A multimodal approach combining tool-pressure and EEG features for laparoscopic skill classification using machine learning

Sebahat Selin Sahin
Ankara University
Cagri Zengin
Ankara University
Hasan Onur Keles
hokeles@ankara.edu.tr

Ankara University https://orcid.org/0000-0001-8493-2582

Research Article

Keywords: Machine Learning , NASA-TLX, Laparoscopy, Surgical Education

Posted Date: November 3rd, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7996713/v1

Additional Declarations: The authors declare no competing interests.

Abstract

Aims: Laparoscopic skill assessment traditionally relies on subjective evaluation, which lacks objectivity and consistency. Developing automated multimodal approaches that integrate tool-pressure and neural data can improve the reliability and scalability of skill assessment. Therefore, our objectives were to: (1) integrate a pressure-sensing unit into existing box-trainer simulators and laparoscopic tools to investigate and validate tool pressure features as objective indicators of laparoscopic skill; (2) combine EEG-derived power and phase-locking value (PLV) features with tool pressure data to evaluate the classification performance of different machine learning models.

Methods: Tool-pressure, EEG, and ECG data, along with task completion time, error counts, and NASA-TLX workload scores, were collected from 10 surgeons and 13 inexperienced students performing a peg-transfer laparoscopic task. The pressure sensors were integrated into the right and left laparoscopic graspers. EEG features were extracted from the four different frequency bands using both power spectral density (PSD) and phase-locking value (PLV) measures. Three machine learning models (Random Forest Classifier (RFC), Gaussian Process Classifier (GPC), and AdaBoost Classifier (ABC)) were used to classify participants into surgeon and non-experienced groups based on these multimodal features

Results: The findings showed that right–left pressure asymmetry was a more reliable indicator of surgical expertise compared to other tool-pressure metrics. Using only the asymmetry feature, RFC achieved up to 78% classification accuracy. The highest performance was obtained when combining theta-band power features with pressure asymmetry during the task, where both RFC and ABC reached 86% accuracy (F1 = 0.83; AUC = 0.92 for RFC and 0.85 for ABC). Theta-band findings support its relevance for surgical skill assessment.

Conclusion: Overall, this multimodal approach combining psychomotor and neurophysiological measures enhances the objectivity of surgical skill evaluation and may support real-time feedback systems for laparoscopic training.

Introduction

Laparoscopy is widely utilized in various surgical procedures today, providing significant benefits such as faster patient recovery, reduced hospital stays, and decreased postoperative pain. However, performing laparoscopic surgery presents unique challenges, demanding surgeons to develop enhanced perceptual and cognitive skills (Bonrath et al., 2013; Buia et al., 2015; Omurtag et al., 2025). Limited visual access and reduced tactile feedback require surgeons to master novel techniques characterized by distinct learning curves, necessitating advanced training methods beyond conventional apprenticeship models. Surgical skill level significantly influences the reduction of medical errors and complications(D. et al., 2025; Healey et al., 2002; Keles et al., 2021)

Objective assessment of surgical skills is crucial in laparoscopic training, as accurate feedback is essential for performance improvement. Recent technological advancements have facilitated more precise and continuous assessments of surgical performance (Shafiei, Shadpour, Mohler, Attwood, et al., 2023; Soangra, Sivakumar, et al., 2022). Psychophysiological measures offer objective and continuous evaluations, overcoming limitations associated with traditional skill assessment methods, which typically rely on behavioural metrics and subjective surveys (Darzi et al., 2001; Omurtag et al., 2025; Zakeri et al., 2020)

Physiological sensor-based technologies, such as electrocardiography (ECG) (Farah et al., 2024; Huaulmé et al., 2025; The et al., 2020), electromyography (EMG) (Soangra, Jiang, et al., 2022; Soto Rodriguez et al., 2023), electroencephalography (EEG) (Manabe et al., 2023; Omurtag et al., 2025) and functional near-infrared spectroscopy (fNIRS) (Gao et al., 2021; Keles et al., 2021; Nemani et al., 2018, 2019) have been employed to assess surgical skills and performance. These measures have revealed significant differences in psychomotor performance between surgeons of varying expertise, providing complementary insights into trainee development. In addition, motion analysis using video cameras (Franco-González et al., 2024; Zia et al., 2018) and eye tracking (Oh & Lau, 2024; Tien et al., 2014) have also been applied for surgical skill assessment.

Surgeon–tool interactions, particularly those involving laparoscopic graspers and similar instruments, are critical to overall surgical performance(Alleblas et al., 2016). To objectively assess these interactions, force sensors have been embedded into various surgical tools, including forceps, graspers, and gloves(Araki et al., 2017; Brown et al., 2017; Rafii-Tari et al., 2017; Sugiyama et al., 2018). Previous studies have demonstrated that novice surgeons tend to exert higher and more variable interaction forces than experts, who apply smoother and more controlled force patterns(Golahmadi et al., 2021; Sugiyama et al., 2018). These force dynamics serve as informative indicators of technical proficiency, reflecting the surgeon's ability to regulate grip strength, minimize abrupt load changes, and coordinate bimanual actions effectively. The assessment of bimanual motor skills is critical in surgical training, as these skills involve a complex interplay of cognitive, decision-making, and psychomotor components(Gao et al., 2021; Nemani et al., 2018). Our multimodal approach enables a comprehensive evaluation of these integrated processes, overcoming the limitations of conventional assessments that often fail to capture the full spectrum of surgical expertise.

In addition, the use of tool pressure features in surgical skill assessment remains underexplored. Prior research has largely focused on force magnitude and variability, often neglecting the coordination between the two hands, which may offer deeper insight into motor control and surgical expertise. Among various tool-pressure features, we specifically examine right–left pressure asymmetry, defined as the imbalance of forces applied by the surgeon's dominant and non-dominant hands during laparoscopic manipulation. This metric is particularly relevant in laparoscopic procedures that demand precise coordination and fine motor control. Previous studies using Force-Sensing Bipolar Forceps force-sensing surgical gloves, and laparoscopic graspers (Araki et al., 2017; Sugiyama et al., 2018) have shown the feasibility of pressure-based skill assessment, but asymmetry has rarely been quantified explicitly.

To address this gap, the present study introduces a tool-pressure asymmetry metric and integrates it with neurophysiological signals to develop a multimodal framework for surgical expertise classification. While machine learning techniques have been applied to surgical skill assessment using different modalities and data type (Dockter et al., 2017; Ebina et al., 2024; Gao et al., 2020; Natheir et al., 2023; Power et al., 2025; Shafiei, Shadpour, Mohler, Attwood, et al., 2023), the combination of EEG and tool pressure data remains largely unexplored. To the best of our knowledge, no prior studies have systematically investigated this integration for classifying surgical expertise.

This study aims to bridge that gap by combining sensor-based behavioral features with EEG-based neurophysiological data, using machine learning models to classify the surgical skill. By examining the relationship between bimanual force coordination and corresponding brain activity patterns, we propose a multimodal, objective approach to skill assessment that may significantly enhance training and evaluation frameworks in minimally invasive surgery.

Methods

Participants

Data from 10 surgeons with varying laparoscopic experience (range 2–120 procedures; mean 26.7 ± 33.0, median 15) and 13 students with no prior laparoscopy experience were included. All participants provided written informed consent at the beginning of the experimental session. This study was approved by the Ankara University Human Research Ethics Committee and was conducted at Neuroscience and Neurotechnology Center of Excellence. All participants were provided with essential information regarding the purpose of the study, associated risks, and potential benefits to ensure that they could give their voluntary and informed consent to participate.

Experimental Design and Data Collection

An overview of the experimental design is provided in Fig. 1. Participants performed standard training tasks including peg transfer using a laparoscopic trainer box. Data were recorded including time to completion, error rate. After subjects completed each task, they completed the NASA task load index (NASA-TLX) questionnaire. At the beginning of the experiment, two 2-minutes-long videos that demonstrates the tasks were shown on a computer screen to the subject. After the video session the 15 min training was repeated by each subject for participants.

In this study, we use the peg transfer task of the FLS (Fundamentals of Laparoscopic Surgery) This task is an essential exercise that forms the foundation of laparoscopic surgical skills and aims to improve surgeons' ability to objects using their hands. It is designed to enhance surgeons' capability to perform precise and coordinated movements with laparoscopic instruments. In this task, participants were required to transfer three rings from one location to another to complete the task. Each ring was picked up with the left tool (grasper), transferred mid-air to the right tool (grasper), and then placed in a new

location with the right tool. Dropping any of the rings onto the platform during the transfer was considered an error. Participants began the task after a two-minute rest period.

As part of the experimental setup, pressure sensors built by the research team were integrated into the right and left grasper tools. These sensors enabled precise measurement of the pressure applied by participants' thumbs on the tools while completing the task. We used a series of metrics aimed at capturing distinctions in skill level and training(Omurtag et al., 2019). Active time segments were characterized as those periods when pressure exceeded a specified threshold, indicating engagement. For measuring bilateral coordination, we used the Right-Left Overlap metric, defined as the duration in which pressures from both the right and left sides were simultaneously active. Additionally, we used the Right-Left Asymmetry metric, which is calculated as |R-L| / (R + L). Here, R and L represent the time-averaged amplitudes of the right and left tool pressures, allowing us to quantify any imbalance between sides. These metrics provide robust insight, and we confirmed that variations in threshold settings had minimal impact on their reliability. Furthermore, we verified that the calculated measures remained stable across a wide range of threshold values, ensuring consistent results regardless of threshold adjustments.

In this study, sensor data from the tool device was collected simultaneously with electroencephalography (EEG) and electrocardiography (ECG) recordings throughout the experiment Fig. 2. To measure brain activity, the 8-channel version of the Mentalab Portable EEG device was used. One channel was assigned for recording ECG data, while the remaining seven channels were utilized to gather EEG data pertaining to brain activities. The brain activity data were acquired from electrodes positioned at Fz, F3, F4, C3, C4, P3, and P4. This electrode arrangement facilitated the examination of neural activities across various cortical regions, including the frontal lobe (associated with decision-making and attention), the central lobe (involved in motor activities), and the parietal lobe (responsible for sensory processing). Additionally, heart rate variability (HRV) was quantified as the standard deviation of successive interbeat intervals (SDNN), derived from the time differences between consecutive R-peaks in the ECG signa

Participants completed the NASA Task Load Index (NASA-TLX) questionnaire, which was developed by the Human Performance Group at NASA Ames Research Center, after completing the experiment. This multidimensional assessment tool for mental workload comprises six categories, each rated on a 20-point visual analog scale. The initial five categories—mental demand, physical demand, temporal demand, effort, and frustration—range from ‘low’ to ‘high’. The sixth category evaluates participants' performance, ranging from 'satisfied' to 'failure'. The first three subscales relate to the demands placed on the participant, while the last three subscales pertain to the participant's interaction with the task.

Data Analysis

Behavioral and Performance Data

Completion time for the laparoscopic task was recorded for each participant. The number of errors, defined as instances where a peg was dropped was also quantified. Participants completed the NASA Task Load Index (NASA-TLX) to evaluate six workload dimensions: mental, physical, and temporal demands, effort, frustration, and perceived performance. Each dimension was rated on a scale from 1 to 20. The overall mental workload score was obtained by summing the individual criterion scores.

EEG Processing

The raw EEG data underwent a series of pre-processing steps to remove non-brain signals and preserve the brain signal for further analysis. Segments of data containing high-frequency, high-amplitude waves, likely originating from gross body movements during EEG recording, were identified and deleted using a sliding window. Additionally, electrodes exhibiting kurtosis values exceeding 5 were considered invalid channels and removed from the data. The signals were then band-pass filtered between 0.16 Hz and 40 Hz to reduce slow drifts and high-frequency artifacts, and subsequently down sampled to 200 Hz to optimise computational and storage requirements. EEG-based features were computed from the frequency band power (PSD), phase locking value (PLV), Initially, the spectrogram was calculated using short-time Fourier transform method with windows of 1 s and half window size overlapping and frequency resolution of 1 Hz. The power was calculated in eight frequency bands each with a width of 4 Hz in the range 0 to 32 Hz. The ranges are referred to by their conventional labels: delta (0–4 Hz), theta (4–8), alpha (8–12) and beta (12–30Hz). EEG frequency band power for each epoch was extracted by integration of the corresponding power over each frequency band. We imposed the 32 Hz cutoff since higher frequencies in scalp EEG are generally not considered informative about cortical activity (Muthukumaraswamy, 2013). PLV is a measure of phase synchrony between two distinct neuronal populations, which is computed between two selected EEG electrodes as an estimate the inter-area synchrony (Vinck et al., 2011). PLV was estimated between electrode pairs that were selected to assess the synchrony. PLV was computed for each narrow (1 Hz wide) band, then averaged across four bands of interest ([0–4], [4–8], [8–12], [12–30] Hz). We used 7 EEG channel (21 pairs) representing connections: FZ–F4, FZ–C3, FZ–C4, FZ–P4, FZ–F3, FZ–P3, F4–C3, F4–C4, F4–P4, F4–F3, F4–P3, C3–C4, C3–P4, C3–F3, C3–P3, C4–P4, C4–F3, C4–P3, P4–F3, P4–P3, and F3–P3.

Classification

EEG and tool-pressure data were used to train multiple machine learning algorithms to distinguish between surgeon and control groups across different frequency bands. Model performance was evaluated using 3-fold, 5-fold, and 7-fold cross-validation schemes. Fewer folds may not provide sufficient data partitions to yield stable estimates, whereas an excessive number of folds may reduce the number of samples per partition, leading to unreliable training. As the results were comparable across configurations, a 5-fold stratified cross-validation scheme was adopted as a practical compromise. This approach ensured that each subset served once as the validation set while maintaining class balance. Leave-one-out cross-validation was not considered, as it produces a single

partition and may yield biased performance estimates. All features were standardized by centering and scaling according to the mean and standard deviation of each column. Given the limited sample size, feature selection was performed using the *SelectKBest* method with chi-square and mutual information criteria to evaluate the statistical dependence between features and class labels. The top five ranked features were retained to improve model accuracy and reduce computational cost. Hyperparameter optimization was conducted using *GridSearchCV*, which systematically explored predefined parameter combinations through cross-validation. For RFC, the parameters *n_estimators* and *criterion* were optimized; for ABC, the number of estimators and learning rate were tuned; and for GPC, the kernel function and optimizer settings were adjusted. Model performance was assessed using accuracy and F1-score metrics, and the final results were reported as the mean of cross-validation outcomes, ensuring robust and unbiased performance estimation.

Statistical analysis

When conducting a regression analysis comparing 2 numerical variables, linear fit with analysis of variance was used. The descriptive results comparing two groups, such as completion time v asymmetry contained non-paired data. In order to assess the statistical significance of the difference between two groups of non-paired results, we used the non-parametric Kolmogorov test. We did not utilize null-hypotheses whose rejection would have required corrections for multiple comparisons or false discovery.

RESULTS

Twenty-three participants (10 surgeons, 13 non-experienced students; age: M = 27.3 years, SD = 4.9) took part in the study. Each participant first completed a baseline rest period (R1), followed by the peg transfer task (T1), a laparoscopic training task widely used to assess bimanual coordination (Fig. 1A). EEG and tool pressure data were continuously and simultaneously recorded throughout the task. During the task, the completion time and the error was recorded. After task completion, participants completed the NASA-TLX questionnaire to evaluate their perceived cognitive workload. Our analyses proceeded in four main steps. First, we examined participants' behavioral performance metrics, including task completion time and error rates (Fig. 3). Second, we compared sensor-derived force metrics between groups (Fig. 4).*Third, we explored the relationship between behavioral performance, sensor-based force metrics, and self-reported workload subscales from the NASA-TLX (Fig. 5). Finally, we evaluated and compared multiple machine learning classification models sensor-derived features to predict participants' skill level groupings (Fig. 6).

Behavioral and Performance Results

We first compared behavioral performance, heart rate variability (HRV), and subjective workload between participants with laparoscopic experience (Surg) and those without (Non-Exp). As shown in Figure 3, the number of errors during the peg transfer task was significantly lower in the Surg group compared to the Non-Exp group ($p < 0.05$), indicating higher task accuracy among surgeons. Although the completion

time tended to be shorter in the Surg group and HRV values were slightly higher, these differences did not reach statistical significance. Similarly, no significant difference was found between the groups in terms of total NASA-TLX scores, suggesting comparable levels of perceived workload across participants regardless of experience level.

Force Sensor Data Results

We first evaluated a comprehensive set of time-domain features derived from the force signals recorded during the peg transfer task. These included maximum slope, peak and average force amplitudes, force variability (standard deviation), force time integral (FTI), number of peaks, time to peak, and smoothness. None of these features revealed statistically significant differences between the groups, indicating similar temporal and magnitude-based motor characteristics between surgeons and non-experienced participants.

Next, we analysed additional coordination-related force metrics: the asymmetry index, right–left (R–L) overlap, and active force duration (Figure 4). The asymmetry index, which quantifies the normalized pressure imbalance between the left and right tool handles, showed a trend toward significance (p = 0.06). The Non-Experienced group had a higher mean asymmetry (0.087 ± 0.052) compared to the Surgeon group (0.030 ± 0.031), suggesting that surgeons performed the task with more balanced bimanual coordination. The R–L overlap, which measures the duration of simultaneous use of both hands (in milliseconds), was numerically *higher in the Surgeon group (183.69 ± 55.94 ms) than in the Non-Experienced group (168.68 ± 49.94 ms), though this difference was not statistically significant. This suggests that while both groups exhibited similar bilateral engagement times, surgeons may have coordinated hand movements more consistently across the task. Lastly, the active force duration, indicating the total number of samples where force exceeded a predefined threshold, did not differ significantly between groups, supporting the conclusion that overall force application levels were comparable.

Inter-variable Relationship Assessment

To explore the relationship between task performance and perceived workload, we performed linear regression analyses between NASA-TLX subscale scores and asymmetry index (Fig. 5). Regarding coordination, the asymmetry index was significantly associated with physical demand ($p < 0.05$), suggesting that participants who exhibited more imbalanced tool use reported higher levels of physical workload. No significant associations were found between asymmetry and other subscales, although trends were observed for mental demand and temporal demand. Performance, Effort, and Frustration subscales did not exhibit significant linear relationships with asymmetry.

Classification Results

Figure 6 shows the classification performance evaluated using Power Spectral Density (PSD) and Phase Locking Value (PLV) features across four EEG frequency bands (delta, theta, alpha, and beta) under both task and rest conditions. Three classifiers (Random Forest (RFC), Gaussian Process Classifier (GPC),

and AdaBoost Classifier (ABC)) were employed. Three feature sets were evaluated: (1) asymmetry-only, (2) Asymmetry + EEG Power Spectral Density (PSD), and (3) Asymmetry + EEG Phase Locking Value (PLV)

When using only the asymmetry index as the input feature, classification accuracies varied across models. The RFC achieved an accuracy of 78%, followed closely by the AdaBoost Classifier (ABC) with 79%, while the Gaussian Process Classifier (GPC) showed a substantially lower performance with 60% accuracy. These results suggest that asymmetry-based motor coordination features alone provide a moderate discriminative signal for classifying expertise level, particularly when classifiers such as RFC and ABC are employed.

Table 1 presents the classification performance of three machine learning models (RFC, GPC, and ABC) trained using combinations of EEG power band features (δ, θ, α, β) and motor asymmetry metrics under task condition. The models were evaluated using standard metrics: accuracy, F1 score, precision and AUC score. The highest classification performance was observed when combining theta-band power with asymmetry features during the task condition, particularly with RFC and ABC classifiers. Both models achieved an accuracy of 0.86, F1 scores of 0.83, and AUC scores of 0.92 (RFC) and 0.85 (ABC), respectively. Among classifiers, the AdaBoost Classifier (ABC) consistently demonstrated higher and more stable performance, particularly in terms of F1 and AUC scores. In contrast, the Gaussian Process Classifier (GPC) exhibited lower accuracy and generalization, especially in the beta-band under task conditions (AUC = 0.45). Models trained with task-related EEG features yielded higher performance compared to those using resting-state data. The performance during task conditions was generally lower for PLV-based features compared to spectral power-based combinations. However, the beta band during task still produced relatively good performance with RFC (Accuracy = 0.76, AUC = 0.83) and ABC (AUC = 0.88), suggesting some utility of phase-based features even under active task conditions.

Table 1. Performance comparison of PSD and PLV asymmetry features across classifiers.

Features	Classifier	Accuracy	F1 Score	Precision	AUC Score
Asymmetry + δ Band (PSD)	RFC	0.72	0.73	0.70	0.73
	GPC	0.57	0.61	0.60	0.60
	ABC	0.72	0.70	0.80	0.80
Asymmetry + θ Band (PSD)	RFC	0.86	0.83	0.93	0.92
	GPC	0.80	0.79	0.83	0.79
	ABC	0.86	0.83	0.93	0.85
Asymmetry + α Band (PSD)	RFC	0.72	0.60	0.70	0.88
	GPC	0.58	0.51	0.60	0.77
	ABC	0.81	0.76	0.93	0.85
Asymmetry + β Band (PSD)	RFC	0.58	0.57	0.60	0.62
	GPC	0.58	0.41	0.48	0.45
	ABC	0.82	0.77	0.90	0.82
Asymmetry + δ Band (PLV)	RFC	0.72	0.67	0.90	0.82
	GPC	0.67	0.60	0.80	0.68
	ABC	0.82	0.77	0.90	0.82
Asymmetry + θ Band (PLV)	RFC	0.77	0.75	0.83	0.73
	GPC	0.67	0.69	0.67	0.70
	ABC	0.82	0.77	0.90	0.82
Asymmetry + α Band (PLV)	RFC	0.72	0.67	0.80	0.79
	GPC	0.71	0.71	0.69	0.85
	ABC	0.71	0.72	0.70	0.74
Asymmetry + β Band (PLV)	RFC	0.76	0.75	0.83	0.83
	GPC	0.58	0.59	0.53	0.67
	ABC	0.72	0.68	0.80	0.88

Confusion Matrix Analysis

Figure 7 displays confusion matrices corresponding to the classification results for each frequency band and classifier. Each matrix illustrates the predicted versus actual group labels (Non-Exp vs Surgeon). ABC classifiers (left column) generally achieved balanced classification, effectively distinguishing both control and surgeon groups. In the Alpha-ABC model, each class was correctly identified in 33.3% of cases. GPC classifiers (middle column) frequently predicted all samples as a single class, most often the Non-Exp group. This indicates a strong bias and inability to generalize, as seen in Alpha-GPC and Delta-GPC, where the surgeon class was not recognized at all. RFC classifiers (right column) tended to show high accuracy for the surgeon group but occasionally misclassified control participants. In Theta-RFC, the surgeon group was correctly identified in 66.7% of cases, whereas control accuracy dropped to 33.3%

Figure 8 presents the confusion matrices obtained from the combined phase-locking value (PLV) and asymmetry features across the alpha, beta, theta, and delta bands. Three machine learning classifiers (ABC, GPC and RFC) were trained to distinguish surgeons from non-experienced participants. Overall, the

AdaBoost model demonstrated the most balanced performance across all frequency bands, achieving higher true-positive rates for the surgeon group, particularly in the alpha and theta bands (up to 50%). The RFC model showed moderate discrimination ability, with relatively consistent classification of surgeons (≈ 66.7%) across all bands, but lower accuracy for non-experienced participants. In contrast, the GPC classifier exhibited the weakest generalization, often misclassifying control trials as surgeon class, suggesting model overfitting or limited feature separability.

DISCUSSION

The present findings demonstrate notable performance differences between experienced surgeons and non-experienced during simulated task execution. Although not all group differences reached statistical significance, trends consistently favored the surgical group in terms of faster completion times, lower error rates, and reduced subjective workload. These results are consistent with prior studies reporting superior psychomotor efficiency among trained surgical professionals (Kim et al., 2014; Zakeri et al., 2020)

Interestingly, while both groups reported similar overall workload on the NASA-TLX, subscale analyses revealed more pronounced associations between task duration and perceived effort, frustration, and temporal demand in the novice group. These stronger correlations suggest that novices may be more sensitive to time pressure and cognitive overload during task performance, which could reflect less automaticity and greater reliance on controlled processing. In contrast, surgeons exhibited more stable workload ratings regardless of task duration, potentially indicating more efficient mental resource allocation and adaptive coping strategies under time constraints.(Keles et al., 2021)

Moreover, the reduced number of errors in the surgeon group further supports the idea that domain-specific expertise not only enhances efficiency but also promotes accuracy and reduces the likelihood of performance breakdowns under pressure. This aligns with the concept of "cognitive unloading" observed in expert populations, where task demands are managed with lower subjective strain due to automatized skill execution(Dias et al., 2018; Hannah et al., 2022; Howie et al., 2023). Taken together, these findings underscore the value of experience in moderating the relationship between task demands and subjective workload, with potential implications for training protocols, performance monitoring, and workload management in high-stakes domains such as surgery.

The combination of theta PSD with asymmetry metrics improved classification accuracy compared to using either modality alone. This suggests that integrating neurophysiological and behavioral measures provides a more comprehensive representation of surgical expertise. While theta-band activity captures cognitive control and error monitoring processes, asymmetry values reflect motor coordination and execution (Akkad et al., 2021; Balkhoyor et al., 2020). Their fusion enables the model to distinguish between groups based on both cognitive and motor dimensions of performance. This multimodal approach aligns with recent findings emphasizing the value of combining behavioral and electrophysiological markers for skill assessment. Recent findings in literature suggest that combining

neurophysiological and behavioral/performance features can enhance classification performance in expertise assessment (Shafiei, Shadpour, Mohler, Sasangohar, et al., 2023). In our study; our approach introduced tool-pressure features as a behavioral modality to capture fine-grained aspects of motor execution. This integration of EEG and tool-pressure data provides a more comprehensive representation of surgical performance by linking cognitive control with motor precision.

This multimodal integration leverages complementary information: while theta oscillations capture the cognitive and attentional demands of the task, asymmetry values reflect fine motor control and sensorimotor coordination. Such findings are consistent with recent studies demonstrating that multimodal approaches, particularly those combining EEG activity with behavioral indicators, significantly enhance discriminative power in skill-level classification. Therefore, the fusion of EEG theta features and asymmetry may serve as a reliable biomarker framework for distinguishing expert and novice performers in immersive surgical tasks. The superior performance of the AdaBoost model suggests its robustness in capturing complex, non-linear feature interactions within EEG data. Furthermore, the higher classification accuracy obtained from task-related EEG features supports the notion that neural activity during active engagement provides richer information for distinguishing expertise. This may be related to enhanced neural differentiation during cognitive and motor execution in skilled participants. ABC emerged as the most reliable classifier, followed closely by RFC. The GPC model consistently underperformed, often collapsing to biased predictions. Furthermore, features extracted from task conditions yielded better classification outcomes compared to rest, reinforcing the importance of recording under cognitive/motor engagement for expertise decoding.

Our study had several limitations. The primary limitation was the relatively small number of participants, which may have reduced the statistical power to detect certain effects. A larger sample size would help confirm or refute some of the observed trends. Additionally, the surgical experience of the participants varied considerably and we did not administer any standardized aptitude or psychomotor skill tests prior to the experiment. This variability may have influenced performance measures and introduced uncontrolled individual differences.

The integration of EEG and sensor data for objective and automatic assessment of surgical skills offers significant advantages for both surgical education and clinical practice. Unlike traditional evaluation methods that rely on subjective observation, these data-driven approaches provide quantifiable and individualized feedback during laparoscopics tasks. Such neurophysiological and behavioral metrics can be used to design personalized and adaptive training programs, allowing trainees to identify their specific weaknesses and improve more efficiently. Moreover, longitudinal monitoring of EEG-based performance markers enables the tracking of learning curves and skill transfer from simulation to the operating room. Ultimately, these automated systems can enhance surgical proficiency, reduce human error, and contribute to safer and more effective clinical outcomes.

References

1. Akkad H, Dupont-Hadwen J, Kane E, Evans C, Barrett L, Frese A, Tetkovic I, Bestmann S, Stagg CJ (2021) Increasing human motor skill acquisition by driving theta–gamma coupling. ELife 10:e67355. https://doi.org/10.7554/eLife.67355
2. Alleblas CCJ, Vleugels MPH, Nieboer TE (2016) Ergonomics of laparoscopic graspers and the importance of haptic feedback: the surgeons' perspective. Gynecol Surg 13(4):379–384. https://doi.org/10.1007/s10397-016-0959-z
3. Araki A, Makiyama K, Yamanaka H, Ueno D, Osaka K, Nagasaka M, Yamada T, Yao M (2017) Comparison of the performance of experienced and novice surgeons: measurement of gripping force during laparoscopic surgery performed on pigs using forceps with pressure sensors. Surg Endosc 31(4):1999–2005. https://doi.org/10.1007/s00464-016-5153-x
4. Balkhoyor AM, Awais M, Biyani S, Schaefer A, Craddock M, Jones O, Manogue M, Mon-Williams MA, Mushtaq F (2020) Frontal theta brain activity varies as a function of surgical experience and task error. BMJ Surg Interventions Health Technol 2(1):e000040. https://doi.org/10.1136/bmjsit-2020-000040
5. Bonrath EM, Dedy NJ, Zevin B, Grantcharov TP (2013) Defining technical errors in laparoscopic surgery: a systematic review. Surg Endosc 27(8):2678–2691. https://doi.org/10.1007/s00464-013-2827-5
6. Brown JD, O'Brien CE, Leung SC, Dumon KR, Lee DI, Kuchenbecker KJ (2017) Using Contact Forces and Robot Arm Accelerations to Automatically Rate Surgeon Skill at Peg Transfer. IEEE Trans Biomed Eng 64(9):2263–2275. https://doi.org/10.1109/TBME.2016.2634861
7. Buia A, Stockhausen F, Hanisch E (2015) Laparoscopic surgery: A qualified systematic review. World J Methodol 5(4):238–254. https://doi.org/10.5662/wjm.v5.i4.238
8. Amanda DBJFFJ, Mary O, Justin OMCARNA, Mousumi D, B.,J.O., B. N (2025) Surgical Skill and Complication Rates after Bariatric Surgery. N Engl J Med 369(15):1434–1442. https://doi.org/10.1056/NEJMsa1300625
9. Darzi A, Datta V, Mackay S (2001) The challenge of objective assessment of surgical skill. Am J Surg 181(6):484–486. https://doi.org/10.1016/S0002-9610(01)00624-9
10. Dias RD, Ngo-Howard MC, Boskovski MT, Zenati MA, Yule SJ (2018) Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. Br J Surg 105(5):491–501. https://doi.org/10.1002/bjs.10795
11. Dockter RL, Lendvay TS, Sweet RM, Kowalewski TM (2017) The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. Int J Comput Assist Radiol Surg 12(7):1151–1159. https://doi.org/10.1007/s11548-017-1610-9
12. Ebina K, Abe T, Yan L, Hotta K, Highuchi M, Iwahara N, Furumido J, Kon M, Murai S, Kurashima Y, Komizunai S, Tsujita T, Sase K, Chen X, Senoo T, Shinohara N, Konno A (2024) Development of Machine Learning-Based Assessment System for Laparoscopic Surgical Skills Using Motion-Capture. *2024 IEEE/SICE International Symposium on System Integration (SII)*, 1–6. https://doi.org/10.1109/SII58957.2024.10417615

13. Farah E, Desir A, Marques C, Hegde SR, Abreu A, Polanco PM, Holcomb C, Scott DJ, Sankaranarayanan G (2024) Heart rate variability: an objective measure of mental stress in surgical simulation. Global Surg Educ - J Association Surg Educ 3(1):25. https://doi.org/10.1007/s44186-023-00220-7
14. Franco-González IT, Lappalainen N, Bednarik R (2024) Tracking 3D motion of instruments in microsurgery: A comparative study of stereoscopic marker-based vs. deep learning method for objective analysis of surgical skills. Inf Med Unlocked 51:101593. https://doi.org/https://doi.org/10.1016/j.imu.2024.101593
15. Gao Y, Kruger U, Intes X, Schwaitzberg S, De S (2020) A machine learning approach to predict surgical learning curves. Surgery 167(2):321–327. https://doi.org/10.1016/j.surg.2019.10.008
16. Gao Y, Yan P, Kruger U, Cavuoto L, Schwaitzberg S, De S, Intes X (2021) Functional Brain Imaging Reliably Predicts Bimanual Motor Skill Performance in a Standardized Surgical Task. IEEE Trans Bio Med Eng 68(7):2058–2066. https://doi.org/10.1109/TBME.2020.3014299
17. Golahmadi AK, Khan DZ, Mylonas GP, Marcus HJ (2021) Tool-tissue forces in surgery: A systematic review. Annals Med Surg (2012) 65:102268. https://doi.org/10.1016/j.amsu.2021.102268
18. Hannah TC, Turner D, Kellner R, Bederson J, Putrino D, Kellner CP (2022) Neuromonitoring Correlates of Expertise Level in Surgical Performers: A Systematic Review. Front Hum Neurosci 16:705238. https://doi.org/10.3389/fnhum.2022.705238
19. Healey MA, Shackford SR, Osler TM, Rogers FB, Burns E (2002) Complications in Surgical Patients. Arch Surg 137(5):611–618. https://doi.org/10.1001/archsurg.137.5.611
20. Howie EE, Dharanikota H, Gunn E, Ambler O, Dias R, Wigmore SJ, Skipworth RJE, Yule S (2023) Cognitive Load Management: An Invaluable Tool for Safe and Effective Surgical Training. J Surg Educ 80(3):311–322. https://doi.org/https://doi.org/10.1016/j.jsurg.2022.12.010
21. Huaulmé A, Tronchot A, Thomazeau H, Jannin P (2025) Automated assessment of non-technical skills by heart-rate data. Int J Comput Assist Radiol Surg 20(3):561–568. https://doi.org/10.1007/s11548-024-03287-9
22. Keles HO, Cengiz C, Demiral I, Ozmen MM, Omurtag A (2021) High density optical neuroimaging predicts surgeons's subjective experience and skill levels. PLoS ONE 16(2):e0247117. https://doi.org/10.1371/journal.pone.0247117
23. Kim HJ, Choi G-S, Park JS, Park SY (2014) Comparison of surgical skills in laparoscopic and robotic tasks between experienced surgeons and novices in laparoscopic surgery: an experimental study. Annals Coloproctology 30(2):71–76. https://doi.org/10.3393/ac.2014.30.2.71
24. Manabe T, Rahul FNU, Fu Y, Intes X, Schwaitzberg SD, De S, Cavuoto L, Dutta A (2023) Distinguishing Laparoscopic Surgery Experts from Novices Using EEG Topographic Features. Brain Sci 13(12). https://doi.org/10.3390/brainsci13121706
25. Muthukumaraswamy SD (2013) High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. Front Hum Neurosci 7:138. https://doi.org/10.3389/fnhum.2013.00138

26. Natheir S, Christie S, Yilmaz R, Winkler-Schwartz A, Bajunaid K, Sabbagh AJ, Werthner P, Fares J, Azarnoush H, Del Maestro R (2023) Utilizing artificial intelligence and electroencephalography to assess expertise on a simulated neurosurgical task. Comput Biol Med 152:106286. https://doi.org/https://doi.org/10.1016/j.compbiomed.2022.106286
27. Nemani A, Kruger U, Cooper CA, Schwaitzberg SD, Intes X, De S (2019) Objective assessment of surgical skill transfer using non-invasive brain imaging. Surg Endosc 33(8):2485–2494. https://doi.org/10.1007/s00464-018-6535-z
28. Nemani A, Yücel MA, Kruger U, Gee DW, Cooper C, Schwaitzberg SD, De S, Intes X (2018) Assessing bimanual motor skills with optical neuroimaging. Sci Adv 4(10). https://doi.org/10.1126/sciadv.aat3807
29. Oh J, Lau N (2024) Quantitative Analysis of Eye-gaze Metrics in Differentiating Surgical Expertise. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *68*(1), 624–626. https://doi.org/10.1177/10711813241264200
30. Omurtag A, Roy RN, Dehais F, Chatty L, Garbey M (2019) Chapter 16 - Tracking Mental Workload by Multimodal Measurements in the Operating Room. In H. Ayaz & F. Dehais (Eds.), *Neuroergonomics* (pp. 99–103). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-811926-6.00016-6
31. Omurtag A, Sunderland C, Mansfield NJ, Zakeri Z (2025) EEG connectivity and BDNF correlates of fast motor learning in laparoscopic surgery. Sci Rep 15(1):7399. https://doi.org/10.1038/s41598-025-89261-0
32. Power D, Burke C, Madden MG, Ullah I (2025) Automated assessment of simulated laparoscopic surgical skill performance using deep learning. Sci Rep 15(1):13591. https://doi.org/10.1038/s41598-025-96336-5
33. Rafii-Tari H, Payne CJ, Bicknell C, Kwok K-W, Cheshire NJW, Riga C, Yang G-Z (2017) Objective Assessment of Endovascular Navigation Skills with Force Sensing. Ann Biomed Eng 45(5):1315–1327. https://doi.org/10.1007/s10439-017-1791-y
34. Shafiei SB, Shadpour S, Mohler JL, Attwood K, Liu Q, Gutierrez C, Toussi MS (2023) Developing surgical skill level classification model using visual metrics and a gradient boosting algorithm. Annals Surg Open: Perspect Surg History Educ Clin Approaches 4(2). https://doi.org/10.1097/as9.0000000000000292
35. Shafiei SB, Shadpour S, Mohler JL, Sasangohar F, Gutierrez C, Toussi S, M., Shafqat A (2023) Surgical skill level classification model development using EEG and eye-gaze data and machine learning algorithms. J Robotic Surg 17(6):2963–2971. https://doi.org/10.1007/s11701-023-01722-8
36. Soangra R, Jiang P, Haik D, Xu P, Brevik A, Peta A, Tapiero S, Landman J, John E, Clayman RV (2022) Beyond Efficiency: Surface Electromyography Enables Further Insights into the Surgical Movements of Urologists. J Endourol 36(10):1355–1361. https://doi.org/10.1089/end.2022.0120
37. Soangra R, Sivakumar R, Anirudh ER, Reddy Y, S. V., John EB (2022) Evaluation of surgical skill using machine learning with optimal wearable sensor locations. PLoS ONE 17(6):e0267936.

https://doi.org/10.1371/journal.pone.0267936

38. Soto Rodriguez NA, Arroyo Kuribreña C, Hernández P, Gutiérrez-Gnecchi JD, Pérez-Escamirosa JA, Rigoberto F, Minor-Martinez M-M, A., Lorias-Espinoza D (2023) Objective Evaluation of Laparoscopic Experience Based on Muscle Electromyography and Accelerometry Performing Circular Pattern Cutting Tasks: A Pilot Study. Surg Innov 30(4):493–500. https://doi.org/10.1177/15533506231169063
39. Sugiyama T, Lama S, Gan LS (2018) Forces of Tool-Tissue Interaction to Assess Surgical Skill Level. JAMA Surg 153(3):234–242. https://doi.org/10.1001/jamasurg.2017.4516
40. The A-F, Reijmerink I, van der Laan M, Cnossen F (2020) Heart rate variability as a measure of mental stress in surgery: a systematic review. Int Arch Occup Environ Health 93(7):805–821. https://doi.org/10.1007/s00420-020-01525-6
41. Tien T, Pucher PH, Sodergren MH, Sriskandarajah K, Yang G-Z, Darzi A (2014) Eye tracking for skills assessment and training: a systematic review. J Surg Res 191(1):169–178. https://doi.org/10.1016/j.jss.2014.04.032
42. Vinck M, Oostenveld R, van Wingerden M, Battaglia F, Pennartz CMA (2011) An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. NeuroImage 55(4):1548–1565. https://doi.org/10.1016/j.neuroimage.2011.01.055
43. Zakeri Z, Mansfield N, Sunderland C, Omurtag A (2020) Physiological correlates of cognitive load in laparoscopic surgery. Sci Rep 10(1):12927. https://doi.org/10.1038/s41598-020-69553-3
44. Zia A, Sharma Y, Bettadapura V, Sarin EL, Essa I (2018) Video and accelerometer-based motion analysis for automated surgical skills assessment. Int J Comput Assist Radiol Surg 13(3):443–455. https://doi.org/10.1007/s11548-018-1704-z

Figures

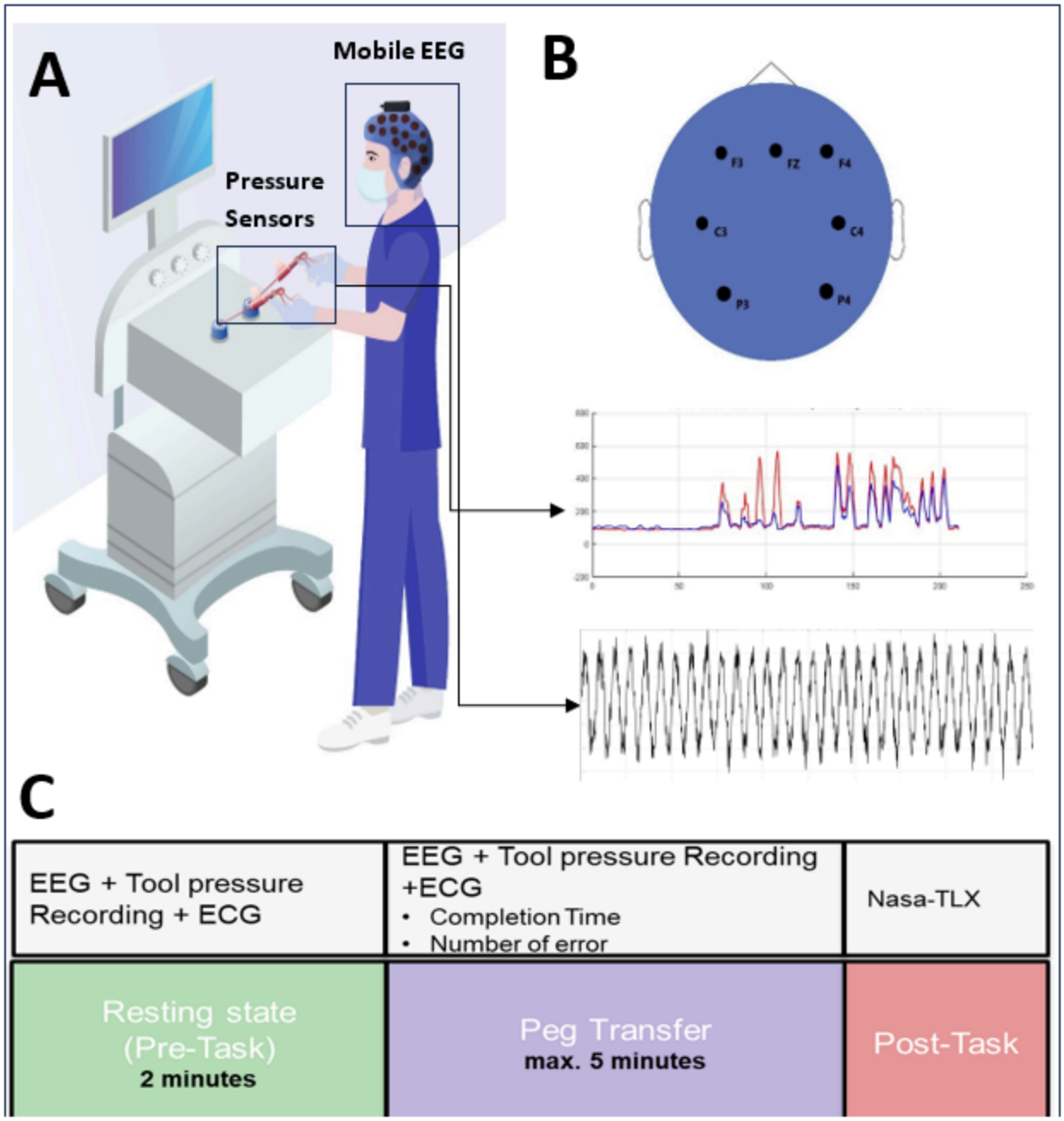

Figure 1

Experimental setup for multimodal assessment of surgical skills. (A) A laparoscopic training box equipped with custom pressure sensors was used to record tool–surgeon interaction forces from both grasper handles (in Newtons). A mobile EEG system was employed to capture neural activity with electrodes placed over frontal and parietal regions (F3, F4, C3, C4, P3, P4, Fz). Representative signals include force profiles from right and left graspers, their difference (Right–Left), and EEG waveforms. The combined signals enabled the computation of Right–Left Asymmetry (RLA) and EEG spectral features to characterize motor performance and cognitive load during laparoscopic simulation tasks. C) Experimental design includes Resting and Task (Peg Transfer)

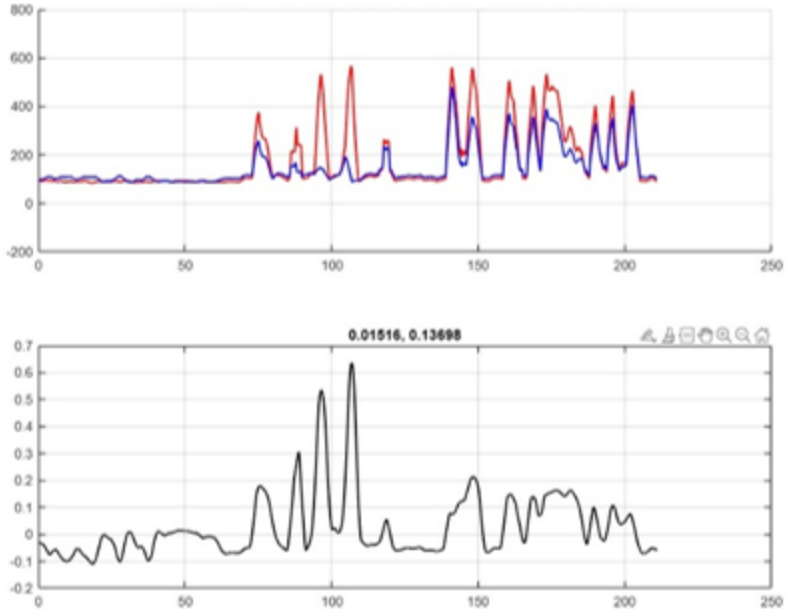

Figure 2

Force signals recorded from laparoscopic graspers during the task. The red line represents the right-hand grasper and the blue line represents the left-hand grasper. The bottom panel shows the difference between right and left force signals, indicating a clear dominance of the hand throughout the task. The x-axis represents time (s) and the y-axis represents force (N).

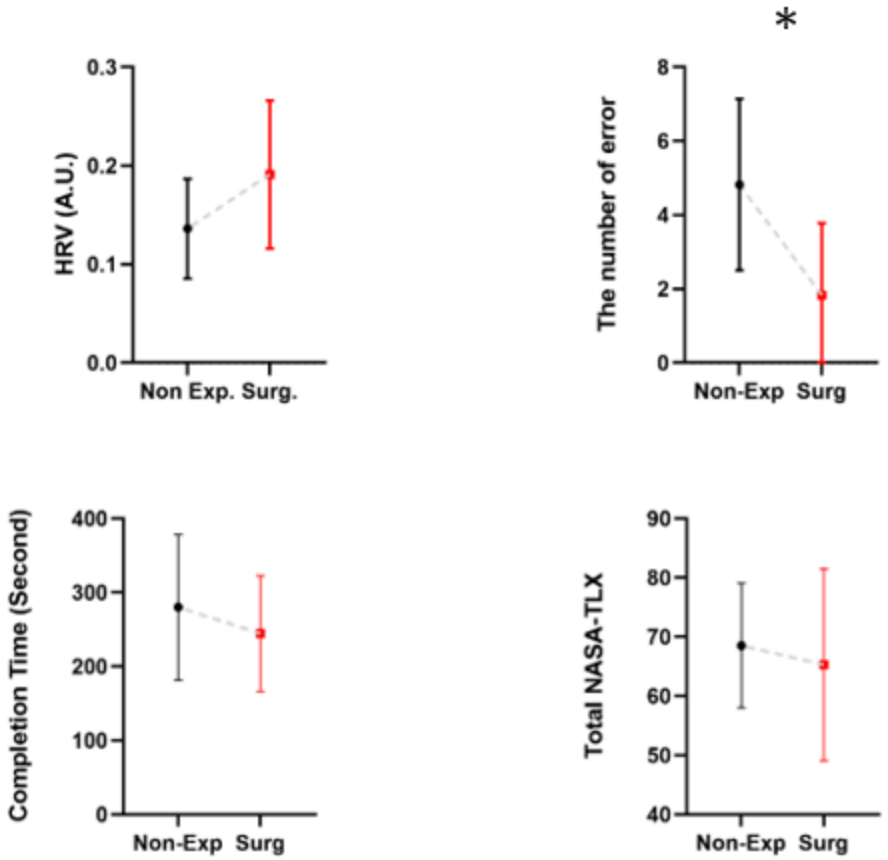

Figure 3

This figure shows the comparison between non-experienced (Non-Exp) and surgeons (Surg) across four measures: HRV, number of errors, task completion time, and total NASA-TLX scores Metrics include heart rate variability (HRV), number of errors, task completion time, and total NASA-TLX scores. Only the number of errors showed a significant group difference ($p < 0.05$), with surgeons making fewer errors during the peg transfer task. Error bars indicate standard deviation.

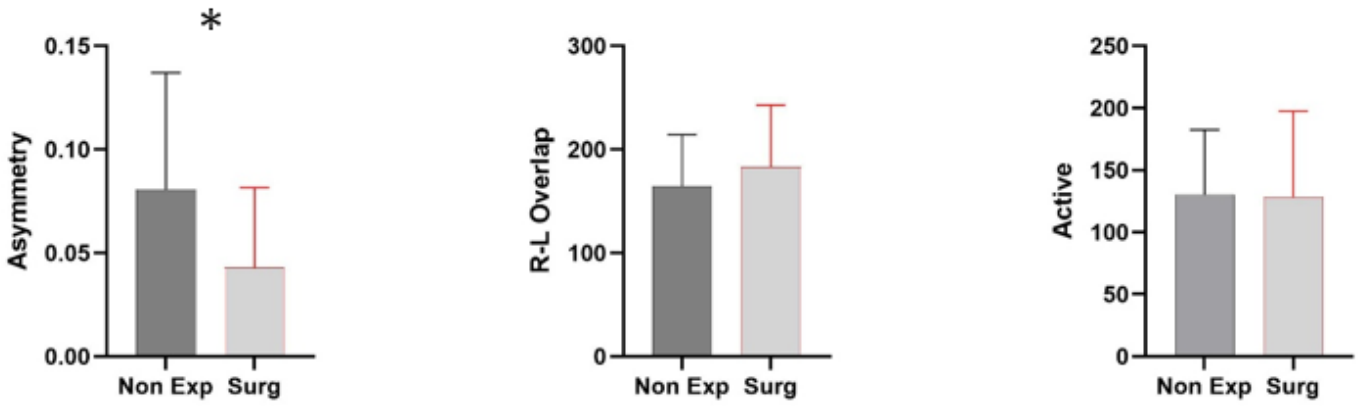

Figure 4

Sensor Comparison of sensor-derived force metrics between surgeons (Surg) and non-experienced participants (Non-Exp). A trend toward lower asymmetry was observed in the Surg group (p = 0.06), suggesting more balanced bimanual coordination. No significant differences were found for R–L overlap duration or active force duration. Error bars indicate standard deviation.

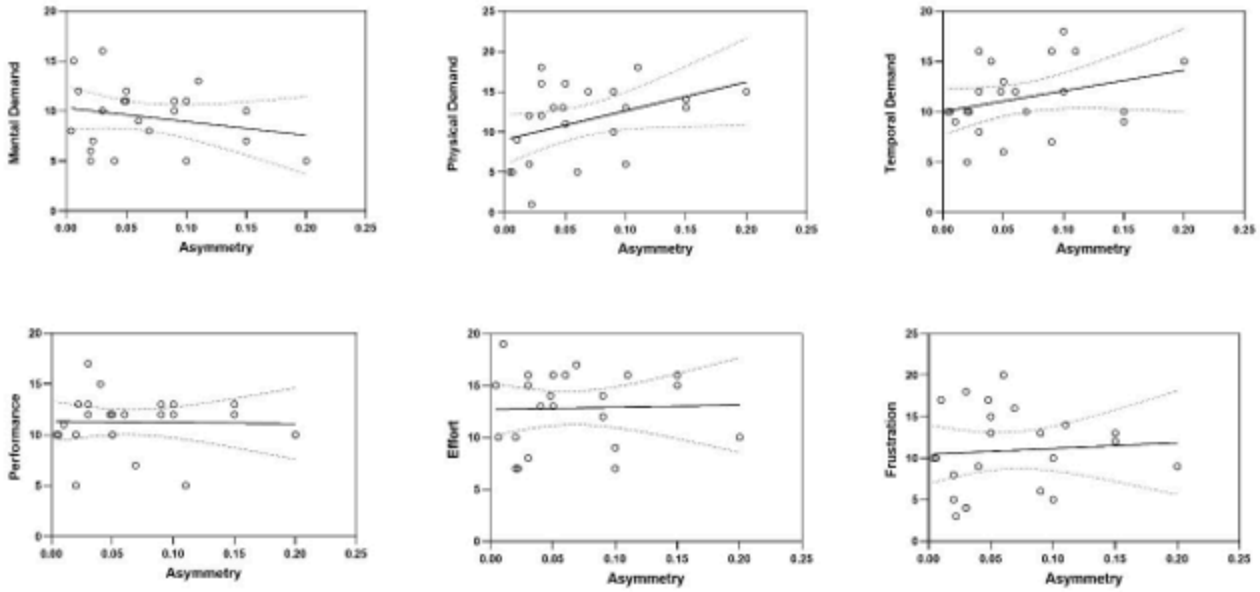

Figure 5

Simple Linear regression analyses examining the relationships between NASA-TLX subscale scores and asymmetry index.

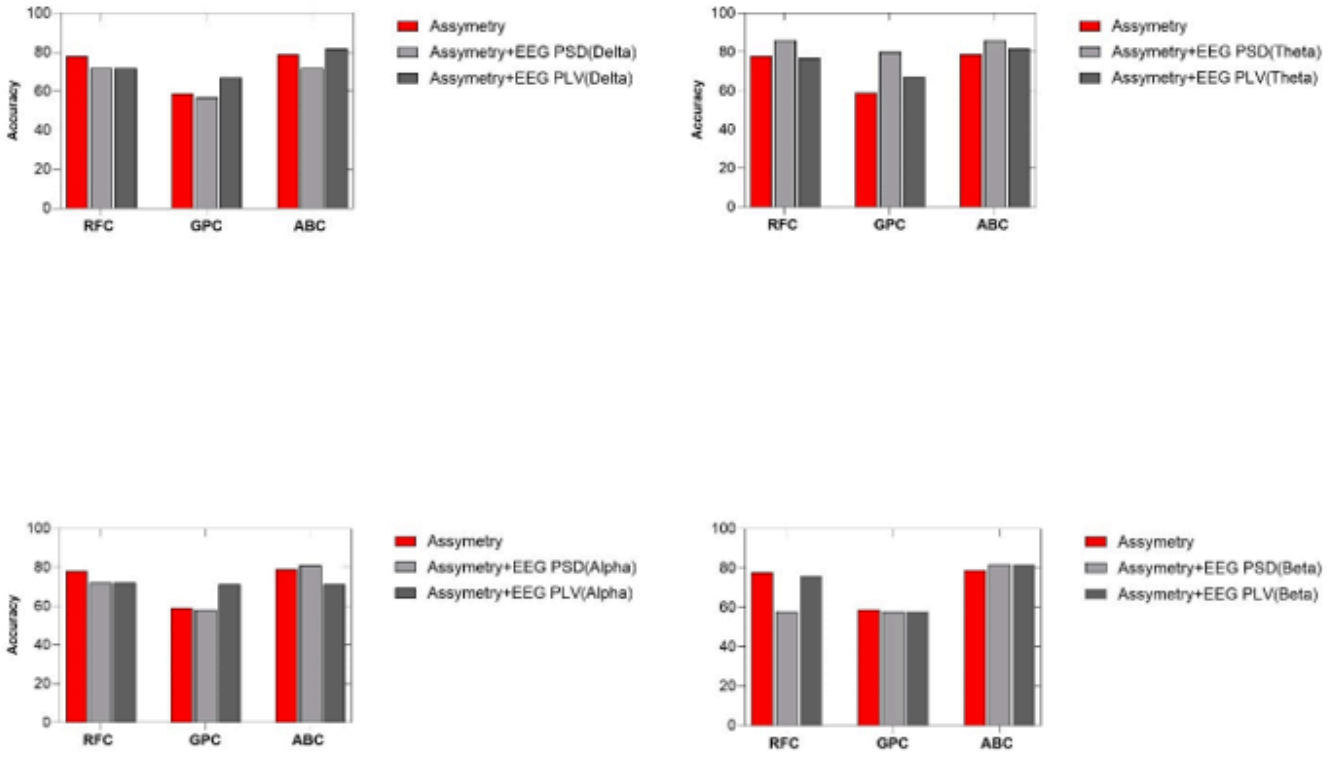

Figure 6

Classification accuracies across four EEG frequency bands (delta, theta, alpha, and beta) using three different classifiers: RFC, GPC, and ABC. Three feature sets were compared: asymmetry-only (red), EEG power spectral density (PSD) combined with asymmetry (light gray), and EEG phase locking value (PLV) combined with asymmetry (dark gray).

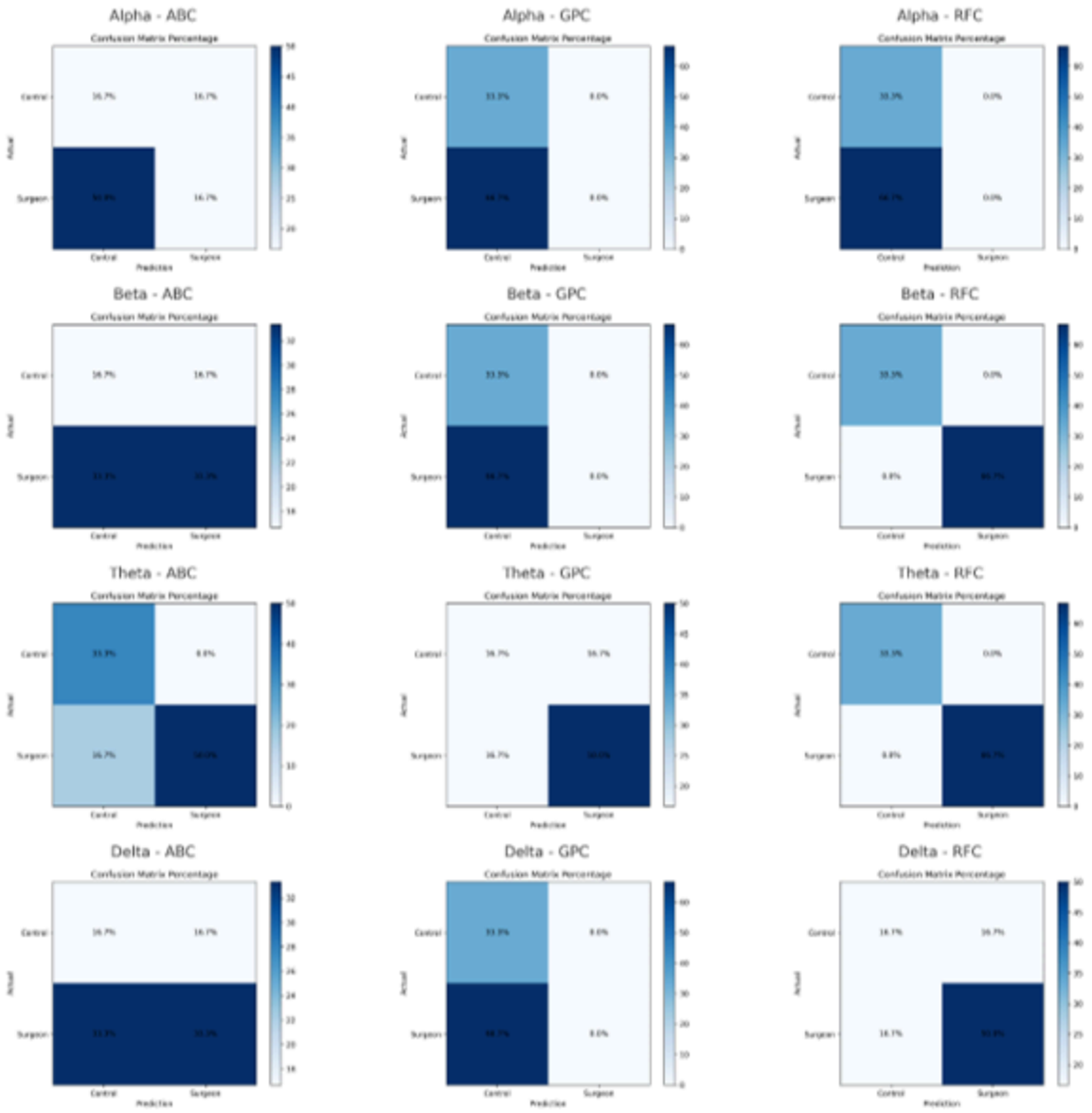

Figure 7

Confusion matrices illustrating the classification performance of ABC, GPC, and RFC models. The input features combined EEG PSD derived from each frequency band and asymmetry. Rows represent the actual classes (Non experienced, Surgeon), while columns indicate predicted labels. Color intensity corresponds to the percentage of correctly and incorrectly classified samples within each model–band pair.

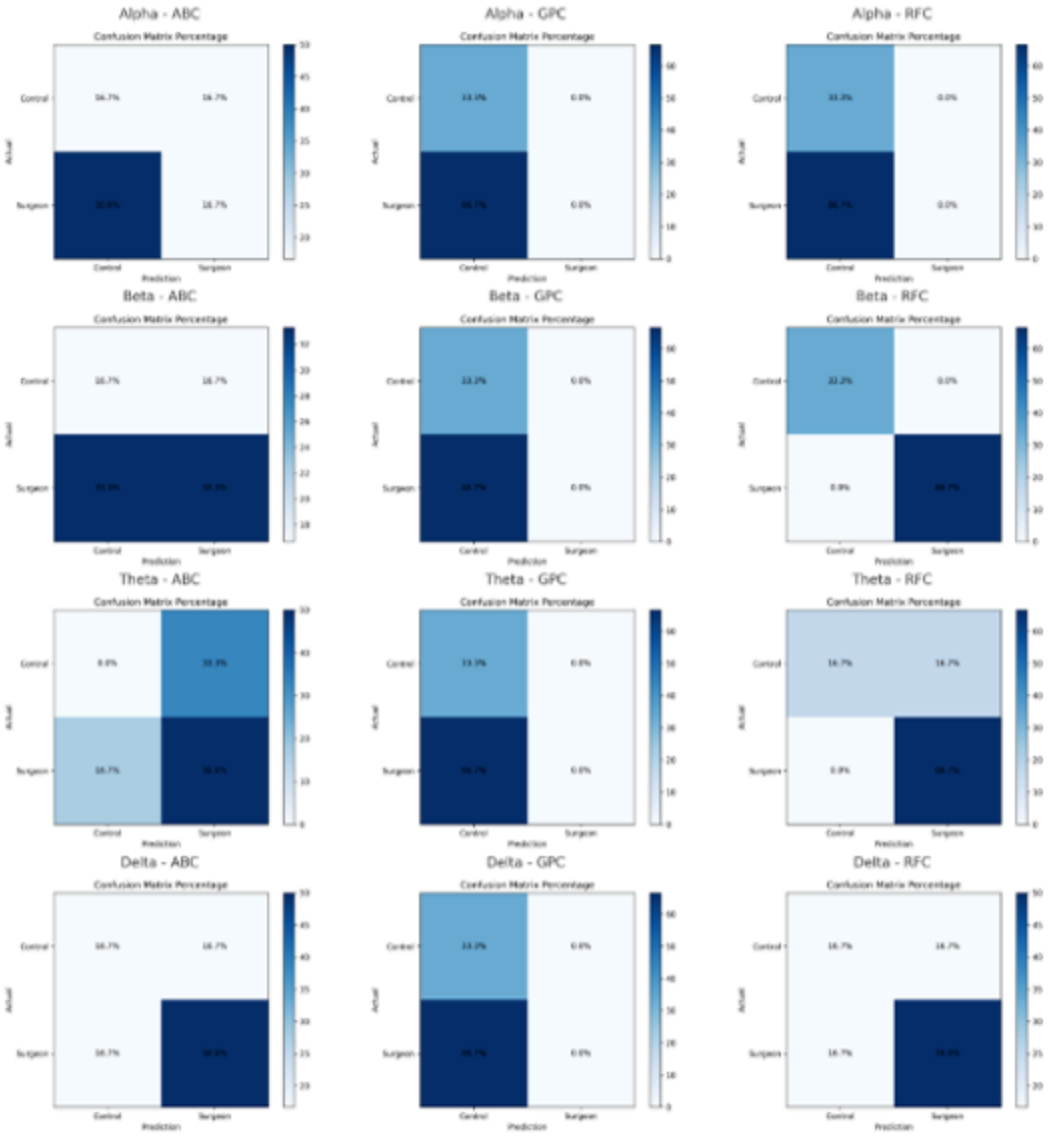

Figure 8

Confusion matrices illustrating the classification performance of ABC, GPC, and RFC. The input features combined EEG PLV derived from each frequency band and asymmetry. Rows represent the actual classes (Non experienced, Surgeon), while columns indicate predicted labels. Color intensity corresponds to the percentage of correctly and incorrectly classified samples within each model–band pair.