation functions have been established in the literature, although the rectified linear unit (ReLU), sigmoid, and the hyperbolic tan (Tanh) functions are very common.[103] The role of the bias term, on the other hand, is to mathematically shift the activation function left or right, effectively increasing the potential complexity of the relationships mapped. For example, whereas changing the weight to multiply the summed values by may not affect a hypothetical final sum of 0, adding a bias would enable this final sum to be mappable to any other value.

When the metrics reach the output layer, a loss function which is defined *a priori* is applied. This loss function quantifies the distance between the current prediction and the true value. A popular loss function for classification is binary cross entropy, whereas a popular loss function for regression may be mean squared error (MSE).[104] Then, the backpropagation technique, derived from the Chain Rule of calculus, is used in conjunction with a technique known as stochastic gradient descent (SGD) to smoothly adjust the weights between each neuron so as to decrease the current loss. This is done progressively, with multiple passes (epochs) of the training data going through the neural network in batches.[105]

Various metrics exist to assess the success of a neural network. For example, accuracy, the area under the received operating curve (AUROC), F-measure, and mean absolute error (MAE), are all used to this end.[77] However, certain metrics are thought to be better than others in certain circumstances. For instance, accuracy may not be a good measure of success given extreme class imbalances, as the correct prediction of only a few overrepresented classes would then enable the achievement of a relatively high accuracy at the expense of the underrepresented classes. In this case, a metric which takes this sort of bias into account, such as AUROC, is advised.[106]

Neural networks are characterized by their inherent flexibility and the relatively large number of possible parameters to adjust.[107] These parameters, not learned during the training process and instead set manually by the machine learning practitioner, are known as hyperparameters. Usually, a procedure is set whereby hyperparameters are tuned by trying various combinations iteratively while seeking to minimize each model's loss. Furthermore, they are also characterized by their relatively large number of potential trainable parameters, with some

networks being trained with upwards of a trillion parameters.[108] **Figure 7** illustrates an example neural network.

Deep Learning

The latest wave of interest in artificial intelligence has been a result of the success of deeper neural networks.[109] To distinguish this sort of artificial intelligence from its more basic machine learning counterpart, the term deep learning was coined.[109] Although there exists controversy over what exactly constitutes a deep model, a rough guideline posits that it is a neural network with three or more hidden layers.[110] The relationship between deep learning and artificial intelligence as well as machine learning is illustrated in **Figure 6**.

Traditional machine learning workflows include a time-consuming and unscalable step known as feature engineering. Feature engineering is the manual selection and curation of input metrics for a model and constitutes up to 80% of total modelling time.[77] Deep learning models, in contrast, are able to circumvent the need for this step by extracting metrics from less processed data.[109] However, this circumvention contributes to the increased difficulty with which these models may be understood due to the selection of non-human interpretable abstractions and may introduce elements of bias that are carefully avoided with manual feature selection.[77]

Deeper networks are usually associated with increased predictive power and accuracy. However, they are also associated with increased complexity and difficulty in interpretability due to their potential abstract representations of the input metrics. This is particularly challenging in healthcare, where a dilemma may exist between using a more powerful deep learning model, and the need to be able to justify the decision made. However, there are ways to control for this, such

as partial feature engineering where a DL practitioner designs several metrics but also allows the model flexibility to generate its own.[77] Furthermore, the field of model interpretability seeks to address this issue.

Model Interpretability

Traditionally, a key limitation of more complex machine and deep learning models has been the inability to interpret their decision-making process.[111] As a result, many machine learning models, especially those associated with deep learning, were classified as black boxes. This has unique implications in healthcare, where decisions can have grave consequences and as such patients often demand that individual healthcare providers be able to take responsibility for them.[112] Trust in these systems by healthcare professionals may be undermined by the potential for bias, the lack of domain technical expertise, and lack of knowledge pertaining to machine learning principles.[112,113] Historically this issue was circumvented by limiting oneself to more simple models, which were relatively simple to interpret.[111] However, a dilemma has arisen where a balance must be struck between the use of more complex models, which were often more powerful yet less interpretable, and the ability to explain decisions to patients.[112]

To address this issue, statisticians and computer scientists have developed tools to allow for the interpretation of the decision-making process of more complicated machine learning models in a trend known as interpretable ML.[112] A model in healthcare is classified as interpretable if it is possible to evaluate the reasoning behind its output, thus empowering healthcare providers with the ability to accept or reject its recommendations on a more informed basis.[114]

Interpretability methods can be classified in several ways. A common approach is to classify them based on which step they are employed within the training process. Pre-model interpretability methods are applied before a model is built and include techniques such as descriptive statistics and clustering methods.[112] By understanding the characteristics of the data at hand, and the model's inherent dependence on said data, one can extrapolate the potential biases the model would pick up on. It is then possible to alter the properties of the data at hand through various techniques, such as subsampling, to arrive at a less biased model. In an alternative approach to pre-model interpretability, one may train a model that is inherently simple to interpret due to its structural properties, such as a decision tree, or limit the potential complexity of a model by introducing interpretability constraints such as sparsity.[112] Sparsity refers to a constraint on the maximum number of input metrics that are used in the final classification, thus only allowing for the learning of relatively less complex relationships. In contrast, post-hoc, or post-model, methods facilitate the interpretation of a model after its training. Various techniques exist, such as the training of an intrinsically interpretable model to mimic the predictions of the complex model and then interpreting this relatively more interpretable model. If the approximation is sufficiently similar to the original model, it will preserve its statistical properties, thus making this a valid method of interpretation.[115] Alternatively, one may examine the learned model weights.[115]

Interpretability methods can be classified as model-specific or model-agnostic. Model-specific methods derive explanations that are dependent on a model's internal structure. An example of this type of method was proposed by Julian Olden in 2004 and is known as Olden's Method or the Connection Weight Product method.[116] It is specific to neural networks.[116] In this method, each metric of a neural network is traced through the associated weights and neurons to determine its contribution in the output layer. See **Equation 1** and the associated explanation in

the Model Interpretability Case Study section of the Discussion for a more detailed account of this method, and **Figure 8** for a sample interpretation. In contrast, some have argued for the use of model-agnostic methods due to their relative versatility.[117] These are generally post-hoc model interpretation methods that operate on the principle that a feature's importance can be determined by measuring the change in a model's accuracy after the feature is prodded. More important metrics will have larger effects on a model's accuracy with only slight modifications.[115]

Interpretability methods have also been classified based on their scope of interpretability. Global interpretability methods aim to explain a model's outputs over a population in general terms. In contrast, local interpretability explains how a model comes to any one particular output.[118] Both are important and have their applications in healthcare.[112] For instance, one might elect to run a global interpretability method upon initial modelling to ensure that no striking biases exist and that the model does not operate in counter-intuitive terms. Then, upon the use of a model in clinical practice, one may interpret the decision-making process of the model on a one-by-one basis before presenting their results to the patient. This way, these interpretability methods serve to provide layers of safety for the end user.

It is now, at least in the case of smaller models, possible to determine the relative importance of each input metric to the model's final classification, in an explanation measure called feature importance.[111] When combined with a carefully curated set of meaningful metrics, this represents an opportunity to better understand a model's decision-making process.[119] Two especially well-known approaches exist in this regard: Local Interpretable Model-agnostic Explanations (LIME)[120] and Shapley Additive Explanations (SHAP).[111] The first type depends on the production of a model that can approximate a model's output for only one particular instance.

It works by perturbing the input data and assessing the resulting changes in the model output.[120] In contrast, the Shapley explanations method treats the features of a machine learning problem as players in a coalitional game from game theory. A specific value, called a Shapley value, is assigned to each feature, and demonstrates its contribution to the result. However, while Shapley values produce high quality explanations, their exact computation can be implemented efficiently only in certain (decision tree-based) model types, whereas they must be approximated when using other models.[111] See **Equation 2** for the mathematical formulation of the Shapley values and **Figure 9** for a sample model interpretation using Shapley values.

THESIS HYPOTHESIS AND SPECIFIC OBJECTIVES

The challenges of 21st century medicine, including the advent of novel surgical treatments and the discovery of novel diseases and advancements in surgical technologies, against a backdrop of increasing litigation and ensuing paperwork, has called for more learning in less time on the part of surgical residents.[1] In addressing these challenges, traditional surgical education has been supplemented with surgical simulation.[10] Virtual reality simulation presents several key advantages in this regard, including its propensity towards reusability, the absence of animal sacrifice and related ethical dilemmas, and the lack of risk of potential infections associated with animal or cadaver models.[14]

In order to maximize the efficiency of such a system, and in line with the trend towards competency-based medical education, objective methods of classifying surgical expertise are required.[27] The gold-standard of these systems, known as the objective structured assessment of technical skills (OSATS),[30] consists of both a surgery-specific checklist and a global rating five-point Likert scale.[30,31] This system has had some success but is relatively costly to administer since it requires the independent assessment of surgical performance by multiple surgical personnel.[31] To mitigate this issue, several automated methods are currently under development to track surgical performance in the hopes of objectively quantifying surgical expertise.[29] These methodologies have relied on tracking surgeons via several key technologies, such as cameras,[30,32] physiological sensors,[33] tool motion sensors,[34] tissue manipulation sensors,[35] and electroencephalography (EEG).[36,37]

Relatively less focus has been placed on the potential of electroencephalography (EEG), the recording of neural electrical activity using electrodes, in this regard. This is despite the fact

that EEG has the marked advantage over other biosensors of being predictive of movement,[47] which could potentially be utilized to prevent fatal errors in surgery, whilst being relatively low cost and having a relatively high sampling rate.[40] Furthermore, EEG has the potential to be used to improve technical skills using a technique known as neurofeedback, where EEG frequency bands are fed back to participants in order to increase their psychological self-awareness and facilitate active change in EEG activity.[121]

Although EEG has been used in assessing surgical performance,[122] it has not been applied to classify surgical expertise. A large data analytical task such as this lends itself to applications in artificial intelligence (AI), the mimicking of human behaviour using digital computers. Specifically, a branch of AI known as machine learning (ML) allows for algorithms to be trained using data rather than by programming predefined rules. It has been been applied extensively in neurosurgical care and EEG analysis, where it has been used to facilitate the surgical treatment of epilepsy, brain tumors, Parkinson's disease, and brain injury.[123] Furthermore, our group has already utilized ML methodologies to construct Intelligent Tutoring Systems for surgical education by classifying surgical expertise using tool motion sensors.[34,78,119]

Such complex systems have traditionally been difficult to implement in healthcare settings due to the difficulty with which their classifications are interpreted and the need to ensure objectivity for the highest quality of patient care.[112] However, recent advances in model interpretability methods have enabled exploration of the underlying decision-making systems in complex algorithms.[112,115] One example of this is the Shapley model interpretability method, which aims to assign a relative value to each input metric in a given classification task.[111]

This study addresses multiple questions. Is it possible to accurately distinguish between different levels of surgical expertise in an artificial reality surgical simulation using only EEG data? Which machine learning model type is best equipped to handle such a classification task? What are the differences in EEG frequency band power levels between skilled and less skilled surgeons? What is the relative contributions of each EEG frequency band to a final skilled classification?

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. We thus set several objectives:

1. To compare the most common types of machine learning models in their ability to classify expertise using EEG data
2. To generate a model capable of accurately distinguishing between skilled and less-skilled participants on a virtual neurosurgical simulation
3. To statistically analyze differences in EEG frequency bands between skilled and less-skilled participants
4. To use model interpretability methods to assess the relative contribution of each EEG frequency band to final expertise classification

This pilot study aims to provide insight on the utility of EEG data in surgical expertise classification, with the goal of ascertaining the educational utility of EEG in a modern surgical residency program. Using virtual reality, artificial intelligence, and model interpretability methods, we shed light on the psychological profile of surgical trainees and how this knowledge may be used in training programs. Finally, we build on our results to provide guidelines for how a neurofeedback mechanism for training surgical residents may be established.

STUDY

Utilizing Artificial Intelligence and Electroencephalography to Assess Expertise on a Simulated Neurosurgical Task

Sharif Natheir[a], BMSc, Sommer Christie[a], PhD, Recai Yilmaz[a], MD, Alexander Winkler-Schwartz[a], MD, PhD, Khalid Bajunaid[b], MD, MSc, Abdulrahman J Sabbagh[c,d], MD, Penny Werthner[e], PhD, Rolando Del Maestro[a], MD, PhD

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.

Manuscript submitted to NeuroImage.

Abstract

Virtual reality surgical simulators have facilitated neurosurgical education by providing a safe training environment. Electroencephalography (EEG) has been employed to assess neuroelectric activity during surgical performance. Machine learning (ML) has been widely applied in analyzing EEG data split into frequency bands. Modern model interpretability techniques, such as those utilizing Shapley values, have enabled the analysis of more complex ML models. 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 (post-graduate years 4-6) and the less-skilled group was composed of six junior residents (post-graduate years 1-3) and five medical students. A total of 13 metrics from EEG frequency bands and ratios (e.g., theta, alpha, beta, theta/beta ratio) were generated. Machine learning models were trained using EEG activity to differentiate between skilled and less-skilled groups. The relative importance of each EEG metric was calculated using Shapley values. Seven model types were trained, with the artificial neural network achieving a 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.

KEY WORDS: Electroencephalography, Artificial Intelligence, Machine Learning, Education, Virtual Reality, Neurofeedback

1.0 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.[124] Technical errors in this complex bimanual psychomotor skill include subpial vessel hemorrhage and injury to adjacent normal cortex can result in significant patient morbidity.[28,124] To aid learners in the mastery of this technical skill necessary to safely and efficiently carry out these procedures our group has helped develop[24] and validate virtual reality simulators[125] along with creating complex and realistic virtual reality tumor resection tasks.[126] Virtual reality surgical simulators employed in neurosurgical education provide a safe training environment and allow for self-guided learning.[15] These learning tools are particularly relevant during times when trainees have less clinical interaction such as during the present COVID-19 pandemic.[12]

Electroencephalography (EEG), the use of electrodes to assess neural electrical activity, has been used to continuously assess brain activity during surgical performance.[37] 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.[62,127] Theta frequencies, for example, are associated with learning and memory, whereas alpha frequencies are associated with tranquillity.[49] 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.[61]

EEG has been used in surgery,[36] sport,[61,128,129] and flight simulation training[68] to predict expertise and/or to improve performance using neurofeedback. Christie et al. (2020) utilized EEG neurofeedback to improve ice hockey shooting performance and reported significantly higher rates

of improvement in the treatment group compared to controls.[61] Neurofeedback training is based on the principles of operant conditioning, where EEG is recorded, analysed and fed-back live in the form of visual and/or audio cues to the participant.[63] This feedback enables the participant to progressively learn how their internal mental state correlates with the neural signal, facilitating voluntary control over entry to, and maintenance of, particular states.[62]

Neurofeedback has been shown to increase the amount of white and grey matter in the brain in addition to significantly enhancing visual and auditory sustained attention.[64] Despite it's use in other fields that require expert performance, few studies have explored neurofeedback in surgery.[36,130] This may be related to the focus of present surgical training curricula on the development of competence in trainees, rather than expertise.[69,131] Technical surgical skills education is evolving from an apprenticeship model where trainees working with intraoperative surgical educators are progressively given more operative responsibility towards a competency-based quantifiable framework. Linking neurosurgical psychomotor bimanual skill performance in virtual reality simulator scenarios to resident specific training in operating room environments continues to be difficult.[10] However, the utilization of EEG monitoring during virtual reality neurosurgical procedures to outline the EEG frequency band composites of expertise and the utilization of these composites in neurofeedback may result in new formative paradigms for surgical education.[62]

Large EEG data sets can be analysed by artificial intelligence to deconstruct the frequency bands important in skilled bimanual performance.[132] 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.[86]

There are many different types of machine learning algorithms, which are based on different mathematical analytical methods applied to the data.[87] 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.[107] Machine learning has been applied in neurosurgical care, to assist in the surgical treatment of epilepsy, brain tumors, Parkinson's disease, and brain injury.[123] 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.[123] Our group has employed a number of machine learning algorithms to assess and train surgical learners.[34,119,133,134] Machine learning algorithms can be utilized to classify groups into different levels of surgical expertise with greater granularity and precision than previously demonstrated.[34,119]

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[90,135] and for less complex models, it is possible to determine the relative importance of each input metric to the model's final classification.[111] 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.[111]

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.

2.0 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.[136] Since previous research suggests differential EEG patterns between left- and right-handed individuals,[137] 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 **Figure 10** 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)[138] and the best practices for Machine Learning to Assess Surgical Experience (MLASE) reporting guidelines.[139]

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 (**Figure 1A**). The virtual reality simulated brain tumor resection scenario consisted of identical tumors and stiffness with random bleeding points.[20] The color, stiffness, and elliptical structure chosen for each tumor was a simulated glioma-like brain tumor embedded in a simulated cortical surface (**Figure 1B**). 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[38] and referenced to linked ears (**Figure 1A**). 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 collection began with a 2-minutes eyes-closed, and a 2-minutes eyes-open baseline recording. Following this baseline, participants resected 6 simulated brain tumors on the NeuroVR™ platform (CAE Healthcare, Montreal, Canada) (**Figure 1B**). 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 (**Figure 1C**).[136] Participants began by resecting tumors 1 and 2 (2 minutes were allocated per tumor resection). This was followed by a 90-second rest

period in 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.[20] 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, thus totalling 6 minutes of data per participant. Future research will explore differences in expertise under simulated stress. 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) was used to separate the raw EEG signal into various power spectra bandwidths using Biograph Infinity software (Thought Technology Ltd., Montreal, Canada). The 13 metrics 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 (15–8 Hz), beta 2 (19–22 Hz), beta 3 (23–36 Hz). 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.[50] 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 2** 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.[20] 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 age, years of practice, gender, 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.[140]

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).[141] These models represent the seven most common algorithms used in the field of artificial intelligence in healthcare.[99] Leave-one-out cross validation was used (**Figure 11**). 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 **Figure 7** 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.[111] Shapley values were plotted (**Figure 9**). All modelling and interpretability were performed using Scikit-Learn, Tensorflow, and Keras, in Python code written by the authors. A Tensorflow/Keras Sequential Model was utilized for the artificial neural network, whereas Scikit-Learn models were used for all other model types and for the pre-processing of data.

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.

3.0 Results

Demographic information is presented in **Table 3**. 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 hours, 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, 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 EEG data exhibited high fidelity, except for the previously mentioned subject that was removed due to inadequate EEG placement (**Figure 10**), and all background rhythms were normal. The classification accuracies for training and testing are illustrated in **Table 4** and the final confusion matrices of the ANN modelling are shown in **Figure 12**. 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 **Figure 9**. 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 2**). 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 ($p=0.0484$).

4.0 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. We utilized the Shapley model interpretability technique[111] 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 (**Figure 9**). A statistical analysis in average EEG bands provided the differences between skilled and less-skilled EEG activity (**Table 2**).

Since low alpha is associated with calmness and neural efficiency,[49] our findings suggest that skilled participants may have acquired abilities resulting in operating with greater composure and purpose than less-skilled participants.[67] 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.[142] This result reinforces the concept that skilled surgical performance involves cognitive elements such as enhanced composure and focus.[143]

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 2**), suggesting that skilled participants may more consistently operate with greater attention and problem-solving abilities.[144] 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.[50] Our model did not put a high emphasis on the TBR (lowest Shapley value of the 13 metrics, **Figure 9**), 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.

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.[145] 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-minute tumor resections data, allowing for personalized post-hoc feedback training. 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 metrics the model finds most useful in classifying specific skilled operative performance and alerting learners to these metrics as per neurofeedback methodology, surgical trainees could self-modify their own EEG metrics to approximate these EEG frequency metrics and improve task execution.[146] A proposed neurofeedback system using the algorithms developed in this study is outlined in **Figure 5**. It is possible to collect EEG data concurrently and integrate these with other artificial intelligence derived biometrics performance platforms[34] 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.[34]

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,[60] 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.[51] Disadvantages included the inability to assess conduct EEG spatial analysis, which outlines physiological brain locations underlying the EEG information. This is a notable limitation as spatial analysis is standard protocol in practice and has numerous advantages, as outlined in the relevant section on Temporal, Spatial and Spectral Analyses. 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. 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[91] rather than EEG timeseries information was felt to be easier for trainees to understand and learn. Since EEG[137] and hand ergonomics[25] exhibit differences between left and right-handed

individuals, left-handed participants were excluded from this investigation 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.

Another limitation of this study was the lack of control groups, which would ensure that what was classified was in fact the neural underpinnings of surgical expertise rather than potential confounders such as “familiarity” or “pleasure”. A control group of expert video game players, for example, could have been recruited and required to perform the same procedures as the participants but without having any of the requisite surgical expertise, thus elucidating the difference between general technical expertise and surgical expertise. Furthermore, the same participants could have served as their own controls during a sham procedure where they would be instructed to move their instruments stochastically across the surgical view and not operate at all, thus elucidating the difference between neural signals during surgical operation and those during other motor movements.

It has been shown that higher participant age is associated with changes in EEG patterns, such as increasing beta activity and decreased alpha activity.[147,148] 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 (**Tables 5** and **6**). Moreover, alpha peak frequency (IAF), a robust metric of brain maturation,[149] 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. One may strategically place the electrodes to minimize their number and thus associated time and costs while maximizing the resulting spatial acuity, such as by placing just 6 symmetrical electrodes on F3, F4, P3, P4, T3, and T4 (**Figure 4**). 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[78] for competency evaluation could be assessed in randomized controlled trials. 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.

5.0 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.

DISCUSSION

Model Interpretability Case Study: Olden's vs SHAP

Model interpretability techniques, as discussed in the dedicated section under Artificial Intelligence, offer different methods of understanding the inner workings of a machine learning algorithm. The Neurosurgical Simulation and Artificial Intelligence Learning Centre has traditionally made use of a model-specific method known as Olden's method or the Connection Weight Product (CWP) method in interpreting models.[34,119,150] More recent modelling in the literature has included a model-agnostic method based on Shapley values known as Shapley Additive Explanations (SHAP).[111]

A comparison can be made between the underlying theories behind Olden's method and SHAP. The CWP traces the contribution of each input metric to the final output nodes via the trained model weights. It is mathematically defined in **Equation 1**, where the CWP of a particular input metric *x* in a total of *m* metrics is equal to the sum of all product weights connecting input node *x* with hidden layer node *y* and output layer node *z*. $W_{x,y}$ and $V_{y,z}$ thus represent weights between the input node and hidden layer, and the hidden layer node and output layer node, respectively. The sign of the connection weight product indicates whether a metric's z-score value should be positive (if sign is positive) or negative (if sign is negative) to increase the likelihood of classification in the corresponding group.

$$CWP_x = \sum_{y=1}^{m} w_{x,y}\, v_{y,z}$$

Equation 1: Connection Weight Product

In contrast, SHAP treats the input metrics of a machine learning problem as players in a coalitional game from game theory. A coalition game, also known as a cooperative game, is a game wherein groups of players compete to maximize the output value.[151] Thus, Shapley values aim to answer the question, how much did each individual member of a game contribute to the final output value?[111] This can be a difficult question to answer due to interactions between members. To calculate a Shapley value, one must find the marginal contribution of a player in each possible permutation of players that includes this player. A marginal contribution is the difference between an output value that includes and that which excludes a particular player. The mean marginal contribution is taken as the Shapley value for that player. This is done for all members and is not a trivial calculation given the factorial nature of permutations. In fact, their exact computation can be implemented efficiently only in certain (decision tree-based) models, whereas they must be approximated using other models.[111] Mathematically, the Shapley value $\Phi_i(v)$ of player *i* in a coalition game (N, *v*) is shown in **Equation 2**, with σ representing a player set (i.e. a subset of the input metrics) and the set of permutations of N (i.e. all input metrics) denoted as Π(N).[152] The summated term, $v(P_\sigma(i) \cup \{\mathrm{i}\}) - v(P_\sigma(i))$, represents the marginal contribution of a player *i* to a coalition, as the difference between the output value that includes ($v(P_\sigma(i) \cup \{\mathrm{i}\})$) and the output value that excludes ($v(P_\sigma(i))$) player *i*.

$$\Phi_i(v) = \frac{1}{n!} \sum_{\sigma \epsilon \Pi(\mathrm{N})} v(P_\sigma(i) \cup \{\mathrm{i}\}) - v(P_\sigma(i))$$

Equation 2: Shapley Value

A comparison can also be made between the results from the Shapley value method (**Figure 9**) and the Connection Weight Product method (**Figure 8**), which in the context of the

present study reveals striking similarities and differences. For instance, they both ranked beta 1 as more important than beta 3. This was to be expected, as although beta 1 was significantly different across expertise groups ($p = 0.0141$), beta 3 was not ($p = 0.3635$) (**Table 2**). Furthermore, the lower frequency beta bands, as opposed to the higher frequency beta 3 band, are more associated with positive cognitive processes, such as attention, memory formation, and performance (**Table 1**). In contrast, beta 3 is associated with relatively negative cognitive traits such as worry and anxiety (**Table 1**). This suggests that more positive cognitive traits characterize the difference between skilled and less-skilled individuals.

Moreover, both interpretability methods ranked TBR relatively low (**Figures 8** and **9**). Although it was expected that TBR rank relatively high based on conventional neuroscientific data, as well as its relative difference across expertise groups as illustrated in **Table 1**, an analysis through an artificial intelligence lens reveals that this is not the case. However, as a composite metric, TBR is composed of both theta and beta metrics, which are both found in their pure form in the neural network as well. As such, the final neural network model could have learned to assign a low weight to the TBR input neuron, as it could derive the same information from other neurons. Indeed, this is precisely what the relatively low rank assigned to TBR by the Connection Weight Product seems to indicate.

Furthermore, low alpha, the highest ranked metric by Shapley standards is the third highest on CWP, and even though it is almost double the second highest metric according to Shapley analysis (**Figure 9**), it is relatively similar to the first and second highest metrics under Olden's method (**Figure 8**). Low alpha was found to be significantly different across expertise groups ($p = 0.0443$) and as such this was to be expected. Also, the relatively high appraisal of one metric by

two independent systems gives the machine learning practitioner more confidence in the result, while also giving credence to both methods as valid forms of understanding algorithms.

Given this finding, it may be tempting to claim that the primary difference between skilled and less-skilled participants is that skilled participants expend less mental resources in achieving the same motor action. However, this is an erroneous conclusion that arises from an incorrect understanding of model interpretability tools. Rank order lists of metrics (feature importance lists) generated via model interpretability methods cannot be used to conclusively elucidate the importance of each metric to a list. Rather, such interpretability methods only make claims about the particular model upon which they were applied. Thus, in this case, it may be concluded that the final neural network places particularly high influence on the low alpha metric, and that as a result a neurofeedback regime that makes use of this particular model should be focused on the low alpha metric for training purposes. However, to generalize this to all possible models would be erroneous.

Although model interpretation methods have generally tended to rely more on Shapley values in recent times,[117] the Connection Weight Product has merit for simpler models. It is able to provide directed feedback as to the contribution of each individual metric, while being much simpler and more efficient to calculate. Thus, there is still a role for the CWP method as simple models may sometimes outperform their more complex counterparts, particularly when there is relatively less data upon which to train a model, or when the situation modeled is simpler. However, CWP does have limits. For example, deeper learning methods such as Long-Short Term Memory (LSTM) type models cannot be interpreted using the CWP model as connections between neurons are more complicated. Furthermore, since it is model-specific, it cannot be used on any non-neural network model type.

Timeseries-based analysis of EEG

As mentioned in the Temporal, Spatial, and Spectral Analyses section on EEG, there are a variety of potential ways to analyze EEG data. The present study made use only of spectral analysis due to present limitations of electrode number and computational availability. However, a more comprehensive machine learning classification approach would factor in data from all three analysis types. Given the fact that there are numerous advantages associated with the use of only one electrode, additional work may focus on temporal analysis at the expense of spatial analysis. Furthermore, temporal analytic data has been shown to be the most important in output classification amongst the three types of analysis.[58]

Different types of events in surgery may be investigated. One such analysis may involve timestamping surgical tool usage, while others may involve timestamping the phase of a relatively linear surgical operation. Thereafter, it would be possible to compare the event response in a timeseries EEG analysis, by considering the shape of waveforms, as opposed to analyzing EEG band means. This would add another layer of analysis to expertise that makes use of relatively more data, and as such, may provide more accurate classification of expertise. An example of a potential temporal analysis of EEG is found in **Figure 3**.

However, this sort of analysis may be more difficult to use for the purposes of neurofeedback. Although the metrics inputted into a machine learning algorithm may theoretically be abstract to the point of incomprehension, they should generally be developed such that they can be understood and taught by surgical educators. This is particularly important if the ultimate goal of the research is to provide workable training routines via neurofeedback, as neurofeedback depends on the availability of trainable metrics, and it may be relatively more difficult to inculcate particular waveforms in trainees rather than train to meet particular EEG band thresholds.

Thesis Conclusion

Summary

The present thesis represents evidence for the utility of electroencephalography (EEG) in surgical education. Our results are consistent with the hypothesis that EEG recordings during the virtual reality simulated surgery assessed can accurately classify the surgical expertise of participants using machine learning algorithms. Furthermore, all four objectives were achieved. Firstly, seven model types were trained, corresponding to the seven most common model types in the applications of machine learning in healthcare and were compared in their ability to classify expertise using EEG data. The Support Vector Machine (SVM) and the artificial neural network (ANN) were the best performing models (AUROC = 0.833 and 1.0 respectively), while considerably poorer results were achieved with other model types, such as the Naïve Bayes and Random Forest (AUROC = 0.639 and 0.583 respectively). Next, our artificial neural network fulfilled our criteria of a model capable of accurately distinguishing between skilled and less-skilled participants on a virtual neurosurgical simulation, although a study with a larger sample size is needed to validate its accuracy further. Thirdly, a statistical analysis of differences in EEG frequency bands between skilled and less-skilled participants revealed significant differences. Specifically, skilled surgeons displayed higher ($p = 0.044$) low alpha (8–10 Hz) than the less-skilled group. Furthermore, skilled surgeons displayed significantly lower theta/beta ratio (TBR) ($p = 0.048$), confirming the literature on this ratio as a marker of technical expertise. Skilled surgeons also displayed 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. Finally, Shapley model interpretation identified low alpha (8–10 Hz), which is identified with neural efficiency, as the most important metric for classifying expertise.

Artificial intelligence has been advancing at a faster rate than legislation or research on its ethics and morality.[98] Therefore, in order to best facilitate integration of such machine learning systems into the medical or surgical education curricula, it would be prudent to conduct more studies into the ethical and social consequences of the use of AI in medical or surgical education. The social ramifications of an approach to objectively quantify expertise using a myriad of biosensors are unclear. Surgical trainees and other healthcare practitioners have to prepare for a large number of examinations and as such the advent of self-administered expertise examinations may only add to their heavy workload. In addition, the state of being equipped with many sensors ranging from tool motion sensors, physiological heart and breathing rate sensors, and EEG and the ensuing vulnerability and personal accountability remains unexplored. Furthermore, certain unprecedented policies may begin to be explored as quantifying expertise becomes easier. For instance, the disqualification of medical students pursuing surgical education based on their natural talent or lack thereof as assessed by an AI algorithm, may begin to be justified by the potential for decreased harm to patients and decreased stress to students, though it is important to not undermine the incredible human potential for personal growth and development here. Such approaches to medical education have wide-reaching consequences and may alter the criteria by which technical specialists are recruited in general.

The author is not an advocate for the total replacement of human interaction in favor of a computerized approach to surgical expertise assessment. While self-administered tests and training procedures present their own advantages, such as convenience and relative economics, a hybrid

approach combining human feedback with AI may be optimal. This is a general guideline in AI research and is not limited to healthcare, although the unique combination of art and science present in healthcare makes it a prime candidate.[153] While an AI algorithm may be able to teach the science of a surgical procedure, for example, bedside manners will likely remain in the purview of surgical educators.[153]

Future Directions

Several studies may be done to validate the findings of this pilot study. This study may be replicated with a larger sample size to ensure generalizability. The use of more than one electrode would enable spatial analysis in addition to the possible temporal and spectral analyses that may be conducted with only one electrode. In this case, it may be preferable to put emphasis on electrodes on the temporal and occipital lobes as EEG data from these areas has previously been shown to elucidate expertise[154].

Although it was previously unknown which neurofeedback training protocol may be optimal in surgical training, the present results suggest that alpha neurofeedback may be the most beneficial. Thus, a neurofeedback protocol that makes use of this ordering may be performed and compared with an alternative order in its effectiveness.

As suggested in the aforementioned section on Timeseries-based analysis of EEG, an event-based study may be conducted in which particular phases of a surgery are timestamped and responses between skilled and less-skilled participants recorded as waveforms. This would enable a more granular understanding of psychological expertise as applied to surgery.

Moreover, although this study investigated the resection of a glioma-like brain tumor, other more complex surgical procedures, and those of other surgical specialties along with technical medical procedures could be assessed. This may elucidate the differences, for example, between skilled and less-skilled individuals across different gradations of procedural complexity. It is possible that the psychological profile of expert surgeons somehow renders them more immune to abrupt changes or otherwise stressful events in surgery. Such a study may thus enable the training of resistance to, and perseverance in the face of, stressful events.[20]

Finally, in an attempt to lay out a more holistic picture of expertise, it would be interesting to combine the results of classifiers based on several modalities. The NeuroSim group has previously shown that tool motion sensors may be used to accurately predict expertise.[34] Combining this result with the present study, and in addition to other sensors such as physiological heart rate sensors, may result in more accurate classifications and more personalized feedback.

APPENDIX

Figures

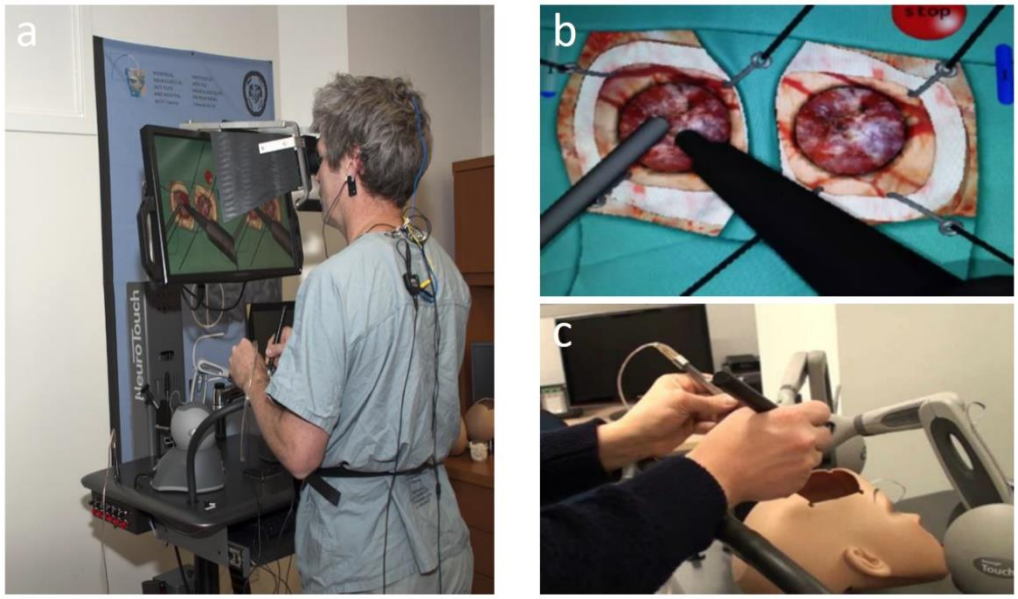

Figure 1. Virtual neurosurgical experimental setup

(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.

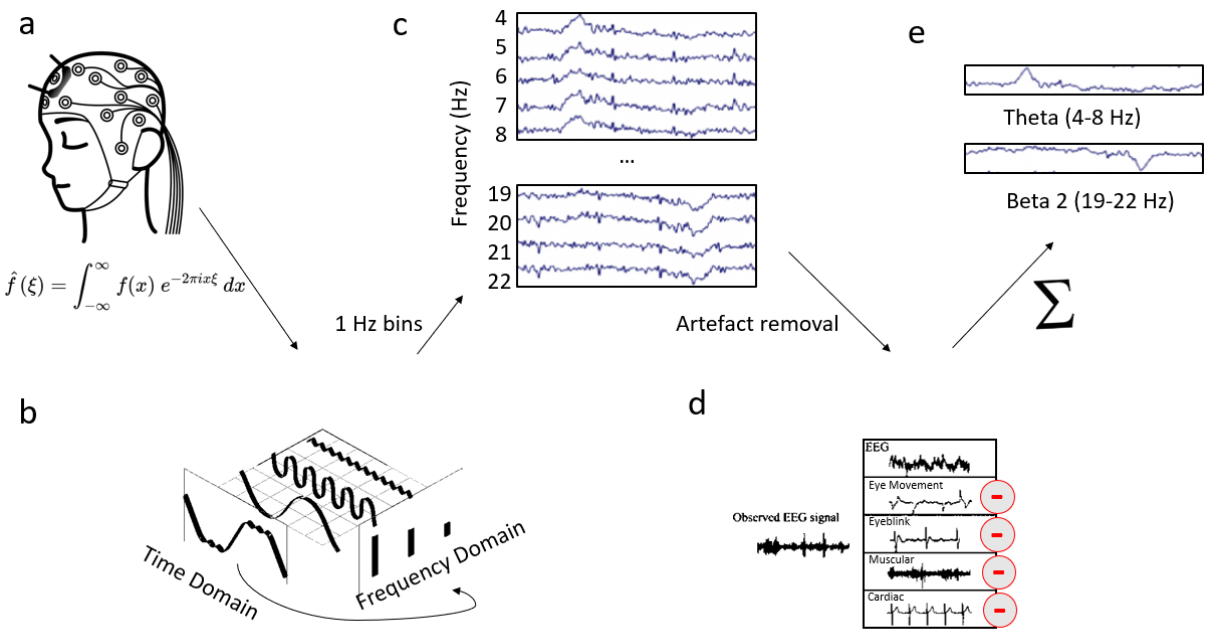

Figure 2. EEG processing workflow

(a) EEG data is collected via electrodes on the scalp. (b) The Fast Fourier Transform (FFT) is used to decompose the raw EEG signal from the time domain into the frequency domain using Biograph® Infiniti software from Thought Technology. (c) The resulting pure EEG frequency domain signal is split into various 1 Hz wide bins. (d) Artefacts in the recorded EEG signal, arising from electrophysiological sources such as eye movements,[41] eyeblinks,[41] muscular activity,[42] and cardiac rhythm,[43] as well as environmental electrical activity, are filtered out using Biograph® Infiniti software or other specialized tools. (e) Finally, various 1 Hz bins are summed together to form EEG bandwidths in accordance with guidelines from the literature. Specifically, this study calculated delta (2-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), and beta (13-30 Hz) bands, as well as the sensorimotor rhythm (SMR, 12-15 Hz). Low (4-6 Hz) and high theta bands (6-8 Hz) and beta 1 (15-18 Hz), 2 (19-22 Hz), and 3 (23-36 Hz) bands were further calculated. Furthermore, a composite metric known as the theta/beta ratio (TBR) was calculated as the square of the theta band divided by the square of the 13-21 Hz band. A full list of calculated metrics is found in **Table 1**.

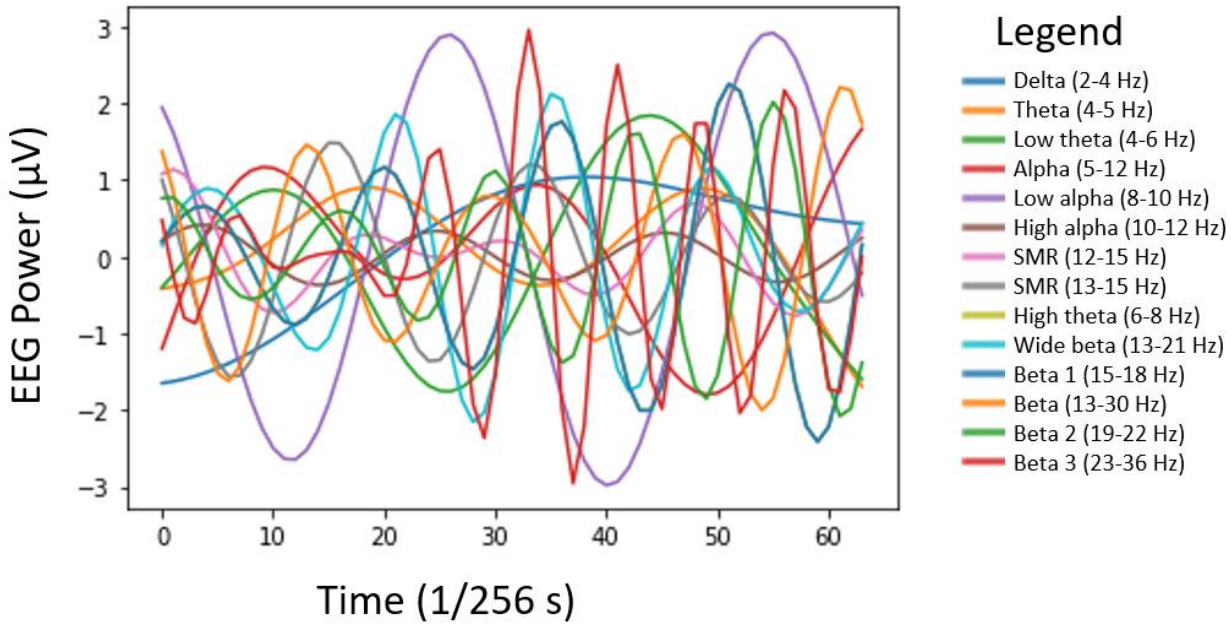

Figure 3. Timeseries-based (temporal) analysis of EEG

An example of a temporal representation of specially curated EEG signal wavelets. The horizontal axis represents time as increments of 1/256 s, in line with a 256 Hz sampling protocol, for a total of a quarter of a second. The vertical axis represents the EEG power in microvolts of each individual metric. Much more information is present here than is available through a spectral analysis and as such, it may be possible to train much more accurate machine learning models using this representation. Furthermore, unlike spectral analysis, models trained using this representation may be able to deliver more classifications in a given unit of time as no averaging of data is required. Lastly, this approach is particularly powerful when timestamped to particular surgical events, such as a phase in a tumor resection, or the use of a certain tool. By averaging out inter-subject differences, it is possible to find the event-related potential (ERP) of a particular component of a surgery.

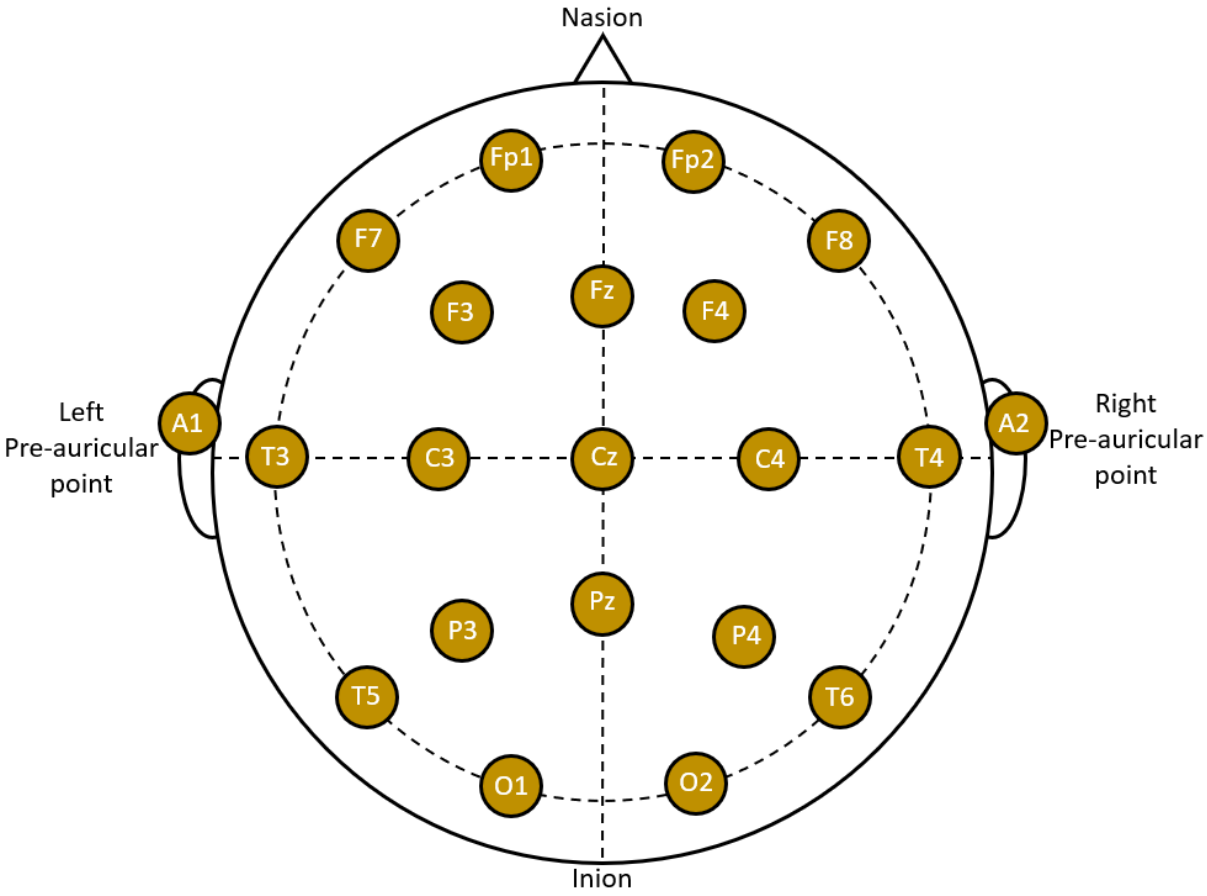

Figure 4. 10-20 System for EEG electrode placement

A standardized set of guidelines on how to place electroencephalography (EEG) electrodes on the human head. This system relies on four key skull markers: the nasion (Nz) at the bridge of the nose; the inion (Iz), the bony extrusion at the back of the head; and the two pre-auricular points beside each ear (LPA and RPA).[59] Electrodes on the left side of the head are characterized by odd numbers, while those on the right side are denoted by even numbers. Fp, F, C, T, P, and O denote prefrontal, frontal, central, temporal, parietal, and occipital electrodes respectively and correspond to cortical regions. Z refers to electrodes placed along the midline sagittal plane of the skull. The 10-20 refers to the fact that electrodes are placed at a distance of 10% and 20% of the total skull width away from each pre-auricular point. Although more granular standardized systems, like the 10-10 and 10-5 systems, exist, they are based on the 10-20 system.[59] The Cz location was used in the present study.

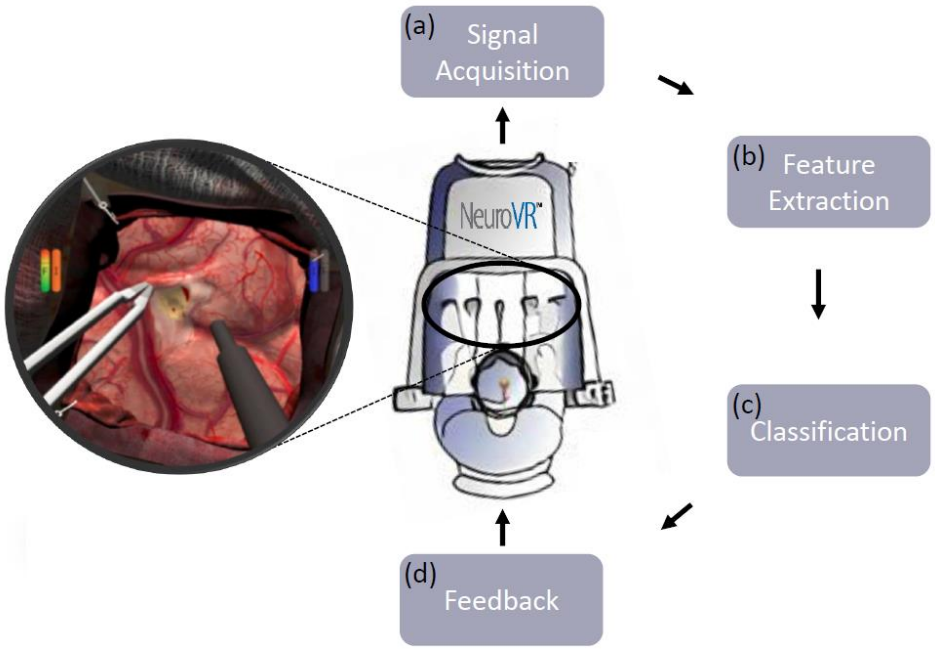

Figure 5. Proposed neurofeedback training protocol in surgical simulation training

(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.

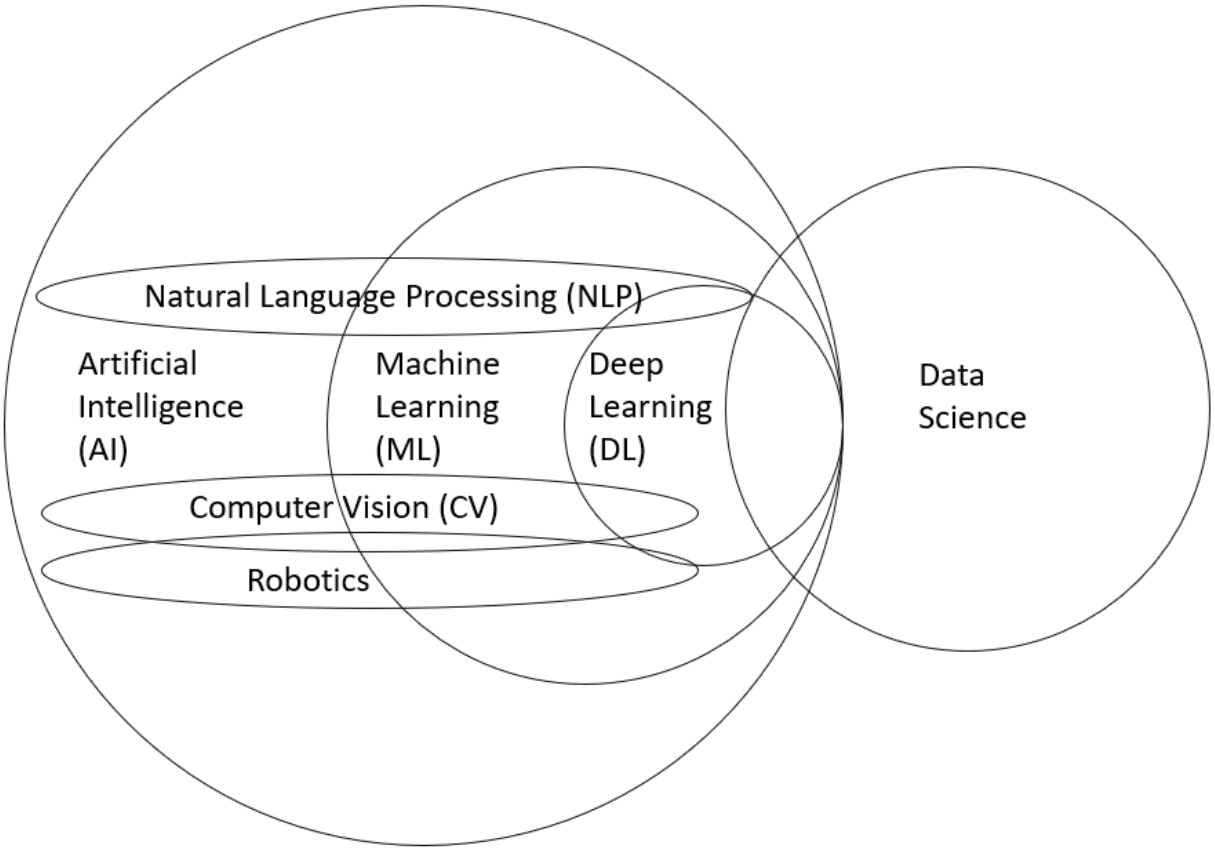

Figure 6. Artificial intelligence Venn diagram

This Venn diagram illustrates the relationship between artificial intelligence (AI), machine learning (ML), deep learning (DL), and data science, as well as several selected branches of AI. Artificial intelligence is the development of knowledge-based systems that mimic human behaviour. Machine learning is a branch of AI that develops such systems by training them on data rather than having them be programmed through expert-based rules. Deep learning is a subset of machine learning that is characterized by more complex algorithms and larger data requirements. Artificial intelligence fields, such as natural language processing (NLP), computer vision (CV), and robotics, may or may not make use of ML or DL. Data science is a modern interdisciplinary field that uses the scientific method, as well as statistics, algorithms, and advanced mathematical principles to derive insight from data. Artificial intelligence and ML/DL play an active role in a data scientist's toolkit.

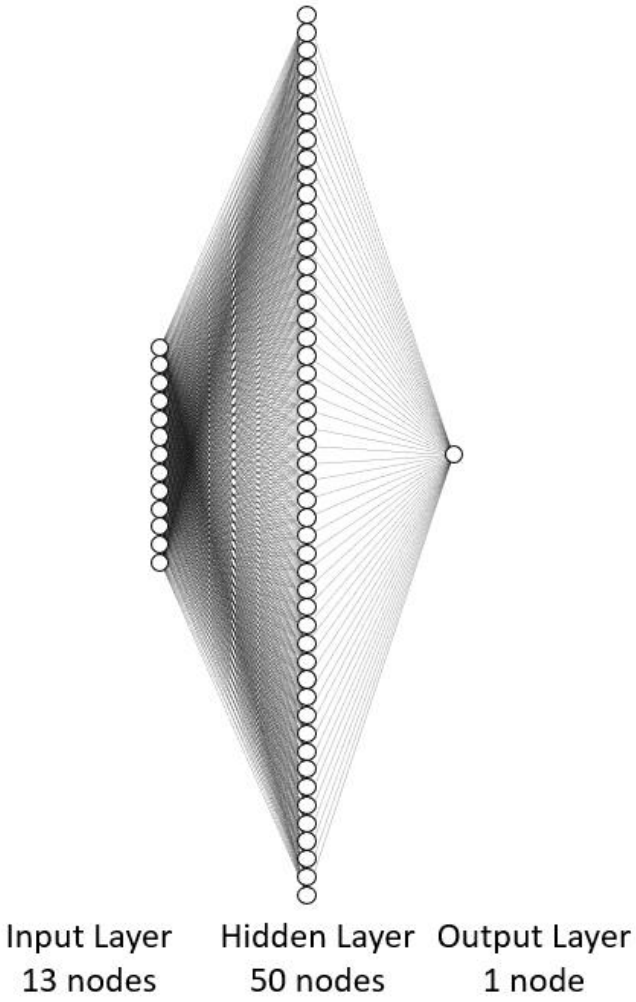

Figure 7. The structure of the final artificial neural network model

This model predicts surgical expertise on a binary basis based on EEG data. Thirteen input neurons are each fed one EEG frequency band, or the theta/beta ratio (TBR), and are fully connected to a dense layer of 50 fully connected neurons, which were in turn fully connected to one output neuron. The output layer predicted the probability that the surgical resection was performed at the less-skilled level. Given that it is a binary classification, a 0.5 threshold was implemented, above of which the less-skilled class was assigned. The ANN model was compiled using the well-known Adam optimizer[155] and the binary cross entropy function as a loss function. It was fit using up to 1000 epochs through the training data with an early stopping procedure implemented to cut the training process short if the model's loss function did not significantly decrease after any given run of 30 epochs.

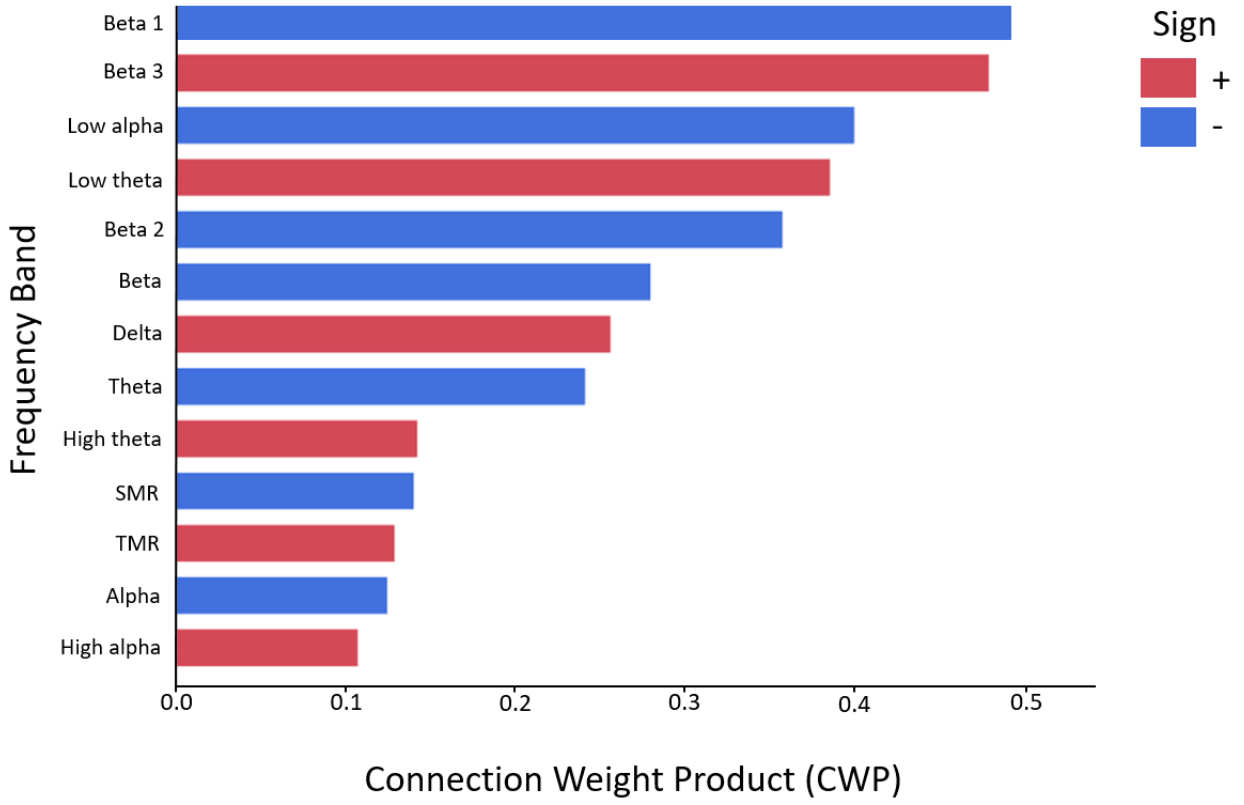

Figure 8. Connection Weight Product (CWP) interpretability plot

Bar plot illustrating the contribution of each frequency band as calculated by Connection Weight Product (CWP) on the final artificial neural network, in descending order.[116] The CWP traces the contribution of each input metric to the final output classification via the trained model weights. A higher final output classification represents less-skilled performance. The sign of the CWP is represented by the color red (positive) or blue (negative). Thus, a negative connection weight product signifies that a high amount of that metric is associated with expertise; whereas a positive connection weight product signifies that a high amount of that metric is associated with less-skilled trainees. Metrics assigned smaller CWPs do not contribute as much to the neural network's expertise classification, relative to those with larger CWPs. Beta 1 was deemed the most important metric to expertise classification, with a higher beta 1 being associated with skilled performance. In contrast, beta 3 was the second most important metric, but a higher amount of beta 3 was associated with less-skilled performance. Low alpha, the most important metric according to the Shapley interpretability method (**Figure 9**), also ranked relatively highly here at third place.

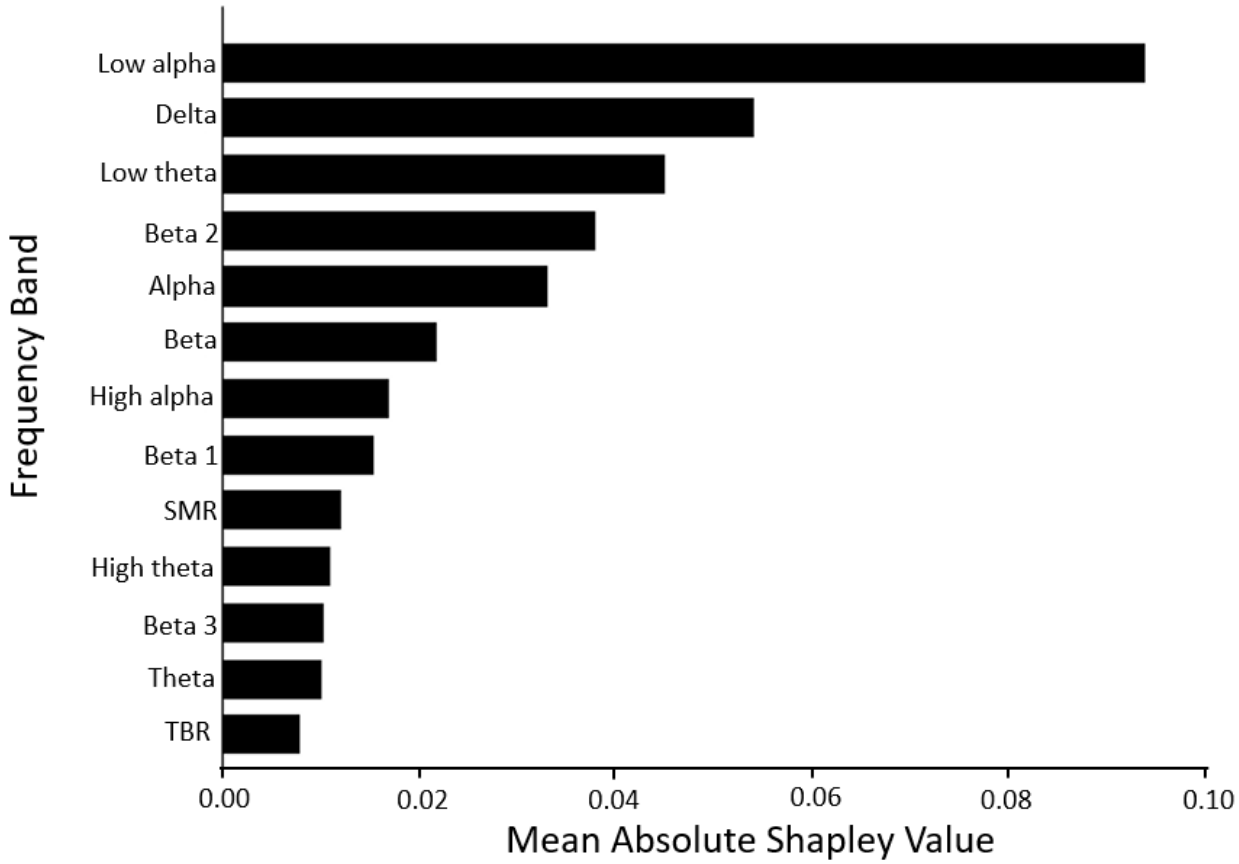

Figure 9. Shapley interpretability plot

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.[111] 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. Shapley values were calculated using the KernalExplainer algorithm from the SHAP (Shapley additive explanations) Python package. The low alpha band was by far the most important factor in expertise classification.

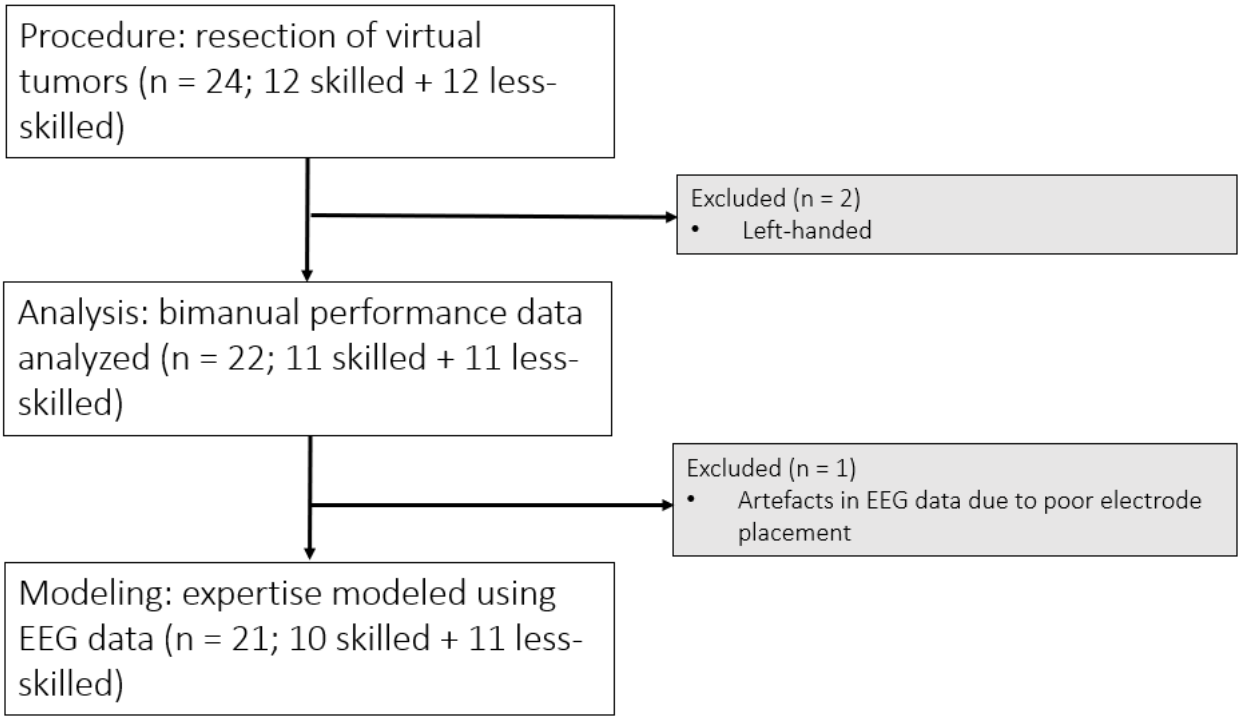

Figure 10. Inclusion and exclusion of participants

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).

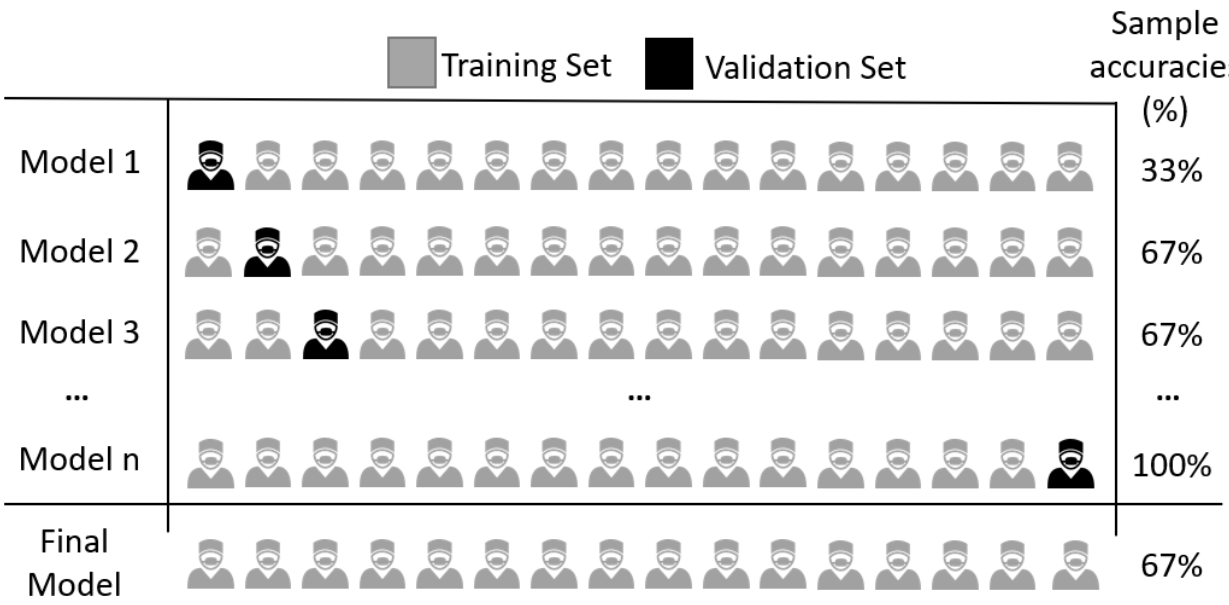

Figure 11. Leave-one-out-cross-validation (LOOCV)

Leave-one-out-cross-validation (LOOCV) is a best-practice method within machine learning research that maximizes the amount of training data available for modelling. The purpose of this procedure, as with all validation splits, is to allow for tuning of untrainable characteristics, such as the number of neurons in a hidden layer and the number of hidden layers, in a process known as hyperparameter tuning. Following the removal of a testing set, the remaining subjects are ordinarily divided into a training set and a validation set. This further decreases the size of the set available for training, which potentially could result in worse modelling outcomes, especially with an already small sample size. To combat this, LOOCV calls for a set of models to be iteratively trained on all but one participant, who is withheld for validation. The model results are averaged to arrive at final training accuracy scores. Lastly, one final model is trained on the whole training data using the hypertuned parameters and then tested on the testing dataset. This method is only practical for models with sufficiently small sample size as it would otherwise require a prohibitive amount of compute resources.

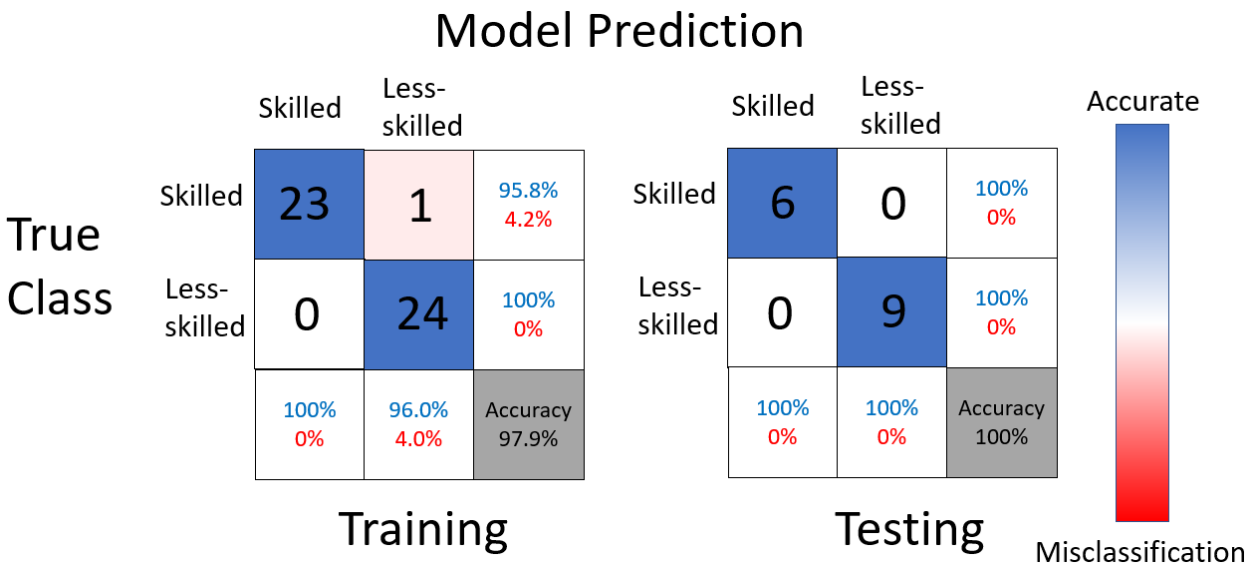

Figure 12. Confusion matrices of the training and testing results of the artificial neural network

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).

Tables

Table 1. EEG frequency band significances

A complete list of the EEG frequency bands used in this study and their respective significances based on a survey of the literature. EEG band frequencies may differ between individuals and are demarcated slightly differently from one source to another, so representative sources have been used.[40] Notably, although beta 2 has been defined as 19-22 Hz in this study, it has also been reported it as 16.5-20 Hz.[156] Likewise, beta 3 has been defined as 23-36 Hz, while it has also been reported as 20.5-28 Hz.[156] Furthermore, different software products calculate the theta/beta ratio (TBR) in slightly different ways.[157]

Name	Frequency (Hz)	Significance
Delta (δ)	0.5-4[158]	Slow-wave sleep[62]
Theta (θ)	4-8[62]	Learning, memory, and intuition[62]
Low Theta	4-6[159]	Memory[160,161]
High Theta	6-8[162]	Spatial attention[163]
Alpha (α)	8-12[164]	Calmness, tranquility,[62] visual processing[127]
Low Alpha	8-10[164]	Neural efficiency,[165] recall[62]
High Alpha	10-12[49]	Optimize cognitive performance[62]
Sensorimotor Rhythm (SMR)	12-15[65]	Mental alertness, immobility[62]
Beta (β)	13-30[51]	Focused attention[51]
Beta 1	15-18[65]	Memory formation[166]
Beta 2	19-22[156]	Energy, anxiety, and performance[167]
Beta 3	23-36[156]	Worry, anxiety[62]
Theta Beta Ratio (TBR)	$(4\text{-}8)^2$ / $(13\text{-}21)^2$ [50]	Creative/intuitive image-based thought versus logical/rational language-based thought[50,168]

Table 2. EEG band means across expertise

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.

EEG Metrics	Skilled (n=10) (uV)	Less-skilled (n=11) (uV)	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 $(4\text{-}8\ Hz)^2/(13\text{-}21Hz)^2$	1.93±0.13	2.67±0.33	0.0484*

Table 3. Participant demographics stratified by expertise level

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.

	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

Table 4. Modelling results

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.

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

Table 5. Regression analysis of EEG bands during eyes closed baseline across age and years in practice

An ANCOVA was conducted to test differences in correlation between age and years in practice with each of the metrics used in the study, as well as the IAF (alpha peak frequency). R^2 coefficients and P values are presented. The final column constitutes p values from a comparison of the two regression lines. Significant differences at the 0.05 level are indicated with an asterisk. We observe correlations between all beta waves plus the TBR with years in practice. In contrast, beta 3 did not significantly correlate with age. Correlation coefficients were generally higher, albeit insignificantly so, in years of practice comparisons. The alpha peak frequency (IAF) is seen as a robust metric of brain maturation,[149] and there is no correlation between IAF and age. This information provides evidence that differences in beta waves between the skilled and less-skilled classes outlined in **Table 3** were more related to their difference in experience than their difference in age.

EEG Bandwidth (mean)	Years in Practice R^2	Years in Practice p Value	Age R^2	Age p Value	Comparison
Delta	0.0632	0.2715	0.1467	0.087	0.9199
Theta	0.0125	0.6298	0.0053	0.7533	0.9272
Low Theta	0.0567	0.2987	0.1554	0.0770	0.6876
High Theta	0.0644	0.2669	0.0053	0.7548	0.7407
Alpha	0.0024	0.8319	0.0009	0.8972	0.3047
Low Alpha	0.0372	0.4023	0.0130	0.6226	0.5093
High Alpha	0.0144	0.6039	0.0289	0.4611	0.3634
SMR	0.0386	0.3931	0.0085	0.6915	0.2998
Beta	0.3105	0.0087*	0.1900	0.0482*	0.5079
Beta 1	0.4021	0.0020*	0.2659	0.0167*	0.6807
Beta 2	0.4433	0.0010*	0.3256	0.0069*	0.8362
Beta 3	0.2623	0.0176*	0.1522	0.0804	0.5036
TBR	0.2780	0.0140*	0.2720	0.0153*	0.4255
IAF	0.1618	0.0707	0.1221	0.1204	0.5732

Table 6. Regression analysis of EEG bands during eyes open baseline across age and years in practice

An ANCOVA was conducted to test differences in correlation between age and years in practice with each of the metrics used in the study, as well as the IAF (alpha peak frequency). R^2 coefficients and P values are presented. The final column constitutes p values from a comparison of the two regression lines. Significant differences at the 0.05 level are indicated with an asterisk. We observe correlations between all beta waves plus the TBR with years in practice. In contrast, neither overall beta nor beta 3 significantly correlated with age, although low theta did. Correlation coefficients were generally higher, albeit insignificantly so, in years of practice comparisons. Furthermore, the alpha peak frequency (IAF) is seen as a robust metric of brain maturation,[149] and there is no correlation between this metric and age. This is evidence that differences in beta waves between the skilled and less-skilled classes observed between the skilled and less-skilled classes in **Table 3** were more related to their difference in experience than their difference in age.

EEG Bandwidth (mean)	Years in Practice R^2	Years in Practice p Value	Age R^2	Age p Value	Comparison
Delta	0.020	0.5383	0.078	0.2218	0.7434
Theta	0.019	0.5477	0.105	0.1518	0.6743
Low Theta	0.084	0.2026	0.219	0.0324*	0.5946
High Theta	0.0002	0.9476	0.032	0.4377	0.8854
Alpha	0.007	0.7155	0.002	0.8597	0.6085
Low Alpha	0.027	0.4740	0.002	0.8667	0.9880
High Alpha	0.0001	0.9614	0.011	0.6476	0.2676
SMR	0.032	0.4380	0.005	0.7624	0.4359
Beta	0.300	0.0101*	0.162	0.0703	0.4443
Beta 1	0.258	0.0186*	0.144	0.0898	0.6067
Beta 2	0.428	0.0013*	0.253	0.0202*	0.6021
Beta 3	0.256	0.0193*	0.139	0.0967	0.3423
TBR	0.278	0.0140*	0.271	0.0155*	0.3291
IAF	0.051	0.3241	0.038	0.4004	0.5733

REFERENCES

1. Soper, N. J. SAGES and surgical education: Assuring that history does not repeat itself. *Surg. Endosc.* **15**, 775–780 (2001).

2. Khan, M. R. & Begum, S. Apprenticeship to simulation - The metamorphosis of surgical training. *J. Pak. Med. Assoc.* **71 1)**, S72–S76 (2021).

3. Pattani, R., Wu, P. E. & Dhalla, I. A. Resident duty hours in Canada: Past, present and future. *Cmaj* **186**, 761–765 (2014).

4. Badash, I., Burtt, K., Solorzano, C. A. & Carey, J. N. Innovations in surgery simulation: A review of past, current and future techniques. *Ann. Transl. Med.* **4**, 1–10 (2016).

5. Hall, J. C., Crebbin, W. & Ellison, A. Towards a hybrid philosophy of surgical education. *ANZ J. Surg.* **74**, 908–911 (2004).

6. Sigaux, N. *et al.* 3D Bioprinting : principles , fantasies and prospects A Review. (2018).

7. Werz, S. M., Zeichner, S. J., Berg, B. I., Zeilhofer, H. F. & Thieringer, F. 3D Printed Surgical Simulation Models as educational tool by maxillofacial surgeons. *Eur. J. Dent. Educ.* **22**, e500–e505 (2018).

8. Palter, V. N. & Grantcharov, T. P. 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**, 443–448 (2014).

9. Kotsis, S. V. & Chung, K. C. Application of the ‘see one, do one, teach one’ concept in surgical training. *Plast. Reconstr. Surg.* **131**, 1194–1201 (2013).

10. Kordowicz, A. G. R. & Gough, M. J. The challenges of implementing a simulation-based surgical training curriculum. *Br. J. Surg.* **101**, 441–443 (2014).

11. Agha, R. A. & Fowler, A. J. The role and validity of surgical simulation. *Int. Surg.* **100**, 350–357 (2015).

12. Mirchi, N., Ledwos, N. & Del Maestro, R. F. Intelligent Tutoring Systems: Re-Envisioning Surgical Education in Response to COVID-19. *Can. J. Neurol. Sci. / J. Can. des Sci. Neurol.* **00**, 1–3 (2020).

13. Massoth, C. *et al.* High-fidelity is not superior to low-fidelity simulation but leads to overconfidence in medical students. *BMC Med. Educ.* **19**, 1–8 (2019).

14. McDougall, E. M. Validation of surgical simulators. *J. Endourol.* **21**, 244–247 (2007).

15. Palter, V. N. & Grantcharov, T. P. 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**, 443–448 (2014).

16. Seymour, N. E. VR to OR: A review of the evidence that virtual reality simulation improves operating room performance. *World J. Surg.* **32**, 182–188 (2008).

17. Kundhal, P. S. & Grantcharov, T. P. Psychomotor performance measured in a virtual environment correlates with technical skills in the operating room. *Surg. Endosc. Other Interv. Tech.* **23**, 645–649 (2009).

18. Endo, K. *et al.* A patient-specific surgical simulator using preoperative imaging data: an interactive simulator using a three-dimensional tactile mouse. *J. Comput. Surg.* **1**, 2–9 (2014).

19. Palter, V. N. & Grantcharov, T. P. Simulation in surgical education. *Cmaj* **182**, 1191–1196 (2010).

20. Bajunaid, K. *et al.* Impact of acute stress on psychomotor bimanual performance during a

simulated tumor resection task. *J. Neurosurg.* **126**, 71–80 (2017).

21. Van Der Meijden, O. A. J. & Schijven, M. P. The value of haptic feedback in conventional and robot-assisted minimal invasive surgery and virtual reality training: A current review. *Surg. Endosc.* **23**, 1180–1190 (2009).

22. Brunozzi, D., McGuire, L. S. & Alaraj, A. NeuroVR^{TM} Simulator in Neurosurgical Training. in (ed. Alaraj, A.) 211–218 (Springer, 2018).

23. NeuroVR Neurosurgical Simulator | CAE Healthcare. https://www.caehealthcare.com/surgical-simulation/neurovr/.

24. Delorme, S., Laroche, D., Diraddo, R. & F. Del Maestro, R. NeuroTouch: A physics-based virtual simulator for cranial microneurosurgery training. *Neurosurgery* **71**, 32–42 (2012).

25. Sawaya, R. *et al.* Virtual reality tumor resection: The force pyramid approach. *Oper. Neurosurg.* **14**, 686–696 (2018).

26. Ross, S., Hauer, K. E. & Melle, E. van. Outcomes are what matter: Competency-based medical education gets us to our goal. *MedEdPublish* **7**, 1–5 (2018).

27. Bhatti, N. I. & Cummings, C. W. Viewpoint: Competency in surgical residency training: Defining and raising the bar. *Acad. Med.* **82**, 569–573 (2007).

28. Stulberg, J. J. *et al.* Association between Surgeon Technical Skills and Patient Outcomes. *JAMA Surg.* **155**, 960–968 (2020).

29. Vedula, S. S., Ishii, M. & Hager, G. D. Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. *Annu. Rev. Biomed. Eng.* **19**, 301–325 (2017).

30. Azari, D. P. *et al.* A Comparison of Expert Ratings and Marker-Less Hand Tracking along

OSATS-Derived Motion Scales. *IEEE Trans. Human-Machine Syst.* **51**, 22–31 (2021).

31. Niitsu, H. *et al.* Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surg. Today* **43**, 271–275 (2013).

32. Lavanchy, J. L. *et al.* Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci. Rep.* **11**, 1–9 (2021).

33. Böhm, B., Rötting, N., Schwenk, W., Grebe, S. & Mansmann, U. A prospective randomized trial on heart rate variability of the surgical team during laparoscopic and conventional sigmoid resection. *Arch. Surg.* **136**, 305–310 (2001).

34. Winkler-Schwartz, A. *et al.* Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation. *JAMA Netw. open* **2**, e198363 (2019).

35. Aizuddin, M., Oshima, N., Midorikawa, R. & Takanishi, A. Development of sensor system for effective evaluation of surgical skill. *Proc. First IEEE/RAS-EMBS Int. Conf. Biomed. Robot. Biomechatronics, 2006, BioRob 2006* **2006**, 678–683 (2006).

36. Ros, T. *et al.* Optimizing microsurgical skills with EEG neurofeedback. *BMC Neurosci.* **10**, (2009).

37. Bocci, T. *et al. How does a surgeon's brain buzz? An EEG coherence study on the interaction between humans and robot*. *Behavioral and Brain Functions* vol. 9 (2013).

38. Jasper, H. H. *Report of the committee on methods of clinical examination in electroencephalography. 1957*. *Electroencephalography and Clinical Neurophysiology* vol. 10 (1958).

39. Ferree, T. C., Luu, P., Russell, G. S. & Tucker, D. M. Scalp Electrode Impedance and EEG Data Quality. *Clin. Neurophysiol.* **112**, 1–9 (2001).

40. Nayak, C. S. & Anilkumar, A. C. EEG normal waveforms. In: StatPearls. *StatPearls* 1–6 (2020).

41. Singh, B. & Wagatsuma, H. A Removal of Eye Movement and Blink Artifacts from EEG Data Using Morphological Component Analysis. *Comput. Math. Methods Med.* **2017**, (2017).

42. Chen, X., He, C. & Peng, H. Removal of muscle artifacts from single-channel EEG based on ensemble empirical mode decomposition and multiset canonical correlation analysis. *J. Appl. Math.* **2014**, (2014).

43. Tamburro, G., Stone, D. B. & Comani, S. Automatic removal of cardiac interference (ARCI): A new approach for EEG data. *Front. Neurosci.* **13**, 1–17 (2019).

44. Hamada, M., Zaidan, B. B. & Zaidan, A. A. IMAGE & SIGNAL PROCESSING A Systematic Review for Human EEG Brain Signals Based Emotion Classification, Feature Extraction, Brain Condition, Group Comparison. *J. Med. Syst.* **42**, 162 (2018).

45. Jiang, X., Bian, G. Bin & Tian, Z. Removal of artifacts from EEG signals: A review. *Sensors (Switzerland)* **19**, 1–18 (2019).

46. Smith, S. J. M. EEG in the diagnosis, classification, and management of patients with epilepsy. *Neurol. Pract.* **76**, (2005).

47. Ahmadian, P., Cagnoni, S. & Ascari, L. How capable is non-invasive EEG data of predicting the next movement? a mini review. *Front. Hum. Neurosci.* **7**, 1–7 (2013).

48. Casson, A. J., Yates, D. C., Smith, S. J. M., Duncan, J. S. & Rodriguez-Villegas, E.

Wearable electroencephalography. *IEEE Eng. Med. Biol. Mag.* **29**, 44–56 (2010).

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

50. Clarke, A. R., Barry, R. J., Karamacoska, D. & Johnstone, S. J. The EEG Theta/Beta Ratio: A marker of Arousal or Cognitive Processing Capacity? *Appl. Psychophysiol. Biofeedback 2019 442* **44**, 123–129 (2019).

51. Morales, J. M., Ruiz-Rabelo, J. F., Diaz-Piedra, C. & Di Stasi, L. L. Detecting Mental Workload in Surgical Teams Using a Wearable Single-Channel Electroencephalographic Device. *J. Surg. Educ.* **76**, 1107–1115 (2019).

52. Foxe, J. J. & Snyder, A. C. The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Front. Psychol.* **2**, 154 (2011).

53. Louis, E. K. S. *et al.* The Normal EEG. (2016).

54. Rogala, J. *et al.* The do's and don'ts of neurofeedback training: A review of the controlled studies using healthy adults. *Front. Hum. Neurosci.* **10**, 1–12 (2016).

55. Palva, J. M., Palva, S. & Kaila, K. Phase synchrony among neuronal oscillations in the human cortex. *J. Neurosci.* **25**, 3962–3972 (2005).

56. Munia, T. T. K. & Aviyente, S. Time-Frequency Based Phase-Amplitude Coupling Measure For Neuronal Oscillations. *Sci. Rep.* **9**, 1–15 (2019).

57. Sur, S. & Sinha, V. K. Event-related potential: An overview. *Ind. Psychiatry J.* **18**, 70 (2009).

58. Lee, S. B. *et al.* Comparative analysis of features extracted from EEG spatial, spectral and temporal domains for binary and multiclass motor imagery classification. *Inf. Sci. (Ny).*

502, 190–200 (2019).

59. Jurcak, V., Tsuzuki, D. & Dan, I. 10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems. *Neuroimage* **34**, 1600–1611 (2007).

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

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

62. Marzbani, H., Marateb, H. R. & Mansourian, M. Methodological note: Neurofeedback: A comprehensive review on system design, methodology and clinical applications. *Basic Clin. Neurosci.* **7**, 143–158 (2016).

63. Schwartz, M. & Andrasik, F. Biofeedback: A Practitioner's Guide - 4th Edition - Mark S. Schwartz. https://www.routledge.com/Biofeedback-A-Practitioners-Guide/Schwartz-Andrasik/p/book/9781462531943 (2016).

64. Ghaziri, J. *et al.* Neurofeedback training induces changes in white and gray matter. *Clin. EEG Neurosci.* **44**, 265–272 (2013).

65. Gruzelier, J. H. Differential effects on mood of 12-15 (SMR) and 15-18 (beta1) Hz neurofeedback. *Int. J. Psychophysiol.* **93**, 112–115 (2014).

66. Williams, R. A. *Invited Commentary Neurofeedback System for Potential Orderly Care of Surgical Residents with Depression and Burnout*.

67. Christie, S. Individual Alpha Peak Frequency in Ice Hockey Shooting Performance. *Front. Psychol.* **8**, 762 (2017).

68. Faller, J., Cummings, J., Saproo, S. & Sajda, P. Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6482–6490 (2019).

69. Grober, E. D. & Jewett, M. A. S. The concept and trajectory of 'operative competence' in surgical training. *Can. J. Surg.* **49**, 238–240 (2006).

70. Krigolson, O. E., Williams, C. C., Norton, A., Hassall, C. D. & Colino, F. L. Choosing MUSE: Validation of a low-cost, portable EEG system for ERP research. *Front. Neurosci.* **11**, 1–10 (2017).

71. Bringsjord, S. & Govindarajulu, N. S. Artificial Intelligence (Stanford Encyclopedia of Philosophy). *Stanford Encyclopaedia of Philosophy* https://plato.stanford.edu/entries/artificial-intelligence/#LogiBaseAISomeSurgPoin (2018).

72. Turing, A. M. Computing machinery and intelligence. *Mind* 1–28 (2012) doi:10.1525/9780520318267-013.

73. O'Leary, D. E. GOOGLE'S Duplex: Pretending to be human. *Intell. Syst. Accounting, Financ. Manag.* **26**, 46–53 (2019).

74. Haenlein, M. & Kaplan, A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *Calif. Manage. Rev.* **61**, 5–14 (2019).

75. Feng, C. *et al.* A novel artificial intelligence-assisted triage tool to aid in the diagnosis of suspected COVID-19 pneumonia cases in fever clinics. *Ann. Transl. Med.* **9**, 201–201 (2021).

76. Nichols, J. A., Herbert Chan, H. W. & Baker, M. A. B. Machine learning: applications of

artificial intelligence to imaging and diagnosis. *Biophys. Rev.* **11**, 111–118 (2019).

77. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 1–10 (2018).

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

79. Coiera, E., Kocaballi, B., Halamaka, J. & Laranjo, L. The digital scribe. *npj Digit. Med.* **1**, 1–5 (2018).

80. Quiroz, J. C. *et al.* Challenges of developing a digital scribe to reduce clinical documentation burden. *npj Digit. Med.* **2**, (2019).

81. Dias, R. & Torkamani, A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med.* **11**, 1–12 (2019).

82. Perez, M. V. *et al.* Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. *N. Engl. J. Med.* **381**, 1909–1917 (2019).

83. Sutton, R. S. & Barto, A. G. Introduction. in *Reinforcement Learning: An Introduction* 3–24 (The MIT Press, 1998).

84. Albalate, A. & Minker, W. Semi-Supervised and Unsupervised Machine Learning: Novel Strategies. *Semi-Supervised Unsupervised Mach. Learn. Nov. Strateg.* (2013) doi:10.1002/9781118557693.

85. Wiggers, K. Supervised vs. unsupervised learning: What's the difference? | VentureBeat. *The Machine: Making Sense of AI* https://venturebeat.com/2021/04/22/supervised-vs-unsupervised-learning-whats-the-difference/ (2021).

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

87. Oladipupo, T. Types of Machine Learning Algorithms. in *New Advances in Machine Learning* (InTech, 2010). doi:10.5772/9385.

88. Coelho, L. P., Richert, W. & Brucher, M. Choosing the right model and learning algorithm. in *Building Machine Learning Systems with Python* (Packt Publishing, 2018).

89. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2**, 1–21 (2021).

90. Elshawi, R., Al-Mallah, M. H. & Sakr, S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med. Inform. Decis. Mak.* **19**, (2019).

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

92. Rajalakshmi, V., Narayanan, M., Ramkumar, M. & Scholar, U. G. An Exclusive Study on Unstructured Data Mining with Big Data An Exclusive Study on Unstructured Data Mining with Big Data. (2017).

93. Jia Deng *et al.* ImageNet: A large-scale hierarchical image database. 248–255 (2009) doi:10.1109/cvprw.2009.5206848.

94. Li, T., Sanjabi, M., Beirami, A. & Smith, V. Fair Resource Allocation in Federated Learning. 1–27 (2019).

95. Veale, M., Binns, R. & Edwards, L. Algorithms that remember: Model inversion attacks and data protection law. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **376**, (2018).

96. Nicholson Price, W. & Glenn Cohen, I. Privacy in the Age of Medical Big Data. *Nat.*

Med. **25**, 37–43 (2019).

97. Caruana, R. *et al.* Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **2015-Augus**, 1721–1730 (2015).

98. Buiten, M. C. Towards intelligent regulation of artificial intelligence. *Eur. J. Risk Regul.* **10**, 41–59 (2019).

99. Jiang, Y. *et al.* A Brief Review of Neural Networks Based Learning and Control and Their Applications for Robots. (2017) doi:10.1155/2017/1895897.

100. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).

101. Vaswani, A. *et al.* Attention Is All You Need. *Conf. Neural Inf. Process. Syst. (NIPS 2017)* **31**, (2017).

102. Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training Recurrent Neural Networks. *30th Int. Conf. Mach. Learn. ICML 2013* 2347–2355 (2012).

103. Nwankpa, C., Ijomah, W., Gachagan, A. & Marshall, S. Activation Functions: Comparison of trends in Practice and Research for Deep Learning. 1–20 (2018).

104. Wang, Q., Ma, Y., Zhao, K. & Tian, Y. A Comprehensive Survey of Loss Functions in Machine Learning. *Ann. Data Sci.* (2020) doi:10.1007/s40745-020-00253-5.

105. Ruder, S. An overview of gradient descent optimization algorithms. 1–14 (2016).

106. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6**, (2019).

107. Jiang, F. *et al.* Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology* vol. 2 230–243 (2017).

108. Fedus, W., Zoph, B. & Shazeer, N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. (2021).

109. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

110. Green, C., Lee, B., Amaresh, S. & Engels, D. W. A Comparative Study of Deep Learning Models for Network Intrusion Detection. *SMU Data Sci. Rev.* **1**, Article 8 (2018).

111. Lundberg, S. M. & Lee, S. I. *A unified approach to interpreting model predictions*. *Advances in Neural Information Processing Systems* vols 2017-Decem (2017).

112. Stiglic, G. *et al.* Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **10**, 1–13 (2020).

113. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (80-.).* **366**, 447–453 (2019).

114. Ahmad, M. A., Teredesai, A. & Eckert, C. Interpretable machine learning in healthcare. *Proc. - 2018 IEEE Int. Conf. Healthc. Informatics, ICHI 2018* 447 (2018) doi:10.1109/ICHI.2018.00095.

115. Du, M., Liu, N. & Hu, X. Techniques for interpretable machine learning. *Commun. ACM* **63**, 68–77 (2020).

116. Olden, J. D., Joy, M. K. & Death, R. G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecol. Modell.* **178**, 389–397 (2004).

117. Ribeiro, M. T., Singh, S. & Guestrin, C. Model-Agnostic Interpretability of Machine

Learning. (2016).

118. Carvalho, D. V., Pereira, E. M. & Cardoso, J. S. Machine learning interpretability: A survey on methods and metrics. *Electron.* **8**, 1–34 (2019).

119. Mirchi, N. *et al.* Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance. *Oper. Neurosurg.* **19**, 65–75 (2020).

120. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should i trust you?' Explaining the predictions of any classifier. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **13-17-Augu**, 1135–1144 (2016).

121. Marzbani, H., Marateb, H. R. & Mansourian, M. Methodological note: Neurofeedback: A comprehensive review on system design, methodology and clinical applications. *Basic Clin. Neurosci.* **7**, 143–158 (2016).

122. Ros, T. *et al.* Optimizing microsurgical skills with EEG neurofeedback. *BMC Neurosci.* **10**, (2009).

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

124. Silbergeld, D., Hebb, A. & Yang, T. The sub-pial resection technique for intrinsic tumor surgery. *Surg. Neurol. Int.* **2**, 180 (2011).

125. Ledwos, N. *et al.* Virtual reality anterior cervical discectomy and fusion simulation on the novel sim-ortho platform: Validation studies. *Oper. Neurosurg.* **20**, 74–82 (2021).

126. Sabbagh, A. J. *et al.* Roadmap for Developing Complex Virtual Reality Simulation Scenarios: Subpial Neurosurgical Tumor Resection Model. *World Neurosurg.* **139**, e220–e229 (2020).

127. Klimesch, W. EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews* vol. 29 169–195 (1999).

128. Park, J. L., Fairweather, M. M. & Donaldson, D. I. Making the case for mobile cognition: EEG and sports performance. *Neurosci. Biobehav. Rev.* **52**, 117–130 (2015).

129. Cheng, M. Y. *et al.* Sensorimotor rhythm neurofeedback enhances golf putting performance. *J. Sport Exerc. Psychol.* **37**, 626–636 (2015).

130. Kratzke, I. M. *et al.* Pilot Study Using Neurofeedback as a Tool to Reduce Surgical Resident Burnout. in *Journal of the American College of Surgeons* vol. 232 74–80 (Elsevier Inc., 2021).

131. Azarnoush, H. *et al.* Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. *Int. J. Comput. Assist. Radiol. Surg.* **10**, 603–618 (2015).

132. Al-Fahoum, A. S. & Al-Fraihat, A. A. Methods of EEG Signal Features Extraction Using Linear Analysis in Frequency and Time-Frequency Domains. *ISRN Neurosci.* **2014**, 1–7 (2014).

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

134. Bissonnette, V. *et al.* Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. *J. Bone Jt. Surg. - Am. Vol.* **101**, (2019).

135. Adebayo, J. & Kagal, L. Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models. (2016).

136. Alzhrani, G. *et al.* Proficiency performance benchmarks for removal of simulated brain

tumors using a virtual reality simulator neurotouch. *J. Surg. Educ.* **72**, 685–696 (2015).

137. Provins, K. A. & Cunliffe, P. The Relationship Between E.E.G. Activity and Handedness. *Cortex* **8**, 136–146 (1972).

138. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature medicine* vol. 26 1364–1374 (2020).

139. Winkler-Schwartz, A. *et al.* Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. *J. Surg. Educ.* **76**, 1681–1690 (2019).

140. Cao, X. H., Stojkovic, I. & Obradovic, Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC Bioinformatics* **17**, 359 (2016).

141. Yuan, J., Li, Y. M., Liu, C. L. & Zha, X. F. 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 LNCS 457–464 (Springer, Berlin, Heidelberg, 2010).

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

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

144. Roohi-Azizi, M., Azimi, L., Heysieattalab, S. & Aamidfar, M. Changes of the brain's bioelectrical activity in cognition, consciousness, and some mental disorders. *Medical Journal of the Islamic Republic of Iran* vol. 31 307–312 (2017).

145. Fried, I., Haggard, P., He, B. J. & Schurger, A. Volition and action in the human brain: Processes, pathologies, and reasons. *Journal of Neuroscience* vol. 37 10842–10847 (2017).

146. Park, J. L., Fairweather, M. M. & Donaldson, D. I. Making the case for mobile cognition: EEG and sports performance. *Neurosci. Biobehav. Rev.* **52**, 117–130 (2015).

147. Zhong, X. & Chen, J. J. Variations in the frequency and amplitude of resting-state EEG and fMRI signals in normal adults: The effects of age and sex. *bioRxiv* 2020.10.02.323840 (2020) doi:10.1101/2020.10.02.323840.

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

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

150. Reich, A. S. *et al.* Journal of Surgical Education Artificial Neural Network Approach to Competency-Based Training.

151. Shams, F. & Luise, M. Basics of coalitional games with applications to communications and networking. *Eurasip J. Wirel. Commun. Netw.* **2013**, 1–20 (2013).

152. Peters, H. The Shapley Value. in *Game Theory - A Multi-Leveled Approach* 241–242 (Springer-Verlag Berlin Heidelberg, 2008). doi:10.7551/mitpress/2954.003.0005.

153. Topol, E. J. Deep medicine : how artificial intelligence can make healthcare human again. 378.

154. Anwar, S. *et al.* A Game Player Expertise Level Classification System Using

Electroencephalography (EEG). *Appl. Sci.* **8**, 18 (2017).

155. Kingma, D. P. & Ba, J. L. *Adam: A method for stochastic optimization*. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015).

156. Porjesz, B. *et al.* Linkage disequilibrium between the beta frequency of the human EEG and a GABAA receptor gene locus. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3729–3733 (2002).

157. Kerson, C. *et al.* EEG Theta/Beta Ratio Calculations Differ Between Various EEG Neurofeedback and Assessment Software Packages: Clinical Interpretation. *Clin. EEG Neurosci.* **51**, 114–120 (2020).

158. Jirakittayakorn, N. & Wongsawat, Y. A Novel Insight of Effects of a 3-Hz Binaural Beat on Sleep Stages During Sleep. *Front. Hum. Neurosci.* **12**, 1–15 (2018).

159. Kwon, G. *et al.* Individual differences in oscillatory brain activity in response to varying attentional demands during a word recall and oculomotor dual task. *Front. Hum. Neurosci.* **9**, (2015).

160. Lega, B. C., Jacobs, J. & Kahana, M. Human hippocampal theta oscillations and the formation of episodic memories. *Hippocampus* **22**, 748–761 (2012).

161. Staudigl, T. & Hanslmayr, S. Theta oscillations at encoding mediate the context-dependent nature of human episodic memory. *Curr. Biol.* **23**, 1101–1106 (2013).

162. Fattinger, S., Kurth, S., Ringli, M., Jenni, O. G. & Huber, R. Theta waves in children's waking electroencephalogram resemble local aspects of sleep during wakefulness. *Sci. Rep.* **7**, 1–10 (2017).

163. Goyal, A. *et al.* Functionally distinct high and low theta oscillations in the human

hippocampus. *Nat. Commun.* **11**, 1–10 (2020).

164. Sigi Hale, T. *et al.* Atypical alpha asymmetry in adults with ADHD. *Neuropsychologia* **48**, 1–7 (2009).

165. Del Percio, C. *et al.* Football players do not show 'neural efficiency' in cortical activity related to visuospatial information processing during football scenes: An EEG mapping study. *Front. Psychol.* **10**, (2019).

166. Scholz, S., Schneider, S. L. & Rose, M. Differential effects of ongoing EEG beta and theta power on memory formation. *PLoS One* **12**, 1–18 (2017).

167. Abhang, P. A., Gawali, B. W. & Mehrotra, S. C. Technical Aspects of Brain Rhythms and Speech Parameters. *Introd. to EEG- Speech-Based Emot. Recognit.* 51–79 (2016) doi:10.1016/b978-0-12-804490-2.00003-8.

168. Arns, M., Conners, C. K. & Kraemer, H. C. A Decade of EEG Theta/Beta Ratio Research in ADHD: A Meta-Analysis. *J. Atten. Disord.* **17**, 374–383 (2013).