Multimodal Assessment of Cognitive and Motor Performance in Immersive Surgical Training

Beatriz Ribeiro, Gil Araújo, Adélio Vilaça, Luis Coelho, and Renato Magalhães

ISEP, Polytechnic of Porto, Porto, Portugal
ICBAS - Santo António Clinical Academic Center, Porto, Portugal
INESC TEC , Institute for Systems and Computer Engineering Technology and Science, Porto, Portugal
LabRP-CIR - Psychosocial Rehabilitation Laboratory - Center for Rehabilitation Research, ESS, Polytechnic of Porto, Porto, Portugal
School of Medicine and Biomedical Sciences (ICBAS), University of Porto, Porto, Portugal
Correspondence: rfm@ess.ipp.pt

DOI: *https://doi.org/10.17979/spu.23.c31*

Abstract: The integration of immersive technologies is reshaping surgical education. This study introduces a performance assessment system for immersive simulators in laparoscopic and arthroscopic training. By combining EEG and accelerometry, it monitors cognitive and motor functions in real time. A multidisciplinary framework enabled the definition of objective metrics for attention, stress, and motor control. A complete pipeline was developed for data acquisition, processing, and analysis, with results accessible via a web interface. Findings suggest that neurofeedback and motion tracking through VR controllers provide a reliable, objective alternative to traditional assessments, enhancing skill acquisition and enabling personalized, practice-based medical training.

1 Introduction

Surgical training, similar to other healthcare education fields, has evolved alongside technological developments. Until the 20th century, the Halstedian model dominated surgical training, emphasizing observation, practice, and experiential learning, where experience reduced errors and improved technical skills. However, acquiring such competencies in a clinical setting involves a learning curve where mistakes are inevitable, risking patient safety and care quality. Likewise, its reliance on unique individual experiences makes it an unstable and inconsistent teaching method (Viglialoro et al., 2021). This has encouraged a reconsideration of training models to enable surgical skill acquisition in more structured, progressive, and safer environments. Concepts like simulation and simulation-based training (SBT) have long been used to "develop competence and confidence in students [...]" in real-life situations. Today, this approach offers opportunities to integrate modern technology into surgical training. Historically, medical simulation has existed for centuries in rudimentary forms. For instance, during China's Song Dynasty (960–1279 AD), physicians used life-sized bronze statues to teach surface anatomy and acupuncture (Bienstock and Heuer, 2022). With this in mind, simulators enable repeated practice of techniques under identical conditions until proficiency, without putting patients at risk. This alone addresses a key limitation of the traditional model. As J.

Leonard Goldner once stated, "Your learning curve is the patient's suffering curve."(Hussain et al., 2025). As a result, surgical education has gradually moved further from the traditional Halstedian model and towards simulation, alongside immersive technologies included in extended reality (XR). Although its effectiveness in developing surgical competencies remains uncertain, studies highlight its benefits in skill acquisition. No consensus exists on the ideal training model, and most clinical studies rely on subjective pre- and post-training questionnaires, underscoring the need for more objective assessment methods. Consequently, a critical question arises: Can neurophysiological changes during skill acquisition be quantified? Recent studies incorporate educational theories and neurophysiological markers into surgical skills programs, showing that monitoring neurophysiological changes provides objective evidence of skill acquisition. Electroencephalography (EEG) can detect such changes with high temporal resolution, tracking fluctuations in alertness, attention, and cognitive workload. Brainwave activity is categorized into five frequency bands: beta (β, 13–30 Hz), alpha (α, 8–13 Hz), theta (ϑ, 4–8 Hz), delta (δ, 0.5–4 Hz), and gamma (γ, 30–45 Hz) (Coelli et al., 2015). Higher magnitude in a frequency band is associated with increased neuronal synchrony, known as event-related synchronisation (ERS), while reduced magnitude indicates event-related desynchronization (ERD). According to current neuroscience knowledge, relative band power can be linked to specific mental states, with higher power corresponding to ERS and lower to ERD. In surgical simulation, such analysis provides a promising foundation for developing neurofeedback-based training systems. Traditionally, technical surgical skills were evaluated by a supervisor who provided feedback directly in the operating room. The Objective Structured Assessment of Technical Skills (OSATS) emerged as a standardized tool, comprising seven criteria scored on a five-point scale. However, OSATS remains qualitative and reliant on human judgment, limiting objective, quantitative measurement of progress. To address these, motion capture and analysis of surgeons' hand movements have been explored as quantitative tools to complement OSATS and assess performance in simulation training. To assess technical performance in surgery can be assessed via movement smoothness using acceleration metrics, with tremor as a key factor. Physiological tremor can be defined as "a rhythmic and involuntary movement of a body part," present in all humans, typically with very low magnitude, inherent to normal neuromuscular function. In general, physiological hand tremor typically ranges from 6 to 15 Hz, while fine voluntary hand movements in microsurgery occur at 0.5–1 Hz (Veluvolu and Ang, 2011). This frequency distinction separates intentional movements from involuntary tremor, helping assess surgical manoeuvres and differentiate novice from advanced practitioners. Keeping that in mind, the purpose of this work is to develop a performance assessment system for an immersive surgical simulator. The system evaluates two core domains of performance simultaneously:

- Psychological performance: attention and stress via EEG (neurofeedback);
- Technical skills: motor control, movement precision, and tremor.

2 Methods

Both core domains, neurofeedback evaluation and technical skill assessment, are developed independently and later integrated into a final web application. This application offers a simple and accessible way to input training data and view results. Each domain requires different methodologies. The web application was developed using Flask as a microframework for its simplicity, alongside Python, CSS, HTML, and JavaScript.

2.1 Psychological Performance Assessment Methodology

To acquire EEG signals, it was first necessary to identify the most suitable device. The Muse S headband, developed by InteraXon, was selected for its ability to record biofeedback data.

To enable the reception, visualisation, recording, and transmission of raw EEG data, the Mind-Monitor application was leveraged. This tool provides essential features for the present work, such as real-time visualisation of raw signals, local storage, and continuous data transmission via the Open Sound Control (OSC) protocol. The collected data is stored in CSV format, from which selected features are transmitted in real time to enable efficient communication between MindMonitor and the developed web application. Each session begins with two recordings, a standard procedure in neurofeedback systems, such as Neurosity, and in the guidelines provided by the Muse developers. The first recording serves as calibration to establish baseline brain activity, followed by a training recording. Once EEG data is acquired and stored in a CSV file, preprocessing is initiated. This step prepares the data for metric and feature extraction at a later stage. Using the MNE-Python library (Gramfort et al., 2013; Larson et al., 2024), raw signals are processed to remove environmental artefacts, such as 50 Hz power line noise in Portugal, and biological artifacts, including heartbeats, eye movements (blinks), and high-frequency noise from muscle activity. Independent Component Analysis (ICA) is applied to identify and remove artifact-related components. ICA decomposes the recorded signal into statistically independent sources, such as eye movements, heartbeats, and muscle activity. It operates under the assumption that noise from independent sources can be separated and is statistically distinct from other components. The relationship between independent components (IC´s) and artifacts is then automatically analyzed and removed using Power Spectral Density (PSD), which quantifies the distribution of signal power across frequencies. For each IC, the PSD is computed in the 0–5 Hz range, where eye-blink artifacts typically exhibit the highest power. The component with the highest total power in this range is identified as the primary source of the artefact. For muscle-related artefacts, PSD analysis is performed within the 20–50 Hz range (UC San Diego, 2025). Once both the baseline and immersive training EEG signals are pre-processed and artefact-free, all conditions are set for extracting the relevant metrics and features. The results are intended to be visualised in two distinct ways. First, a textual interpretation provides an overall assessment of performance, highlighting significant changes in brainwave activity between baseline and training, indicating increases or decreases, and contextualising these changes within neuroscience. Second, the evolution of specific metrics, such as focus and stress, is visualised throughout the training session. Feature extraction is conducted using two complementary approaches, each aligned with a specific objective. The first approach addresses the first objective by analysing both baseline and training inputs as complete signals, providing an overall assessment of performance. The second approach targets the second objective, segmenting the same inputs into four-second windows with a one-second step to evaluate temporal changes. Both the full signals and the extracted segments are then analysed using PSD. This analysis is performed globally across all channels and locally on specific regions, namely the frontal region (AF8 and AF7) and the parietal region (TP10 and TP9). The main goal is to evaluate the energy contribution of the frequency bands of interest relative to the total EEG signal power: Beta (β), Alpha (α), Theta (ϑ), Delta (δ), and Gamma (γ) (Figure 1).

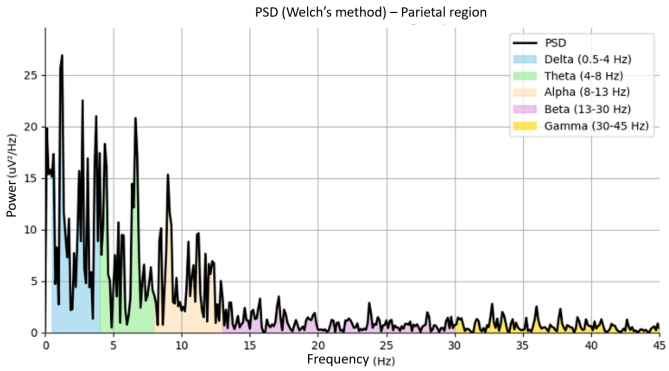

Figure 1: Accumulated power in the frequency bands of interest in the parietal region for stimulus data (Training).

To address the first objective, relative power values are calculated for each frequency band (β, α, ϑ, δ, and γ) and each region of interest (global, frontal, and parietal), across both baseline and training data for thirty immersive samples. These values are then compared using the Wilcoxon test, which addresses the question: "Is there a statistically significant difference between frequency X during the stimulus and the same frequency in the baseline data for the same region?". Based on the three possible outcomes, increase, decrease, or no significant change, all possible result combinations are generated, along with their respective interpretative descriptions in textual format. To address the second objective, which concerns the temporal evolution of metrics, two physiological indices are analysed from the temporal segments: the focus index and the stress index. The focus index (Coelli et al., 2015) is calculated from Equation 22.1, and the stress index is obtained similarly from Equation 22.2.

$$\text{Focus Index} = \frac{\beta}{\alpha + \theta} \qquad (22.1) \qquad\qquad \text{Stress Index} = \frac{\beta}{\alpha} \qquad (22.2)$$

For each temporal segment, the index is computed, smoothed with a 0.2 factor and then normalised. Following smoothing, the time series is normalised according to Equation 22.3.

$$E_{\text{Norm}} = \frac{E_{\text{Smooth}} - E_{\text{min}}}{E_{\text{max}} + E_{\text{min}}} \times 100 \qquad (22.3)$$

This methodological approach follows the procedures outlined by Hassib et al. (2017). The resulting data is then used to generate plots that illustrate the temporal evolution of each metric in the graphical interface.

2.2 Technical Performance Assessment Methodology

Motor data was acquired using the VR controllers of the MetaQuest 3, the same devices intended to perform the immersive simulator exercises. To leverage these built-in functionalities, a Unity component was developed to record selected controller data. Unity allows the creation of a GameObject with a single component containing a C# script that captures controller sensor data and stores it in a CSV file. The OVRInput library provides direct access to measurements like controller position and velocity. During the surgical exercise, motor data is continuously recorded from the start until the session ends, after which it is stored in the headset's internal memory. Upon acquisition, the data is processed to extract insights on motor control during immersive training. The analysis focuses on acceleration values along each spatial axis and the timestamp for each sample. The signal filtering phase focuses the analysis on

relevant frequency ranges. As described earlier, two ranges of interest are considered: 0.5–1 Hz for fine voluntary movements and 6–15 Hz for physiological tremor. To isolate these components, band-pass filtering is applied to remove noise and irrelevant frequencies. Fourth-order Butterworth band-pass filters are implemented for each acceleration axis (X, Y, and Z). The two filters, corresponding to the ranges of interest, are combined into a single filter, which is then applied to the signal. As in the previous methodology, results are visualised in two ways. The first extracts global statistics, namely mean and maximum acceleration for each axis (X, Y, Z) of the filtered signal. Performance is assessed by comparing these values with reference ranges from experienced users. Depending on whether results fall within, below, or above the range, a descriptive evaluation is generated for each axis, indicating axes with greater or lesser control. The second approach focuses on temporal analysis, segmenting the filtered signal into four-second windows with one-second overlap. In each window, PSD estimation via Welch's method is used to quantify power from the area under the curve, distinguishing voluntary movements from physiological tremor. This provides an objective characterisation of motor control based on energy metrics (Figure 2).

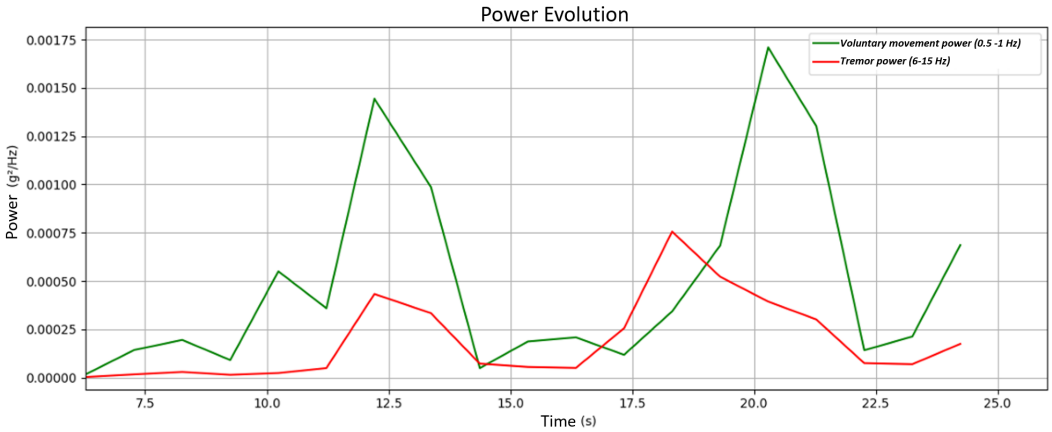

Figure 2: Temporal evolution of spectral power for each component of interest (Voluntary Movement and Physiological Tremor).

These values are used to estimate the tremor index, as defined in Equation 22.4. This index is calculated for each temporal segment.

$$\text{Tremor Index} = \frac{\text{Tremor Power}}{\text{Tremor Power} + \text{Voluntary Power}} \times 100 \tag{22.4}$$

For each temporal window, a percentage value is obtained representing the relative contribution of tremor compared to voluntary movement. This produces a function that tracks the evolution of tremor dominance over fine motor control across time.

3 Results

The final outcome is a web application integrating all implemented algorithms, supporting file uploads and presenting results through an intuitive interface. The homepage provides recording guidelines and feedback on signal quality. Results are displayed in two sections: the default "Summary," with textual outcomes, and the "Graphs" tab, showing temporal evolution, trend interpretation, and average percentages. Figure 3 illustrates the first case.

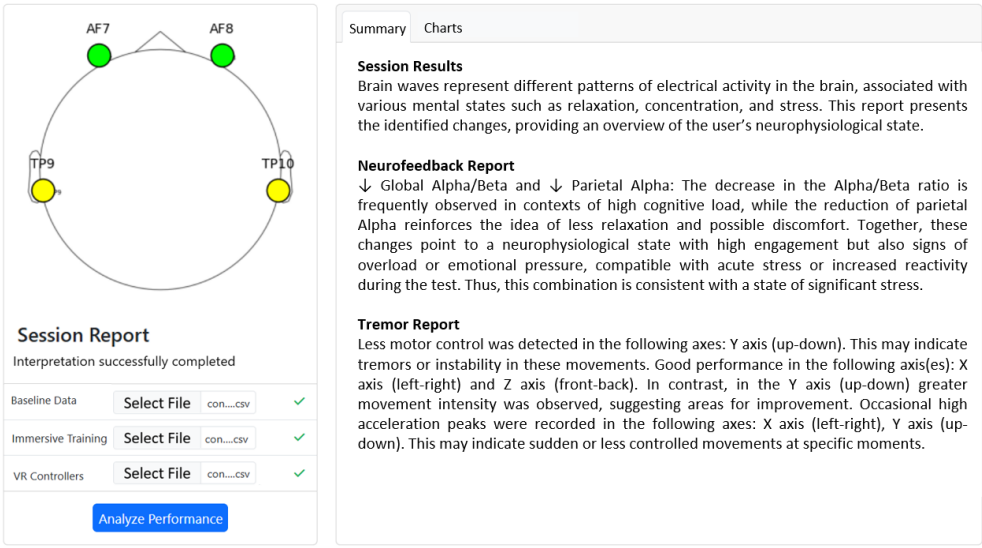

Figure 3: Results displayed in the "Summary" tab of the Web application.

The WebApp monitors signal quality in real time via OSC with MindMonitor. A head diagram shows sensor locations colour-coded as red (poor), yellow (moderate), green (good), or gray (no connection). Data is automatically integrated after recording, with an alert confirming transfer and disabling the upload field. Regarding the graphical component, Figure 4 illustrates the temporal evolution of neurophysiological metrics such as stress and focus. This specific case represents a simulated stress condition, however, these results do not validate the underlying algorithms scientifically, as proper validation would require a clinical trial.

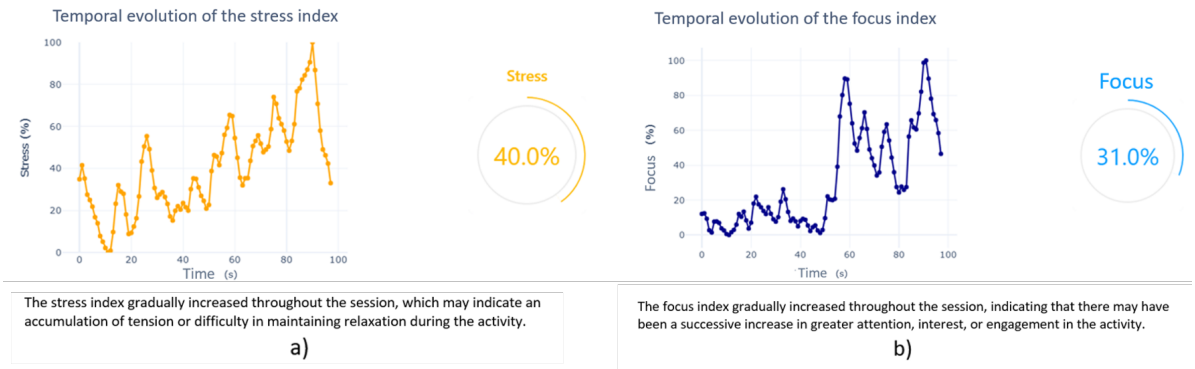

Figure 4: Graphs representing a stress condition: a) Temporal progression of stress. b) Temporal progression of focus.

Regarding motor performance, Figure 5 shows the temporal evolution of hand acceleration. The test consisted of a controlled voluntary movement along the X-axis (left–right), while slight, progressively more frequent tremors were deliberately introduced on the Z-axis (forward–backwards) and Y-axis (up–down).

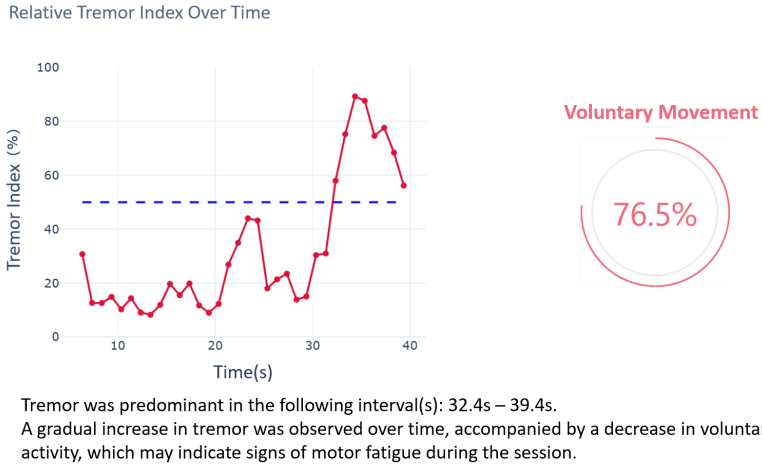

Figure 5: Motor Control Results.

4 Conclusion

This project focused on establishing a solid foundation for developing performance assessment systems grounded in current scientific evidence. EEG data alone, however, is insufficient. Integrating these biometric signals with objective, time-resolved metrics of surgical performance within the simulator is essential. Recent research trends emphasise the use of machine learning for such analyses, but this requires reliable datasets, ideally built from clinical trials, which were not achievable within the project's timeframe. While the prototype is considered complete, it lays the groundwork for future systems that can drive innovation in healthcare education.

Bibliography

J. Bienstock and A. Heuer. A review on the evolution of simulation-based training to help build a safer future. *Medicine*, 101(25):e29503, 2022. doi: 10.1097/MD.0000000000029503. URL *https://doi.org/10.1097/MD.0000000000029503*.

S. Coelli, R. Sclocco, R. Barbieri, G. Reni, C. Zucca, and A. M. Bianchi. Eeg-based index for engagement level monitoring during sustained attention. *Conference proceedings: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015: 1512–1515, 2015. doi: 10.1109/EMBC.2015.7318658.

A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen. Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, 7, 2013. doi: 10.3389/fnins.2013.00267. URL *https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2013.00267*.

M. Hassib, M. Khamis, S. Friedl, S. Schneegass, and F. Alt. Brainatwork: Logging cognitive engagement and tasks in the workplace using electroencephalography. *Conference proceedings: Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia (MUM '17)*, 2017:305–310, 2017. doi: 10.1145/3152832.3152865. URL *https://doi.org/10.1145/3152832.3152865*. Conference location: Stuttgart, Germany.

Z. B. Hussain, G. E. Garrigues, and E. R. Wagner. The use of extended reality to improve upper extremity surgical training and shorten the learning curve. *Hand Clinics*, 41(2):197–206, 2025. doi: 10.1016/j.hcl.2024.12.008. URL *https://doi.org/10.1016/j.hcl.2024.12.008*.

E. Larson, A. Gramfort, D. A. Engemann, J. Leppakangas, C. Brodbeck, M. Jas, T. L. Brooks, J. Sassenhagen, D. Strohmeier, M. Hämäläinen, and et al. Mne-python, Dec 2024. URL *https://doi.org/10.5281/zenodo.14519545*. [Software].

UC San Diego. Iclabel tutorial: Eeg independent component labeling. *https://labeling.ucsd.edu/tutorial/labels*, 2025. [Online; accessed 12-Jul-2025].

K. C. Veluvolu and W. T. Ang. Estimation of physiological tremor from accelerometers for real-time applications. *Sensors*, 11(3):3020–3036, 2011. doi: 10.3390/s110303020. URL *https://www.mdpi.com/1424-8220/11/3/3020*.

R. M. Viglialoro, S. Condino, G. Turini, M. Carbone, V. Ferrari, and M. Gesi. Augmented reality, mixed reality, and hybrid approach in healthcare simulation: A systematic review. *Applied Sciences*, 11(5):2338, 2021. doi: 10.3390/app11052338. URL *https://www.mdpi.com/2076-3417/11/5/2338*.