Original Article | **Open Access**

Artificial Intelligence Surgery

Sahin *et al. Art Int Surg.* 2026;6:171-87 **DOI:**10.20517/ais.2025.109

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

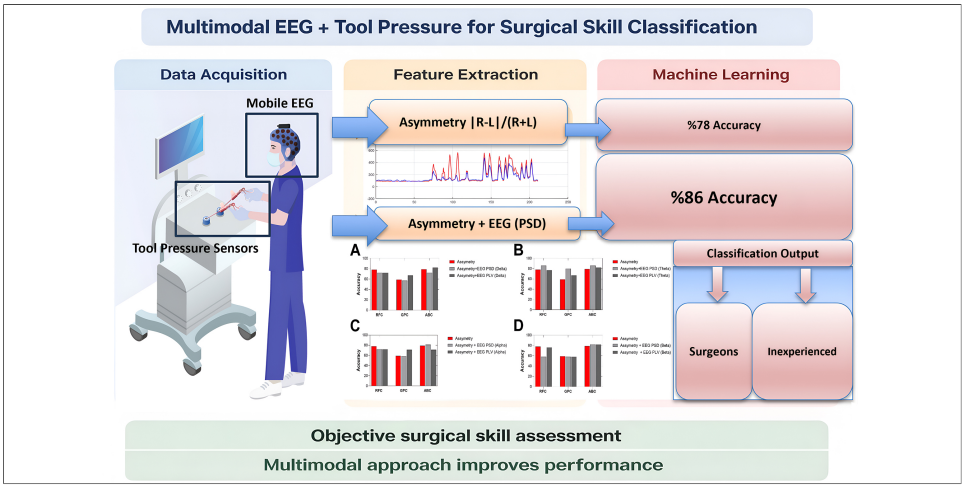
Sebahat Selin Sahin[1], Cagri Zengin[1], Hasan Onur Keles[1,2,3]

Keywords:
EEG, surgical skill, assessment, classification

Citation: Sahin SS, Zengin C, Keles HO. A multimodal approach combining tool-pressure and EEG features for laparoscopic skill classification using machine learning. *Art Int Surg.* 2026;6:171-87. https://dx.doi.org/10.20517/ais.2025.109

Received: 3 Nov 2025
First Decision: 4 Dec 2025
Revised: 4 Mar 2026
Accepted: 18 Mar 2026
Published: 27 Mar 2026

Academic Editor:
Roland Croner
Copy Editor:
Ting-Ting Hu
Production Editor:
Ting-Ting Hu



Abstract

Aim: Laparoscopic skill assessment traditionally relies on subjective evaluation, which lacks objectivity and consistency. Automated multimodal approaches integrating tool-pressure and neural data may improve the reliability and scalability of skill assessment. Therefore, our objectives were to: (1) integrate a pressure-sensing unit into box-trainer simulators and laparoscopic tools to investigate tool-pressure features as objective indicators of surgical skill; and (2) combine electroencephalography (EEG)-derived power spectral density (PSD) 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 National Aeronautics and Space Administration Task Load Index (NASA-TLX) workload scores, were collected from 10 surgeons and 13 inexperienced students performing a peg-transfer laparoscopic task. Pressure sensors were integrated into the right and left laparoscopic graspers. EEG features were extracted from four frequency bands using PSD and PLV. Three machine learning models - random forest classifier (RFC), Gaussian process classifier (GPC), and AdaBoost classifier (ABC) - were used to classify participants into surgeon and inexperienced groups.

[1]Department of Biomedical Engineering, Ankara University, Ankara 06100, Turkey.
[2]Brain Research Center (AUBAUM), Ankara University, Ankara 06340, Turkey.
[3]Neuroscience and Neurotechnology Center of Excellence (NOROM), Ankara University, Ankara 06560, Turkey.

Correspondence to: Assist. Prof. Hasan Onur Keles, Department of Biomedical Engineering, Ankara University, Ankara 06100, Turkey. E-mail: hokeles@ankara.edu.tr

Results: Right-left pressure asymmetry emerged as a reliable indicator of surgical expertise compared with other pressure metrics. Using only this feature, RFC achieved up to 78% classification accuracy. The highest performance occurred when combining theta-band power features with pressure asymmetry, where RFC and ABC reached 86% accuracy [F1 score = 0.83; area under the curve (AUC) = 0.92 for RFC].

Conclusion: 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[1-3]. 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[4-6].

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[7,8]. Psychophysiological measures offer objective and continuous evaluations, overcoming limitations associated with traditional skill assessment methods, which typically rely on behavioural metrics and subjective surveys[1,9,10]. Physiological sensor-based technologies, such as electrocardiography (ECG)[11-13], electromyography (EMG)[14,15], electroencephalography (EEG)[1,16] and functional near-infrared spectroscopy (fNIRS)[5,17-19] 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[20,21] and eye tracking[22,23] 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[24]. To objectively assess these interactions, force sensors have been embedded into various surgical tools, including forceps, graspers, and gloves[25-28]. 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[26,29]. 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[17,19]. 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 (R-L) 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[25,26] have shown the feasibility of pressure-based skill assessment; however, 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 types[8,30-34], the combination of EEG and tool pressure data remains largely unexplored. Recent advances in data-driven systems highlight the importance of machine learning approaches capable of integrating multiple information sources and handling uncertainty, a perspective that is also highly relevant for multimodal surgical skill classification based on heterogeneous signals such as EEG and tool-pressure data[35,36]. 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 ± SD: 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 (Approval No. 2024-000333-1) and was conducted at the 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. All participants were right-hand dominant, and therefore handedness was not included as a covariate in the analyses.

Experimental design

Participants performed standard training tasks, including peg transfer using a laparoscopic trainer box [Figure 1A]. Data were recorded, including time to completion and error rate. After subjects completed each task, they completed the National Aeronautics and Space Administration Task Load Index (NASA-TLX) questionnaire. At the beginning of the experiment, two 2-min videos demonstrating the tasks were shown to the participants on a computer screen. After the video session, each participant completed a 15-min training session. The present study used 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 manipulate objects using both hands. It is designed to enhance the 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 2-min rest period.

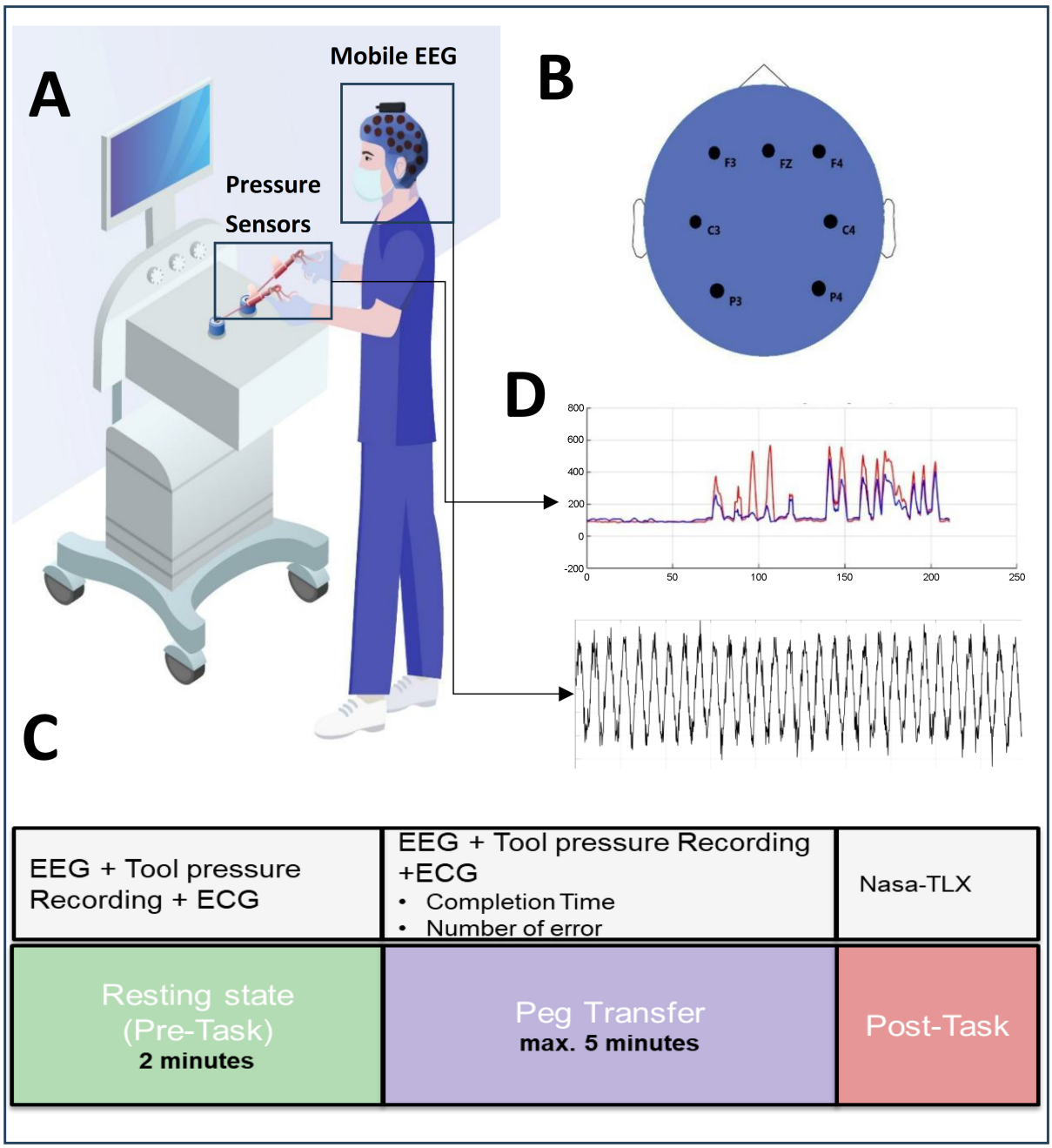


Figure 1. Experimental setup for multimodal assessment of surgical skills. (A) A laparoscopic training box equipped with custom pressure sensors was used to record tool-surgeon interaction forces from both grasper handles (in Newtons); (B) A mobile EEG system was employed to capture neural activity with electrodes placed over frontal and parietal regions; (C) Experimental design includes Resting and Task (Peg Transfer); (D) 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 x-axis represents time (s), and the y-axis represents force (N). The bottom panel displays the corresponding EEG time series. EEG: Electroencephalography; ECG: electrocardiography; Nasa-TLX: Task Load Index.

In this study, sensor data from the tool pressure device was collected simultaneously with EEG and ECG recordings throughout the experiment. To measure brain activity, the 8-channel version of the Portable EEG (Mentalab; Germany) 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 [Figure 1B]. 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). The tool pressure device consisted of thin-film Force-sensing resistors (FSRs) integrated into the handles of the laparoscopic instruments. The sensors were sampled at 20 Hz (f_s = 20 Hz) and digitized using a 10-bit Analog-to-Digital Converter (ADC), providing a measurement range of 0-1,023 arbitrary units (a.u.). Participants completed the NASA-TLX questionnaire, which was developed by the National Aeronautics and Space Administration (NASA) after completing the experiment. An overview of the experimental design is provided in Figure 1C. Two modalities were acquired at different sampling rates (EEG: 200 Hz after preprocessing; tool-pressure: 20 Hz); each signal was processed according to its temporal characteristics [Figure 1D].

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-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.

Tool-pressure data processing

The pressure sensors 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. The sensors measured the grip pressure applied at the instrument handles rather than the direct force at the instrument tip. A series of metrics aimed at capturing distinctions in skill level and training were used[37]. Active time segments were characterized as those periods when pressure exceeded a specified threshold, indicating engagement. For measuring bilateral coordination, we used the R-L Overlap metric, defined as the duration during which pressures from both the right and left sides were simultaneously active. Additionally, we used the R-L 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, respectively, 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.

EEG preprocessing and analysis

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 activity, likely caused by gross body movements during EEG recording, were identified and removed using a 1-s sliding window. Additionally, electrodes exhibiting kurtosis values exceeding 5 were considered invalid channels and excluded from further analysis. The signals were then band-pass filtered between 0.16 and 40 Hz to reduce slow drifts and high-frequency artifacts, and subsequently downsampled to 200 Hz to optimize computational efficiency and storage requirements. Ocular and muscle-related artifacts (e.g., eye blinks, eye movements, and EMG contamination) were not corrected; instead, contaminated data segments were removed based on amplitude- and frequency-based criteria. Finally, the remaining clean EEG data were retained for further analysis.

EEG-based features were computed from frequency band power (power spectral density, PSD) and phase-locking value (PLV). Initially, the spectrogram was calculated using the short-time Fourier transform

method with 1 s windows, 50% overlap, and a 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 Hz), alpha (8-12 Hz) and beta (12-30 Hz). EEG frequency band power for each epoch was extracted by integration of the corresponding power over each frequency band. A cutoff frequency of 32 Hz was applied because higher-frequency components in scalp EEG are generally considered less informative for cortical activity[38]. PLV is a measure of phase synchrony between two distinct neuronal populations, computed between selected EEG electrodes to estimate inter-area synchrony[39]. PLV was calculated for electrode pairs chosen to assess synchrony. It was computed for each narrow (1 Hz) band and then averaged across four bands of interest: 0-4, 4-8, 8-12, and 12-30 Hz. Seven EEG channels (21 channel pairs) were analyzed to represent functional 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.

Feature selection

To ensure the robustness of the predictive models, all features were standardized by centering and scaling based on the mean and standard deviation (SD) of each column. Given the relatively limited sample size, feature selection was performed using the SelectKBest algorithm with mutual information scores to quantify the statistical dependence between features and class labels. To prevent information leakage, the entire preprocessing pipeline including feature standardization, feature selection, and hyperparameter optimization was executed strictly within the training folds of each cross-validation (CV) cycle. The test folds remained entirely unseen during these procedures. For each frequency band, the three highest-ranking features were dynamically identified and retained for model training. In the theta-band analysis, for instance, the top-ranked features consistently included: (1) the R-L pressure asymmetry index; (2) theta-band PSD at Fz; and (3) theta-band PSD at P4. Hyperparameter optimization was conducted using GridSearchCV, which systematically explored predefined parameter grids. Specifically, for the random forest classifier (RFC), the number of estimators and the splitting criterion were optimized; for the AdaBoost classifier (ABC), the number of estimators and the learning rate were tuned; and for the Gaussian process classifier (GPC), the kernel function and optimizer settings were adjusted.

Classification

Multiple machine learning algorithms were trained to distinguish between the surgeon and control groups across different frequency bands. Each participant was represented by a single feature vector comprising band power values from all electrodes within a specific frequency band, integrated with tool-pressure features. Models were trained separately for each frequency band to evaluate their individual discriminative power.

Model performance was evaluated using 3-fold, 5-fold, and 7-fold CV schemes. While 3-fold CV provided larger test sets, 7-fold CV maximized the training data; ultimately, a 5-fold stratified CV scheme was adopted as a practical compromise, ensuring stable performance estimates while maintaining class balance. Leave-one-out cross-validation (LOOCV) was intentionally avoided due to its tendency to yield high-variance results and biased performance estimates in small datasets. No window selection procedure was applied, utilizing the full duration of the recorded tasks. Final model performance was assessed using accuracy and the F1 score, defined as the harmonic mean of precision and recall, and reported as the mean across all cross-validation (CV) folds.

Statistical analysis

Statistical analyses were performed using GraphPad Prism (version 10.0; GraphPad Software, Boston, MA, USA). When conducting a regression analysis comparing 2 numerical variables, linear fit with analysis of

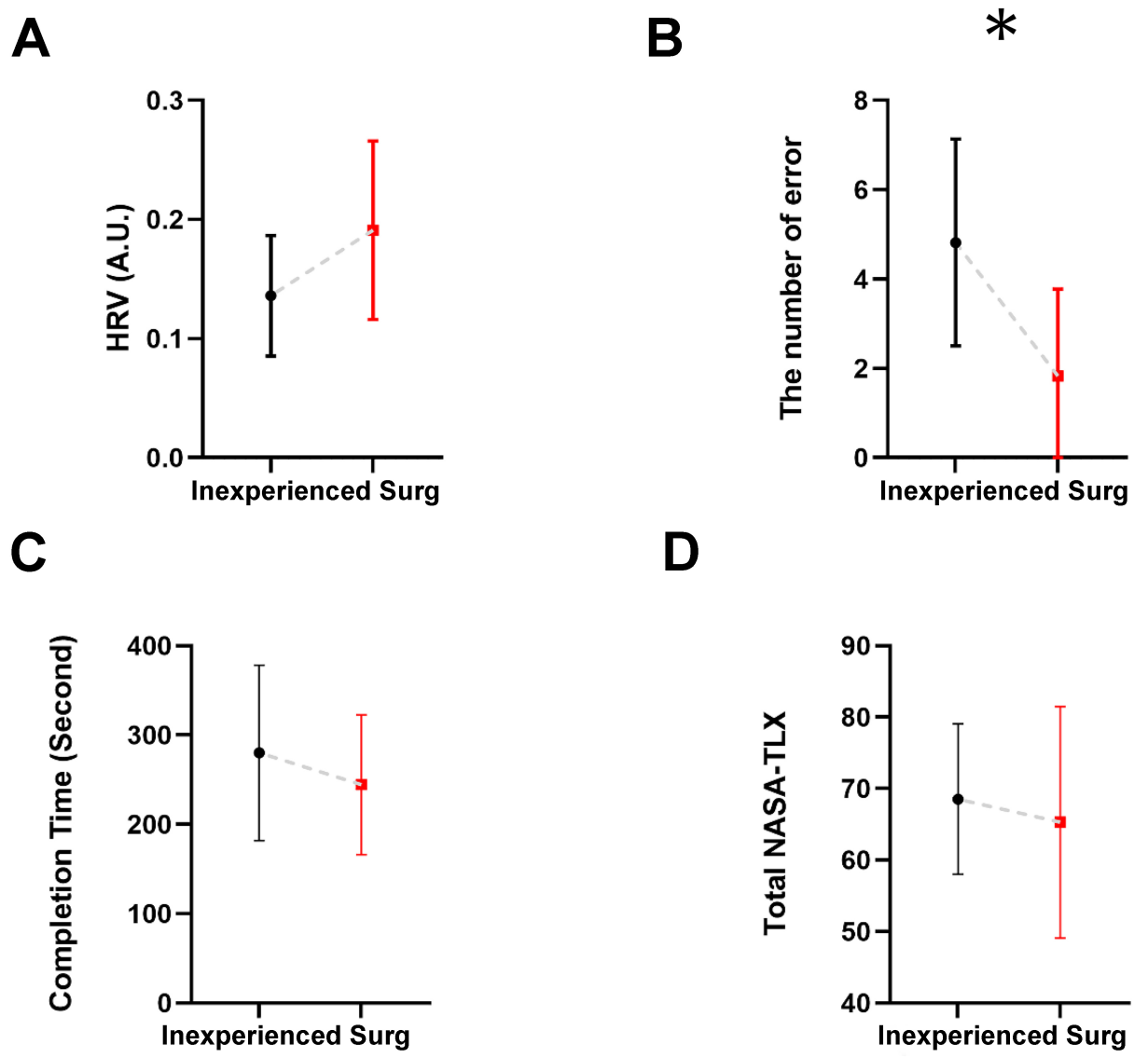


Figure 2. Comparison between inexperienced and Surg across four measures: (A) Heart Rate Variability; (B) Number of errors; (C) Task completion time; and (D) Total NASA-TLX scores. Bars represent group means, and error bars indicate SD. Group differences were evaluated using the Mann-Whitney *U* test. (*) Indicate significant group differences ($P < 0.05$). Surg: Surgeons; HRV: heart rate variability; NASA-TLX: National Aeronautics and Space Administration Task Load Index; SD: standard deviation.

variance was used. The descriptive results comparing the two independent groups (e.g., completion time and asymmetry) were based on unpaired observations. To assess statistical differences between the groups, nonparametric comparisons were performed using the Mann-Whitney *U* test. No additional null hypotheses requiring correction for multiple comparisons or false discovery were applied.

RESULTS

Twenty-three participants (10 surgeons, 13 inexperienced students; age: mean age = 27.3 ± 4.9 years) 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 [Figure 1]. Our analyses proceeded in four main steps. First, we examined participants' behavioral performance metrics, including task completion time and error rates [Figure 2]. Second, we compared sensor-derived force metrics between groups [Figure 3]. Third, we explored the relationship between behavioral performance, sensor-based force metrics, and self-reported workload subscales from the NASA-TLX [Figure 4]. Finally, we evaluated and compared multiple machine learning classification models using sensor-derived features to predict participants' skill level groupings [Figure 5].

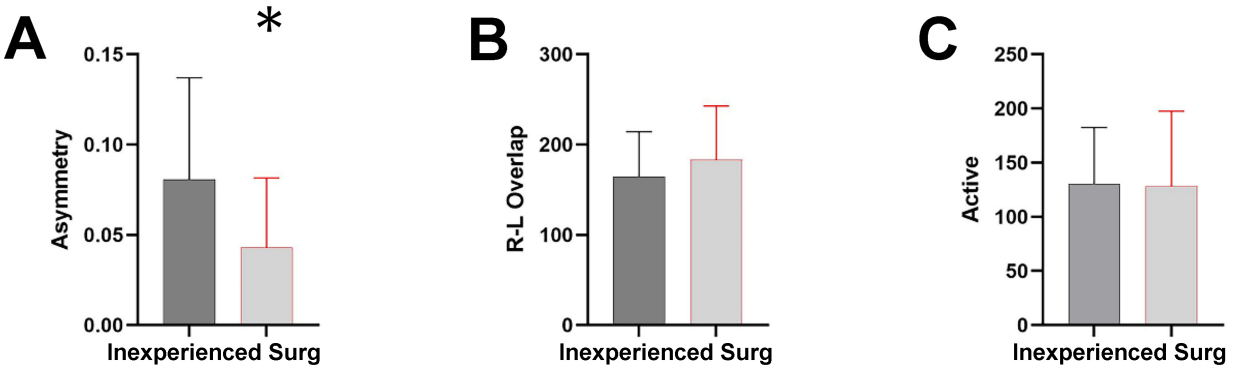


Figure 3. Comparison of sensor-derived force metrics between Surg and inexperienced. (A) Asymmetry index (dimensionless); (B) R-L overlap duration (seconds); and (C) active force duration (seconds). Bars represent group means, and error bars indicate SD. Statistical comparisons were performed using the Mann-Whitney *U* test. (*) indicate significant group differences ($P < 0.05$). Surg: Surgeons; R-L: right-left; SD: standard deviation.

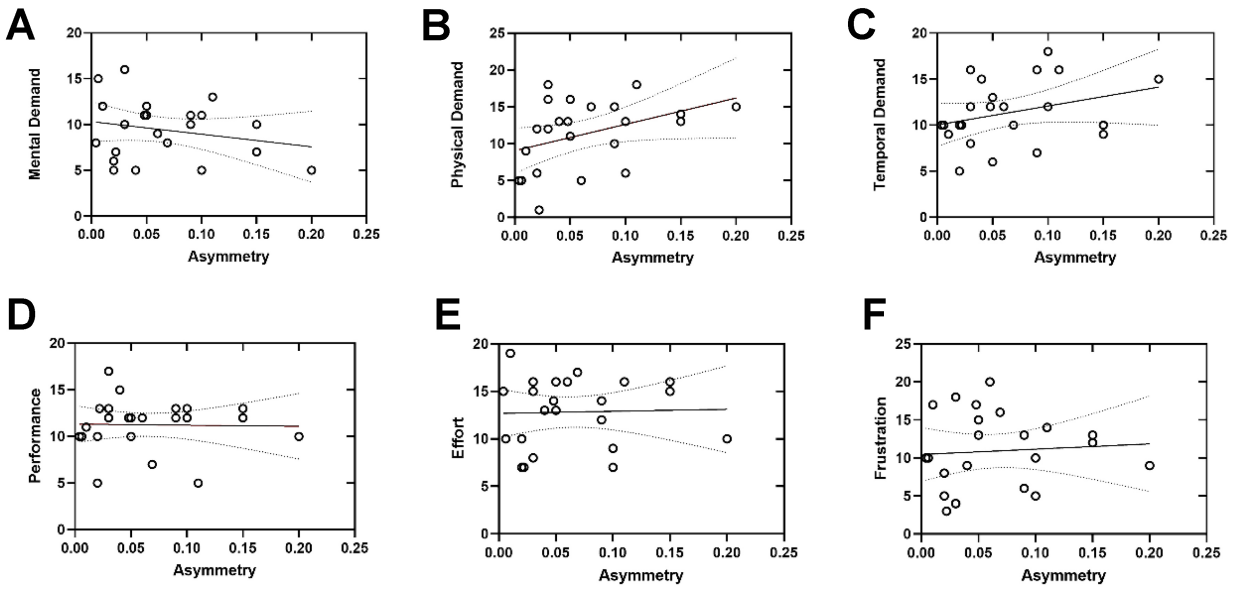


Figure 4. Simple linear regression analyses examining the relationships between the asymmetry index (x-axis, dimensionless) and NASA-TLX subscale scores [y-axis; (A) Mental demand; (B) Physical demand; (C) Temporal demand; (D) Performance; (E) Effort; (F) Frustration]. Each circle represents an individual participant. Solid lines indicate the fitted linear regression models, and dashed lines represent the 95% confidence intervals. Statistical significance was assessed using linear regression analysis, with significance defined as $P < 0.05$. NASA-TLX: National Aeronautics and Space Administration Task Load Index.

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 (inexperienced). As shown in Figure 2, the number of errors during the peg transfer task was significantly lower in the Surg group compared to the inexperienced 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.

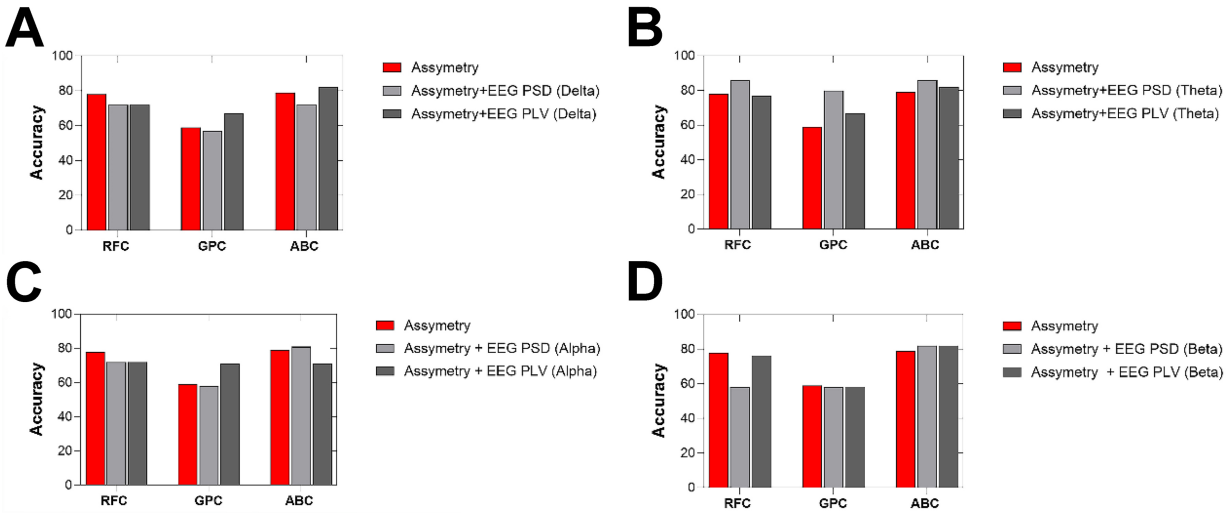


Figure 5. Classification accuracies across four EEG frequency bands using three classifiers: RFC, GPC, and ABC. (A) Delta band; (B) Theta band; (C) Alpha band; and (D) Beta band. The y-axis represents classification accuracy, and the x-axis indicates the classifier type for each frequency band. Three feature sets were compared: asymmetry-only (red), asymmetry combined with EEG PSD (light gray), and asymmetry combined with EEG PLV (dark gray). Accuracy values were obtained using cross-validation procedures as described in the Methods section. EEG: Electroencephalography; RFC: random forest classifier; GPC: Gaussian process classifier; ABC: AdaBoost classifier; PSD: power spectral density; PLV: phase-locking value.

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 (SD), 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 inexperienced participants.

Next, we analysed additional coordination-related force metrics: the asymmetry index, R-L overlap, and active force duration [Figure 3]. The asymmetry index, which quantifies the normalized pressure imbalance between the left and right tool handles, showed a group difference at the threshold of statistical significance ($P = 0.05$). The inexperienced 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 inexperienced 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 the asymmetry index [Figure 4]. 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 5 shows the classification performance evaluated using PSD and PLV features across four EEG frequency bands (delta, theta, alpha, and beta) under both task and rest conditions. Three classifiers (RFC, GPC, and ABC) were employed. Three feature sets were evaluated: (1) Asymmetry-only; (2) Asymmetry + EEG PSD; and (3) Asymmetry + EEG 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 ABC with 79%, while the 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 area under the curve (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 AUCs of 0.92 (RFC) and 0.85 (ABC), respectively. Among classifiers, the ABC consistently demonstrated higher and more stable performance, particularly in terms of F1 and AUC scores. In contrast, the 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.

Confusion matrix analysis

Figure 6 displays confusion matrices corresponding to the classification results for each frequency band and classifier. Each matrix illustrates the predicted *vs*. actual group labels (Inexperienced *vs*. Surgeon). ABC classifiers produced mixed predictions across both classes, but the overall classification accuracy remained limited. In several cases, surgeon samples were misclassified as controls. GPC classifiers frequently predicted all samples as a single class, most often the inexperienced 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 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%.

This presents the confusion matrices obtained from the combined 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 inexperienced 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 inexperienced participants. In contrast, the GPC classifier exhibited the weakest generalization, often misclassifying control trials as the surgeon class, suggesting model overfitting or limited feature separability.

DISCUSSION

In this section, the present findings demonstrate notable performance differences between surgeons and inexperienced participants during simulated task execution. Although not all group differences reached

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

PSD: Power spectral density; PLV: phase-locking value; RFC: random forest classifier; GPC: Gaussian process classifier; ABC: AdaBoost classifier; AUC: area under the curve.

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[9,40].

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[5].

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

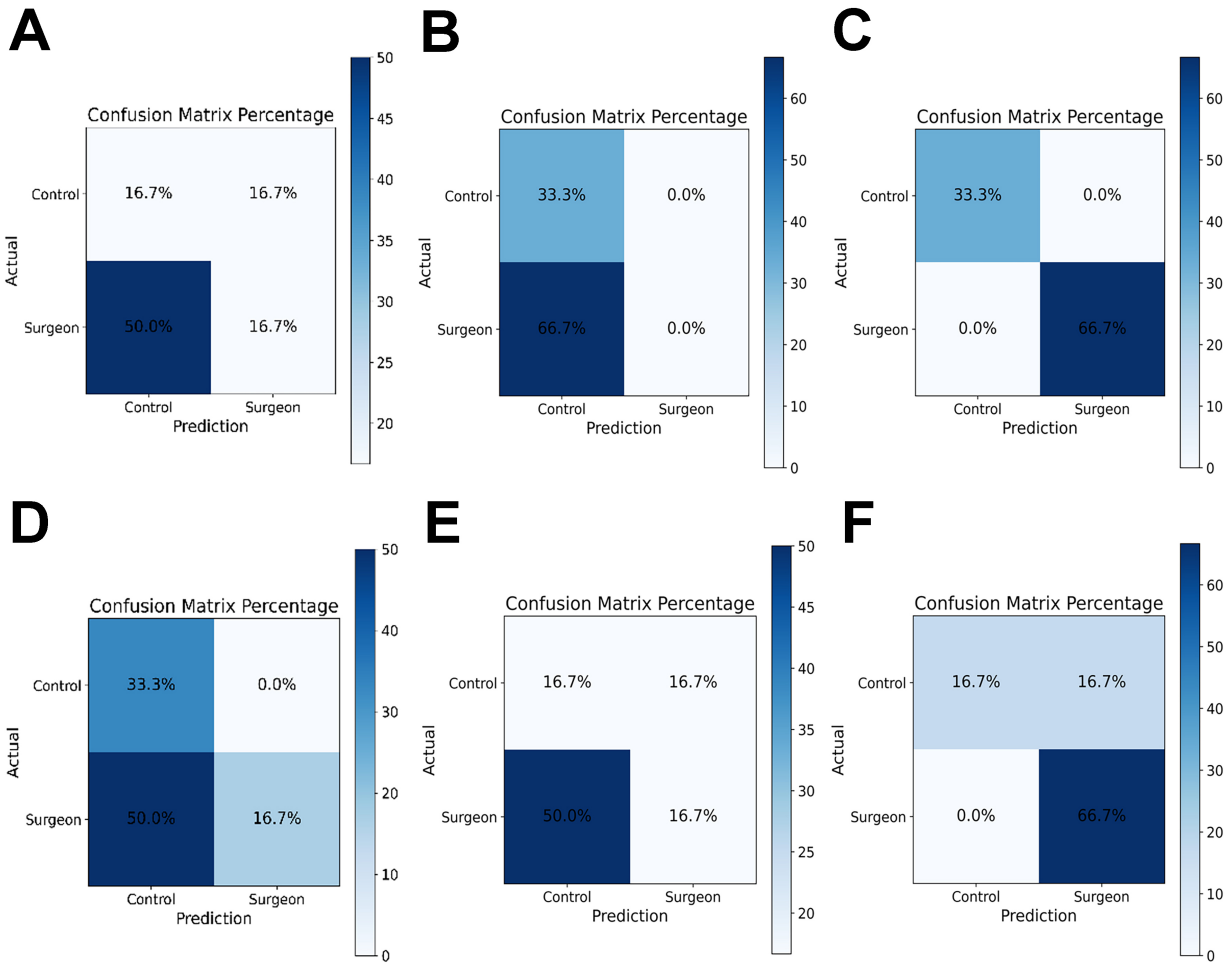


Figure 6. Representative confusion matrices for classification models using EEG features combined with tool-pressure asymmetry. (A-C) Results obtained using EEG-PSD features combined with asymmetry; and (D-F) results obtained using EEG-PLV features combined with asymmetry. Columns correspond to the three classifiers: ABC (A and D); GPC (B and E); and RFC (C and F). Each matrix shows predicted *vs.* actual group labels (Inexperienced *vs.* Surgeon). Each matrix shows predicted *vs.* actual group labels (Inexperienced *vs.* Surgeon), highlighting differences in model generalization across classifiers and feature sets. EEG: Electroencephalography; PSD: power spectral density; PLV: phase-locking value; ABC: AdaBoost classifier; GPC: Gaussian process classifier; RFC: random forest classifier.

execution[41-43]. 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.

Several global force-related features (e.g., maximum slope, force-time integral, and smoothness) did not show significant group differences. This finding may be partly explained by the relatively limited variability in expertise within the surgeon cohort, which primarily consisted of early- to mid-career practitioners, potentially reducing inter-individual differences in global force-production patterns. Moreover, these metrics predominantly capture overall force magnitude and execution dynamics, which may converge across groups when task demands are moderate and do not require substantial force modulation. In contrast, coordination-sensitive metrics such as asymmetry measures may remain more responsive to subtle differences in bimanual motor control. Future studies including a broader expertise spectrum and more demanding task conditions are warranted to further evaluate the discriminative sensitivity of these pressure-derived features.

Physical Demand showed a statistically significant positive association with asymmetry ($P < 0.05$), indicating that participants who exhibited greater bimanual force imbalance tended to report higher levels of perceived physical workload. This finding suggests that increased asymmetry may reflect inefficiencies in bimanual motor coordination that manifest as elevated physical demands during task execution. In contrast, Mental Demand, Temporal Demand, Performance, Effort, and Frustration did not demonstrate statistically significant linear relationships with asymmetry, although weak, non-significant trends were observed for Mental and Temporal Demand. Overall, these results suggest that the asymmetry index is more closely related to motor and force-related aspects of task execution than to broader cognitive or affective workload components, supporting its interpretation as a coordination-sensitive motor performance indicator rather than a global workload measure.

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[44,45]. 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 the literature suggest that combining neurophysiological and behavioral/performance features can enhance classification performance in expertise assessment[46]. 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.

PLV can capture functional coupling between cortical regions; however, in the present study, PLV features showed lower classification performance than PSD features. This outcome may be explained by the higher sensitivity of PLV estimation to noise and limited spatial sampling, particularly in low-density mobile EEG configurations. In contrast, PSD features reflect local oscillatory power and tend to be more robust under sparse electrode layouts and realistic training conditions. Furthermore, the laparoscopic tasks investigated in this study primarily involve motor execution and workload modulation rather than stable inter-regional phase synchronization, which may favor power-based features. In addition, reliable PLV estimation generally requires longer stationary epochs and higher signal-to-noise ratios than the short and dynamic task segments analyzed here.

A key limitation of this study is the relatively small sample size ($N = 23$). This was primarily driven by the logistical challenges of recruiting surgeons and acquiring synchronized multimodal data, including EEG and tool-pressure measurements, during laparoscopic simulation tasks. While this cohort size is consistent with prior exploratory studies in neuroergonomics, it may limit statistical power and the generalizability of the classification results. Therefore, the reported machine learning performance should be interpreted as preliminary. Future work will focus on larger, multi-center datasets to improve robustness and external validity. The relatively limited range of surgical expertise, particularly the small number of highly experienced surgeons with extensive laparoscopic caseloads, may have constrained group separability and classification performance. A further limitation is that participants' video-gaming experience was not systematically assessed, which may influence bimanual motor performance. Future studies should include standardized gaming history measures and non-surgical bimanual control tasks to determine whether the observed effects are specific to surgical expertise. Given the relatively small sample size, the reported accuracy values should be interpreted as preliminary estimates and require validation in larger independent cohorts. A single laparoscopic task was used, which may limit the generalizability of the findings to other surgical procedures with different motor and cognitive demands. Future studies should evaluate the proposed framework across multiple tasks to confirm its broader applicability.

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 laparoscopic 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.

In conclusion, this study examined whether combining tool pressure metric with EEG features can improve the classification of laparoscopic surgical skill. The results indicate that pressure asymmetry provides meaningful information between surgeons and inexperienced participants. When combined with EEG features, particularly in the theta band, classification performance improved. These findings suggest that multimodal neural and motor measures may support more objective assessment of surgical expertise and could contribute to future data-driven surgical training systems.

DECLARATIONS

Acknowledgments

The authors would like to thank Ahmet Omurtag, PhD, for his valuable feedback and guidance during the design and interpretation of this study. They also express their gratitude to the Etlik City Hospital, Departments of General Surgery and Obstetrics and Gynecology, for their support and collaboration during data collection. A preprint version of this manuscript was previously posted on Research Square (https://assets-eu.researchsquare.com/files/rs-7996713/v1/53628bbc-fd96-474a-983b-14113fea504d.pdf?c=1762154335).

Authors' contributions

Conceptualization, methodology, supervision, data analysis, and manuscript writing: Keles HO
Data collection and figure preparation: Sahin SS
Data acquisition and signal processing: Zengin C
All authors read and approved the final manuscript.

Availability of data and materials

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

AI and AI-assisted tools statement

Not applicable.

Financial support and sponsorship

This work was supported by the Türkiye Sağlık Enstitüleri Başkanlığı (TÜSEB) under the Program B Project No. 33115. The funding body had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

This study was approved by the Ankara University Human Research Ethics Committee (Approval No. 2024-000333-1) and conducted in accordance with the principles of the Declaration of Helsinki (2013 revision). Written informed consent was obtained from all participants prior to data collection.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. Omurtag A, Sunderland C, Mansfield NJ, Zakeri Z. EEG connectivity and BDNF correlates of fast motor learning in laparoscopic surgery. *Sci Rep.* 2025;15:7399. DOI PubMed PMC
2. Bonrath EM, Dedy NJ, Zevin B, Grantcharov TP. Defining technical errors in laparoscopic surgery: a systematic review. *Surg Endosc.* 2013;27:2678-91. DOI PubMed
3. Buia A, Stockhausen F, Hanisch E. Laparoscopic surgery: a qualified systematic review. *World J Methodol.* 2015;5:238-54. DOI PubMed PMC
4. Birkmeyer JD, Finks JF, O'Reilly A, et al.; Michigan Bariatric Surgery Collaborative. Surgical skill and complication rates after bariatric surgery. *N Engl J Med.* 2013;369:1434-42. DOI
5. Keles HO, Cengiz C, Demiral I, Ozmen MM, Omurtag A. High density optical neuroimaging predicts surgeons's subjective experience and skill levels. *PLoS ONE.* 2021;16:e0247117. DOI PubMed PMC
6. Healey MA, Shackford SR, Osler TM, Rogers FB, Burns E. Complications in surgical patients. *Arch Surg.* 2002;137:611-7. DOI PubMed
7. Soangra R, Sivakumar R, Anirudh ER, Reddy Y SV, John EB. Evaluation of surgical skill using machine learning with optimal wearable sensor locations. *PLoS ONE.* 2022;17:e0267936. DOI PubMed PMC
8. Shafiei SB, Shadpour S, Mohler JL, et al. Developing surgical skill level classification model using visual metrics and a gradient boosting algorithm. *Ann Surg Open.* 2023;4:e292. DOI PubMed PMC
9. Zakeri Z, Mansfield N, Sunderland C, Omurtag A. Physiological correlates of cognitive load in laparoscopic surgery. *Sci Rep.* 2020;10:12927. DOI PubMed PMC
10. Darzi A, Datta V, Mackay S. The challenge of objective assessment of surgical skill. *Am J Surg.* 2001;181:484-6. DOI PubMed
11. Farah E, Desir A, Marques C, et al. Heart rate variability: an objective measure of mental stress in surgical simulation. *Global Surg Educ.* 2024;3:25. DOI
12. The AF, Reijmerink I, van der Laan M, Cnossen F. Heart rate variability as a measure of mental stress in surgery: a systematic review. *Int Arch Occup Environ Health.* 2020;93:805-21. DOI PubMed PMC
13. Huaulmé A, Tronchot A, Thomazeau H, Jannin P. Automated assessment of non-technical skills by heart-rate data. *Int J Comput Assist Radiol Surg.* 2025;20:561-8. DOI PubMed
14. Soangra R, Jiang P, Haik D, et al. Beyond efficiency: surface electromyography enables further insights into the surgical movements of urologists. *J Endourol.* 2022;36:1355-61. DOI PubMed
15. Soto Rodriguez NA, Arroyo Kuribreña C, Porras Hernández JD, et al. Objective evaluation of laparoscopic experience based on muscle electromyography and accelerometry performing circular pattern cutting tasks: a pilot study. *Surg Innov.* 2023;30:493-500. DOI PubMed

16. Manabe T, Rahul FNU, Fu Y, et al. Distinguishing laparoscopic surgery experts from novices using EEG topographic features. *Brain Sci.* 2023;13:1706. DOI PubMed PMC
17. Nemani A, Yücel MA, Kruger U, et al. Assessing bimanual motor skills with optical neuroimaging. *Sci Adv.* 2018;4:eaat3807. DOI PubMed PMC
18. Nemani A, Kruger U, Cooper CA, Schwaitzberg SD, Intes X, De S. Objective assessment of surgical skill transfer using non-invasive brain imaging. *Surg Endosc.* 2019;33:2485-94. DOI PubMed PMC
19. Gao Y, Yan P, Kruger U, et al. Functional brain imaging reliably predicts bimanual motor skill performance in a standardized surgical task. *IEEE Trans Biomed Eng.* 2021;68:2058-66. DOI PubMed PMC
20. Zia A, Sharma Y, Bettadapura V, Sarin EL, Essa I. Video and accelerometer-based motion analysis for automated surgical skills assessment. *Int J Comput Assist Radiol Surg.* 2018;13:443-55. DOI PubMed
21. Franco-gonzález IT, Lappalainen N, Bednarik R. Tracking 3D motion of instruments in microsurgery: a comparative study of stereoscopic marker-based vs. deep learning method for objective analysis of surgical skills. *Inform Med Unlocked.* 2024;51:101593. DOI
22. Tien T, Pucher PH, Sodergren MH, Sriskandarajah K, Yang GZ, Darzi A. Eye tracking for skills assessment and training: a systematic review. *J Surg Res.* 2014;191:169-78. DOI PubMed
23. Oh J, Lau N. Quantitative analysis of eye-gaze metrics in differentiating surgical expertise. *Proc Hum Factors Ergon Soc Annu Meet.* 2024;68:624-6. DOI
24. Alleblas CC, Vleugels MP, Nieboer TE. Ergonomics of laparoscopic graspers and the importance of haptic feedback: the surgeons' perspective. *Gynecol Surg.* 2016;13:379-84. DOI PubMed PMC
25. Araki A, Makiyama K, Yamanaka H, et al. 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.* 2017;31:1999-2005. DOI PubMed
26. Sugiyama T, Lama S, Gan LS. Forces of tool-tissue interaction to assess surgical skill level. *JAMA Surg.* 2018;153:234-42. DOI PubMed PMC
27. Brown JD, O Brien CE, Leung SC, Dumon KR, Lee DI, Kuchenbecker KJ. Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. *IEEE Trans Biomed Eng.* 2017;64:2263-75. DOI PubMed
28. Rafii-Tari H, Payne CJ, Bicknell C, et al. Objective assessment of endovascular navigation skills with force sensing. *Ann Biomed Eng.* 2017;45:1315-27. DOI PubMed PMC
29. Golahmadi AK, Khan DZ, Mylonas GP, Marcus HJ. Tool-tissue forces in surgery: a systematic review. *Ann Med Surg.* 2021;65:102268. DOI PubMed PMC
30. Dockter RL, Lendvay TS, Sweet RM, Kowalewski TM. The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. *Int J Comput Assist Radiol Surg.* 2017;12:1151-9. DOI PubMed
31. Gao Y, Kruger U, Intes X, Schwaitzberg S, De S. A machine learning approach to predict surgical learning curves. *Surgery.* 2020;167:321-7. DOI PubMed PMC
32. Ebina K, Abe T, Yan L, et al. Development of machine learning-based assessment system for laparoscopic surgical skills using motion-capture. In: Proceedings of the 2024 IEEE/SICE International Symposium on System Integration (SII); 2024 Jan 8-11; Ha Long, Vietnam. New York: IEEE; 2024. pp. 1-6. DOI
33. Power D, Burke C, Madden MG, Ullah I. Automated assessment of simulated laparoscopic surgical skill performance using deep learning. *Sci Rep.* 2025;15:13591. DOI PubMed PMC
34. Natheir S, Christie S, Yilmaz R, et al. Utilizing artificial intelligence and electroencephalography to assess expertise on a simulated neurosurgical task. *Comput Biol Med.* 2023;152:106286. DOI PubMed
35. Yin S, Xiang Z. Multi-objective collaborative path planning for heterogeneous autonomous underwater vehicles in cluttered environments. *Swarm Evol Comput.* 2026;100:102251. DOI
36. Yin S, Xiang Z. Adaptive collision avoidance strategy for USVs in perception-limited environments using dynamic priority guidance. *Adv Eng Inform.* 2025;65:103355. DOI
37. Omurtag A, Roy RN, Dehais F, Chatty L, Garbey M. Tracking mental workload by multimodal measurements in the operating room. Neuroergonomics. Elsevier; 2019. pp. 99-103. DOI
38. Muthukumaraswamy SD. High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Front Hum Neurosci.* 2013;7:138. DOI PubMed PMC
39. Vinck M, Oostenveld R, van Wingerden M, Battaglia F, Pennartz CM. An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *Neuroimage.* 2011;55:1548-65. DOI PubMed PMC

40. Kim HJ, Choi GS, Park JS, Park SY. Comparison of surgical skills in laparoscopic and robotic tasks between experienced surgeons and novices in laparoscopic surgery: an experimental study. *Ann Coloproctol.* 2014;30:71-6. DOI PubMed PMC
41. Dias RD, Ngo-Howard MC, Boskovski MT, Zenati MA, Yule SJ. Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. *Br J Surg.* 2018;105:491-501. DOI PubMed PMC
42. Hannah TC, Turner D, Kellner R, Bederson J, Putrino D, Kellner CP. Neuromonitoring correlates of expertise level in surgical performers: a systematic review. *Front Hum Neurosci.* 2022;16:705238. DOI PubMed PMC
43. Howie EE, Dharanikota H, Gunn E, et al. Cognitive load management: an invaluable tool for safe and effective surgical training. *J Surg Educ.* 2023;80:311-22. DOI PubMed
44. Balkhoyor AM, Awais M, Biyani S, et al. Frontal theta brain activity varies as a function of surgical experience and task error. *BMJ Surg Interv Health Technol.* 2020;2:e000040. DOI PubMed PMC
45. Akkad H, Dupont-Hadwen J, Kane E, et al. Increasing human motor skill acquisition by driving theta-gamma coupling. *Elife.* 2021;10:e67355. DOI PubMed PMC
46. Shafiei SB, Shadpour S, Mohler JL, et al. Surgical skill level classification model development using EEG and eye-gaze data and machine learning algorithms. *J Robot Surg.* 2023;17:2963-71. DOI PubMed PMC

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.


© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.