2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Neuro-Robot Interaction in Robot-Assisted Surgery using EEG and Self-Supervised Graph Transformer

Debashis Das Chakladar*, Foteini Simistira Liwicki and Rajkumar Saini

Abstract— **Robot-Assisted Surgery (RAS) represents a major frontier in the robotics community, blending precision automation with human skill in high-stakes clinical environments. Evaluating surgeon performance in RAS is critical for training and certification, yet current methods rely heavily on video analysis or subjective manual scoring. This study presents a neuro-robotic interaction framework that uses Electroencephalography (EEG)-derived brain connectivity features to classify surgeons' skill levels during RAS tasks. The high dimensionality of EEG data imposes substantial computational cost. Therefore, we first apply Harris Hawks Optimization (HHO) to select an optimal EEG-channel subset, reducing computational cost. Then, functional connectivity feature metrics are extracted from the reduced EEG channel set and used to construct brain graphs, which serve as input to a Self-Supervised Graph Transformer (SSGT). The SSGT model is pre-trained via masked edge reconstruction to capture structural dependencies and fine-tuned for downstream skill-level classification. The proposed SSGT model achieves a classification accuracy of 96.60%, significantly outperforming both traditional machine learning and deep learning baselines. The label-efficient, structurally aware design of SSGT enables scalable and real-time assessment of surgical proficiency. This framework provides a foundation for intelligent robotic tutoring systems and generalizes to broader cognitive monitoring tasks in high-stakes human-robot interaction domains using EEG.**

I. INTRODUCTION

Electroencephalography (EEG) is widely used in surgery [8], cognitive workload level estimation [9]–[11], attention monitoring [12], and learning effects [13] that are crucial in a human-robot interaction setting [14]. In performance–monitoring problems, EEG features have been shown to discriminate expertise levels and task proficiency with competitive accuracy. A robotics transformer trained on large, task-agnostic datasets can transfer knowledge to new robot tasks with minimal data, enabling robots to respond more effectively to human goals during human–robot interaction (HRI) [15]. A deep transformer model serves as a conditional intent predictor, attending jointly to human and robot action sequences to capture their interdependence [16]. Choi *et al.* [17] demonstrated intuitive robotic control during EEG-based motor intentions—such as movement onset or motor imagery tasks, emphasizing its potential to enhance intention-driven HRI. In study [18], the author developed a novel EEG-based vision transformer model enhanced with a 3D spatial representation of EEG electrode topologies, enabling reliable recognition of human trust levels during interactive gameplay with robots. HRI in Robot-Assisted Surgery (RAS) has been shown to benefit from real-time monitoring of surgeons' cognitive states via EEG, enabling adaptive systems to modulate robot behavior when operator workload increases, thus improving safety and performance [19]. However, surgical skill assessment in RAS with a large number of EEG channels (i.e, 128 or 256) makes the model computationally expensive, which causes problems while integrating it with surgical robot systems [2]. Therefore, selection of a task-specific reduced, significant EEG channel set is needed for better performance [20]. Metaheuristic algorithms such as Grey Wolf Optimizer [21], Whale Optimization [22] are widely used for eliminating non-significant channels and finding optimum feature space for EEG-based classification. After removing redundant channels, brain-connectivity features—correlation, coherence [23], and transfer entropy [24]—are extracted from the remaining non-redundant channels to characterize task-related brain dynamics.

A strong connection has been found between surgical skill assessment and EEG-based brain networks, which reveals the neural signature of skill level estimation during robotic surgery [25]. Shafiei *et al.* [7] predicted the skill levels (inexperienced, competent, and experienced) of surgeons while performing different RAS subtasks (blunt, coldsharp, and thermal dissection) by fusing EEG and eye-tracking data. They have achieved the highest accuracy of 88.56% with the Random Forest (RF) model. Lee *et al.* [3] developed a convolutional neural network (CNN)-based method that tracks surgical instruments to evaluate surgeons' skills (novice, intermediate, and expert) during robotic operations. A CNN-based method has been developed to assess surgical skill from raw motion-kinematics in robotic surgery [5]. Wang *et al.* [1] examine the task of urethrovesical anastomosis in robotic-assisted radical prostatectomy using synthetic tissue, where the GEARS (Global Evaluative Assessment of Robotic Skills) score is predicted from video clips. Their method tracks key points of surgical instruments over time and trains a multi-task U-Net neural network to enable robotic-training evaluation by automating skill assessment. Summary of the RAS skill classification studies is mentioned in Table I. Zhou *et al.* [26] combined EEG and heart rate to detect changes in mental workload during RAS, reaching 83.20% accuracy and underscoring the value of EEG for cognitive state monitoring. Shafiei *et al.* [27] compute mental workload levels across multiple RAS-related tasks—matchboard, ring walk, and suturing—using synchronized EEG and eye-tracking data. They achieved a strong prediction result ($R^2 \approx 0.81$–0.83)

*Corresponding author, Email: ddaschakladar@gmail.com
All the authors are with the Machine Learning Group at Luleå University of Technology, 97187 Luleå, Sweden

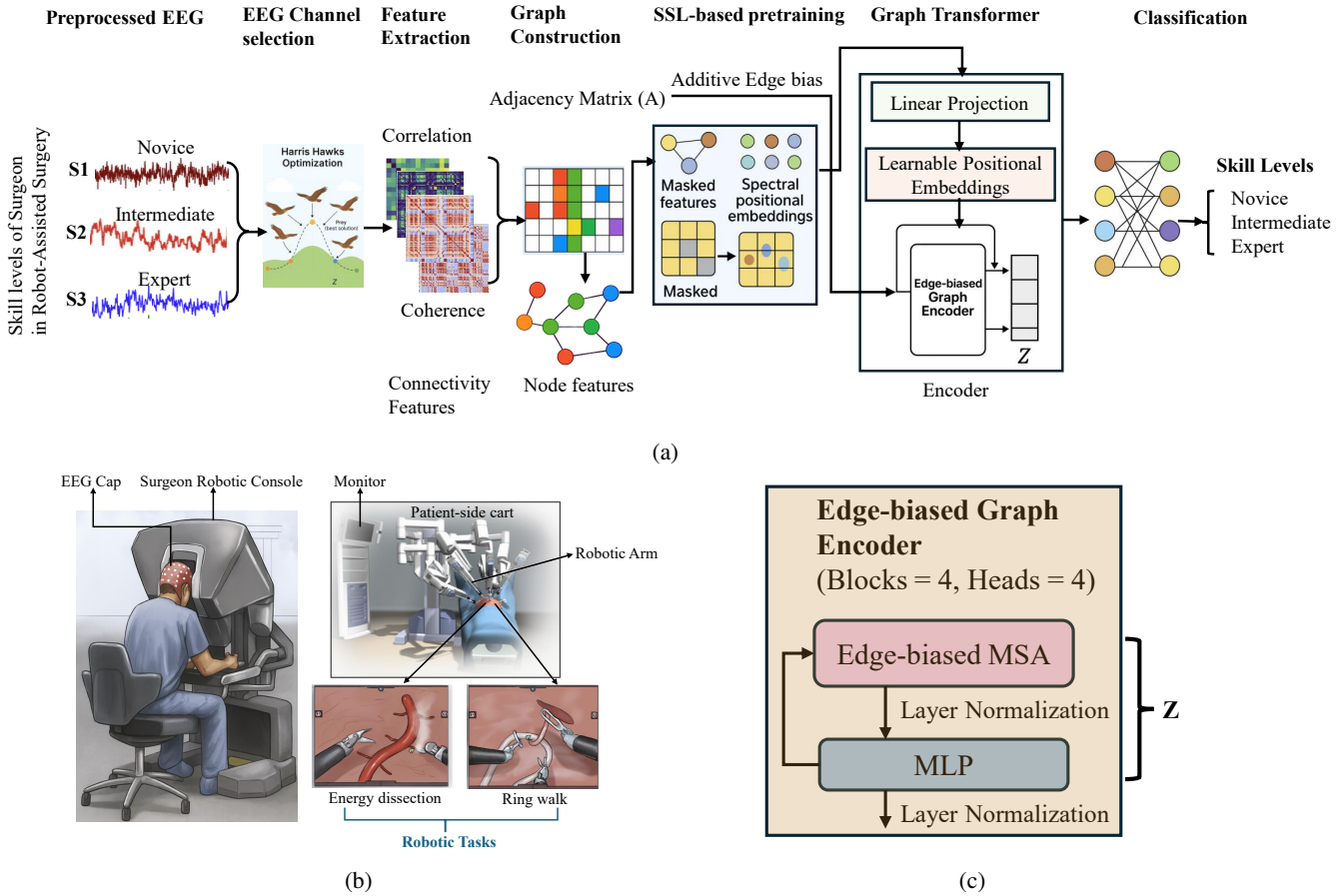


Fig. 1: (a) Self-Supervised Graph Transformer (SSGT) framework for classifying surgeons' skill levels (novice, intermediate, expert) during Robot-Assisted Surgery (RAS) tasks. The encoder of the graph transformer includes **edge-biased graph encoder**. (b) RAS task paradigm. The EEG signal recorded from an EEG cap served as input to the proposed model. The surgeons are denoted S1, S2, and S3 and are grouped by their RAS experience. (c) Block diagram of the **edge-biased graph encoder**. The loop denotes four repeated blocks. The encoder output (Z) feeds the dense classification head in (a). Note: MSA = multi-head self-attention, SSL = self-supervised learning, MLP = multilayer perceptron.

TABLE I: Summary of RAS skill level classification studies. Note: Logistic Regression (LR), k-nearest neighbours (kNN), Support vector machine (SVM), Convolutional Neural Network (CNN), Random Forest (RF)

Study	Participants	Data	Classes	Classification Model	Accuracy (%)
Wang Y.*et al.* [1]	18	Video (RAS simulator)	Skill level: novice / intermediate / expert	U-Net–style encoder–decoder	83.30
Shafiei *et al.* [2]	11	EEG + Eye-gaze (RAS simulator)	Skill level: novice / intermediate / expert	Gradient boosting	93.00
Lee *et al.* [3]	12	Video recordings	Skill level: novice / intermediate / expert	CNN	83.00
Soangra *et al.* [4]	26	Kinematic & EMG (Lap. & RAS)	Skill level: novice / intermediate / expert	RF	58.00
Wang *et al.* [5]	8	Video & kinematic (simulator)	Skill level: novice / intermediate / expert	CNN	95.40
Fard *et al.* [6]	8	Movement trajectory data	Skill level: novice / expert	SVM/LR/k-NN	Max: 89.90 (LR)
Shafiei *et al.* [7]	11	EEG & Eye-tracking	Skill level: inexperienced/competent/experienced	RF	88.56

using an XGBoost regressor.

Despite this progress, several research gaps exist. Most of the existing RAS studies [2], [7] used skill-level assessment with a large set of EEG channels, leading to a computationally expensive model. Therefore, EEG channel reduction has remained largely unexplored in RAS applications. Moreover, existing classification pipelines typically rely on computationally expensive wrapper-based Harris Hawks Optimization (HHO) feature selectors [28] and fully supervised, traditional machine learning (ML) or deep learning (DL) models [2], [3], [7]. In contrast, our HHO-based feature selection utilizes an efficient, unsupervised filter mechanism—maximizing the Shannon entropy of raw EEG amplitude histograms without iterative classifier training. This optimized feature set is then processed by a robust DL model featuring self-supervised pretraining to exploit the structure of EEG connectomes effectively.

In this work, we propose the Self-Supervised Graph Transformer (SSGT)-based method that combines self-supervised learning (SSL) and a graph transformer to estimate the skill levels (novice, intermediate, and expert) of surgeons during RAS tasks. Initially, to reduce redundancy and improve

deployability in the surgical robot system, we select the optimum set of EEG channels using an HHO-based feature selector from the input EEG channels. Then, we build the connectivity features from the optimum channel set and prepare the adjacency graph from those features. Finally, we employ an SSL block to mask and reconstruct the original adjacency matrix of the EEG graph, and then feed the resulting embeddings into a graph transformer classifier that learns both spatial and temporal dependencies for skill level prediction. The framework of the proposed SSGT model is shown in Fig. 1. The contributions of our work are as follows:

1) We proposed a robust DL-based transformer model-SSGT for efficient classification of surgeons' skill levels during different RAS tasks. To the best of our knowledge, this is the first label-efficient EEG-based RAS method that exploits graph structure, complementing earlier fully supervised work.
2) HHO-based channel selection method is developed to find the most significant, non-redundant EEG channels. The reduced set of channels is used for downstream classification, reducing the computational cost of the classification model and supporting real-time HRI feedback.
3) The proposed model outperforms all ML/ DL baseline classifiers in surgical skill assessment in the RAS application.

II. METHODOLOGY

A. Dataset & Preprocessing

In this work, we have used a public RAS dataset [29]. The dataset includes twenty-five participants (age 20–67 years) with diverse RAS experience. Twelve participants had no RAS experience; four had less than 100 hours, four had ≈500 hours, and five had more than 1000 hours of RAS experience. Sampling frequency of the EEG recording was 500 Hz. Each participant completed 27 simulator tasks on the da Vinci system while wearing a 128-channel AntNeuro EEG cap and Tobii eye-tracking glasses. However, in this experiment, we used only the EEG data from the 13 participants with RAS experience. Out of 27 tasks, we used five most significant RAS tasks, such as: Pick and place (task Id-1), Peg board 1 (task Id- 2), Camera targeting 1 (task Id-9), Needle targeting (task Id-15), and Energy switching 1 (task Id-23) [30]–[33]. These tasks span the principal RAS skill domains, which provide complementary, task-resolved evidence of surgical proficiency [30].

Since the dataset lacks a mapping between subjects and their skill levels, we derive the labels directly from the performance scores. For each subject, we compute the *mean* performance score across the selected tasks and discretize this continuous value into three skill categories: **Novice** if the mean score < 75, **Intermediate** if $75 \leq$ mean score ≤ 90, and **Expert** if the mean score > 90. In RAS, small errors can have serious clinical consequences, so a mean score below 75 is considered clearly below the minimum safe competency level and labeled as **Novice**. Only operators with mean scores above 90, reflecting consistently safe, accurate, and efficient performance in demanding RAS tasks, are labeled as **Experts**, and between **Novice** and **Experts** are marked as **Intermediate** [34]. The raw EEG signals are preprocessed using a band pass filter (frequency range: 1-40 Hz) followed by Independent Component Analysis to remove artifacts from the raw EEG signals.

B. Channel Selection and Feature Extraction

To eliminate the **curse of dimensionality** issue, we compute the channel selection process, which selects the optimum channel set from the preprocessed EEG (128 channels) for the experiment. Here, we use **HHO** [35] algorithm for finding optimal EEG channels.

Let $X \in \mathbb{R}^{C\times T}$ denote the preprocessed EEG (channels $\times$ time). We seek a subset $S \subset \{1, \ldots, C\}$ of fixed size $|S| = k$ that *maximizes* the average information content of the selected channels. For a EEG channel c, we estimate Shannon entropy–$H(x_c)$ from a normalized histogram with B number of histogram bins

$$H(x_c) \;=\; -\sum_{b=1}^{B} p_b^{(c)} \,\log\big(p_b^{(c)} + \varepsilon\big), \tag{1}$$

with a small $\varepsilon > 0$ for numerical stability and p_b is the normalized histogram bin probability. The objective is

$$\begin{aligned} \max_{S:\,|S|=k} \;\frac{1}{k}\sum_{c\in S} H(x_c) \;\equiv\; \min_{S:\,|S|=k}\; f(S), \\ \text{where} \quad f(S) = -\frac{1}{k}\sum_{c\in S} H(x_c)\,. \end{aligned} \tag{2}$$

The fitness $f(S)$ is computed for a subset of channels S. For each channel $c \in S$, we estimate its $H(x_c)$ from a histogram of its amplitudes; **Higher-entropy channel subsets** (more informative/less redundant) receive **lower** fitness values and are preferred by HHO. Each candidate (subset S) is scored by $f(S)$. After T iterations, HHO returns the selected set $S^\star$ of k channels, which maximizes the mean entropy of the chosen EEG channels.

After selecting k optimum channels, we extract correlation and coherence features from the selected channels' EEG data. The Pearson correlation coefficient measures the linear relationship between signals from two EEG channels. This coefficient ranges from -1 to 1. It is commonly used to summarize co-fluctuations across EEG channels. We used another feature: coherence, which discusses how consistently two signals co-vary at a given frequency across time. High coherence indicates stable, frequency-specific synchronization of amplitude and/or phase, whereas low coherence indicates weak or inconsistent coupling.

C. Graph Creation

We construct a graph $G = (N, E)$ from each EEG-based connectivity matrix, where each node $n_i \in N$ represents an EEG channel, and the edge $e_{ij} \in E$ denotes the functional connectivity (correlation or coherence) between channel i and j. Given an EEG signal segment, its corresponding

adjacency matrix $A \in \mathbb{R}^{N\times N}$ is computed from either correlation or coherence features. To construct node features, we aggregate three components: (1) the connectivity vector of each node (row of A), (2) the normalized node degree $d_i = \sum_j A_{ij}$, and (3) the spectral positional embeddings $V \in \mathbb{R}^{N\times k}$ obtained by computing the k smallest eigenvectors of the normalized Laplacian $L = I - D^{-1/2}AD^{-1/2}$, where I and D denote the identity and degree matrices, respectively. The final node feature matrix becomes:

$$X' = [A, d, S] \in \mathbb{R}^{N\times(N+1+k)} \tag{3}$$

where d is a column vector of normalized degrees and S captures the spectral structure of the graph.

D. Classification using Self-Supervised Graph Transformer

In this section, we discuss the skill level classification using the SSGT model. While transformers typically consist of both an encoder and decoder, in our architecture, the decoder is employed only during self-supervised pre-training and is not part of the supervised classification pipeline. Thus, once pre-training concludes, the decoder is discarded. The detailed discussion is mentioned in subsequent subsections.

1) Self-Supervised Graph Transformer: The proposed SSGT model combines two modules: Self-Supervised Learning (SSL)-based pretraining and Graph Transformer.

SSL-based pretraining: SSL trains models on unlabeled data by creating supervision from the data itself via pretext tasks (e.g., masking–reconstruction or agreement between augmented views) [36]. By learning domain structure without manual labels, SSL produces rich features that, when fine-tuned with a small labeled set, markedly improve classification by boosting robustness and generalization. To enhance representations on graphs, we adopt **masked edge reconstruction**. During pre-training (100 epochs), random edges in the adjacency matrix A are masked with a symmetric binary mask M, producing a masked/corrupted graph $A' = A\odot(1-M)$, where, $\odot$ denotes the element-wise (Hadamard) product. Node tokens are then built from the masked graph using the following formula:

$$X = \big[A' \parallel d(A') \parallel S(A')\big], \tag{4}$$

where $d(A') = \sum_j A'_{ij}$ is the (normalized) degree and $S(A')$ are *spectral positional embeddings* applied on A'. Here, $\parallel$ denotes the concatenation operator. Thus, the spectral positional embeddings are *tied to the masked structure* (no leakage of the hidden edges) while providing topology-aware coordinates that aid reconstruction. The encoder is trained to reconstruct the original A from X and A', and the pre-training loss is binary cross-entropy (BCE) on the masked edges (i, j) only:

$$\mathcal{L}_{\text{SSL}} = \sum_{(i,j)\in\mathcal{M}} \text{BCE}\big(A_{ij}, \hat{A}_{ij}\big), \quad \mathcal{M} = \{(i,j) : M_{ij} = 1\} \tag{5}$$

Graph Transformer: As depicted in Fig.1, the encoder consumes *only* the outputs of the SSL block: the node tokens $X \in \mathbb{R}^{N\times F}$ and the corresponding adjacency matrix $A \in \mathbb{R}^{N\times N}$, where N and F are the number of nodes and feature dimension per node (obtained from SSL block). Tokens are first linearly projected to the model width d and augmented with a learnable positional embedding for each node. The encoder then applies 4 stacked transformer blocks with 4 heads. In each block, we use a pre-normalized, edge-biased multi-head self-attention (MSA) followed by a Multilayer perceptron (MLP) with hidden dimension $2d = 256$. Throughout these blocks, residual connections and layer normalization are applied. For each attention head, we compute the edge-biased attention scores:

$$\text{MSA}_{\text{edge}}(X, A) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \beta A\right)V \tag{6}$$

Where, Q, K, V are the query, key, and value projections of the input node features, d_k is the head dimension, and βA is the additive edge bias derived from the adjacency matrix. Q, K, V are computed as follows: $Q{=}XW_Q$, $K{=}XW_K$, $V{=}XW_V$. β is a learnable (or fixed) scale and W is the learned weight matrices for the linear projections. This **edge-biased attention** formulation ensures the attention mechanism prioritizes connected nodes in the graph structure. Stacking the blocks yields node embeddings $Z \in \mathbb{R}^{N\times d}$, which are passed to the final classification layer.

2) Classification using SSGT: The node embeddings $Z \in \mathbb{R}^{N\times d}$ capture global graph structure and serve as a robust backbone for the downstream classification model. These embeddings are aggregated via global average pooling:

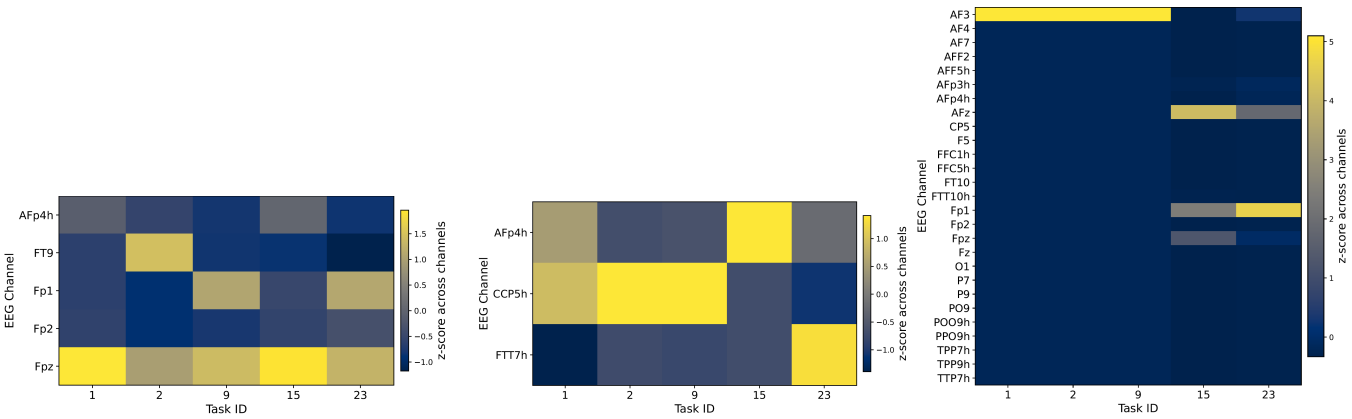
$$z = \frac{1}{N}\sum_{i=1}^{N} Z_i \tag{7}$$

The pooled feature vector (z) is then passed through the final dense classification layer to predict the surgeon's skill level (novice, intermediate, expert).

III. Results

A. Brain activation analysis of RAS tasks

In this section, we discussed the channel selection process using the HHO method. The optimization method selects specific channels for each skill level. The RAS task-wise brain activation maps (refer to Fig. 2) were derived by computing optimum channel–wise band power (1–40 Hz) and standardizing the values across channels within each task (z-score), so colors reflect *relative* activation of EEG channel over reference EEG channel rather than absolute power (μV^2). Relative power normalizes global amplitude differences, improving cross-channel/subject comparability and sensitivity to task effects over absolute power [37]. Across the three performance levels, a consistent pattern emerges: the **expert** cohort exhibits a broadly distributed set of highly activated channels spanning frontal brain regions (EEG channels: AF3, AFz, Fp1 refer to Fig. 2(c)). The **intermediate** cohort engages a smaller subset with moderate amplitudes in the EEG channels such as AFp4h, CCP5h, FTT7h (Fig. 2(b)), whereas the **novice** group concentrates activation in only a handful of frontal channels (Fpz/Fp1/Fp2



(a) (b) (c)

Fig. 2: RAS task-wise EEG channel activation plot for surgeons with different skill levels, namely (a) Novice, (b) Intermediate, and (C) Expert.

etc.) (Fig. 2(a)). More specifically, in the novice group, the EEG channel Fpz shows high activation across all tasks. In the intermediate group, CCP5h is dominant for tasks 1, 2, and 9, with AFp4h and FTT7h peaking at tasks 15 and 23, respectively. In the expert group, AF3 shows high activation for tasks 1, 2, and 9, whereas AFz and Fp1 show localized peaks at tasks 15 and 23.

B. Performance Analysis

This section is divided into two subsections: Output analysis and RAS task-wise analysis.

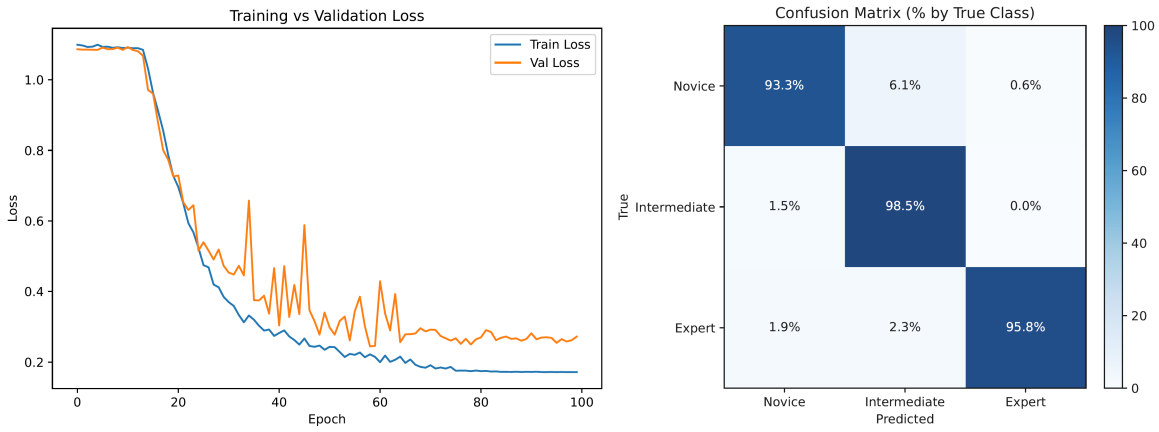
1) Output Analysis: Here, we report the performance analysis of the proposed SSGT model for classifying surgeons' performance levels. The input data is divided into an 80:20 ratio for the training and testing sets; furthermore, 10% of the training data is used for the validation set. The model is trained for 100 epochs with a batch size of 64 using the Adam optimizer (learning rate $= 5 \times 10^{-4}$). The loss curve in Fig. 3(a) show that the training loss decreases steadily and the validation loss closely tracks it, indicating stable learning and minimal overfitting. The model is implemented in TensorFlow and trained on NVIDIA A100 GPUs with 40 GB VRAM.

A confusion matrix summarizes a classifier's performance by counting predictions versus true labels for each class (here, novice, intermediate, and expert). **Recall/True prediction rate** refers to the fraction of correct predictions within each true class. The confusion matrix (Fig. 3(b)) indicates strong class-wise recall across all three skill levels: novice = 93.3%, intermediate = 98.5%, and expert = 95.8%. The intermediate class is the most separable, exhibiting negligible confusion with the other classes (only 1.5% misassigned to novice and 0.0% to expert). Overall, the model is highly accurate. It clearly separates the intermediate class, and little confusion occurs only between novice and expert; adding richer features (e.g., finer spectral or connectivity measures) could reduce these edge-case errors. The overall complexity of the proposed pipeline: $\mathcal{O}\big(PIkT + k^2T + L(k^2d + kd^2)\big)$ where P and I are the HHO population size and iterations, k the number of selected channels, T the number of time samples, L the number of transformer layers, and d the hidden dimension. For all $C = 128$ channels ($C > k$), the corresponding cost: $\mathcal{O}\big(C^2T + L(C^2d + Cd^2)\big)$ Thus, our HHO-based optimization method reduces the dominant connectivity and graph-learning cost by roughly a factor of $(k/C)^2$.

2) RAS task-wise Analysis: The task-wise classification result of the SSGT classifier is shown in Fig. 4. Here, we passed task-wise EEG data of all surgeons for each skill level to the SSGT model and computed the result. A similar process is repeated for other skill levels' data. Across the five RAS tasks (IDs 1, 2, 9, 15, 23), accuracy increases systematically with performance level—novice $\approx$ 60–75%, intermediate $\approx$67–87%, and expert $\approx$ 80–94%. For all skill levels, the maximum accuracy is obtained for Task ID-23. This monotonic ordering indicates that the model captures stable, task-locked neurophysiological signatures that scale with surgical proficiency. This task-level reliability enables objective skill auditing in RAS: instructors can pinpoint weak subtasks for novices/intermediates, while high performers show stable, generalizable patterns suitable for competency checks and adaptive feedback [38].

C. Statistical Analysis

In this section, we evaluate differences among three skill levels (novice, intermediate, and expert) using the Kruskal–Wallis test. For each skill level, we compute the task-wise classification accuracy for each participant using Leave-One-Subject-Out (LOSO) cross-validation (i.e., for each fold, one participant is used in the test set and the remaining subjects are used in the training set). Kruskal-Wallis test ($p < 0.05$) detected a significant difference among the three skill levels—novice, intermediate, and expert: $H(2) = 6.33$, $p = 0.042$, with a large effect size $\varepsilon^2 = 0.62$. In this test, task-averaged classification accuracy per subject is used as a dependent variable, whereas skill level with three groups is represented as an independent variable.



(a) (b)

Fig. 3: Output analysis of SSGT model: (a) Loss curve of the model, (b) Confusion matrix

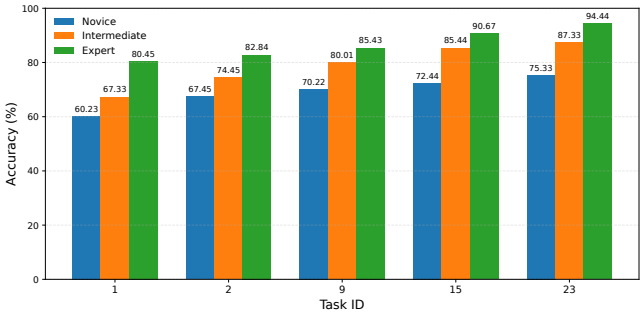


Fig. 4: RAS task-wise classification result of surgeons with three skill levels.

TABLE II: Comparison analysis with ML/DL models. All results are marked with mean ± standard deviation. Note: Vision Transformer — ViT, Swin Transformer — SwinT, Support Vector Machine — SVM, Random Forest — RF, eXtreme Gradient Boosting — XGBoosting

Model categories	Model	Accuracy (%)	Precision (%)	Recall (%)
ML	SVM	85.34 ± 4.12	85.03 ± 4.03	85.12 ± 3.44
	RF	80.34 ± 3.34	81.12 ± 3.22	80.23 ± 4.14
	XGBoosting	86.14 ± 2.26	86.03 ± 3.11	86.11 ± 2.34
DL	ViT	89.23 ± 1.14	89.11 ± 2.23	88.77 ± 4.45
	SwinT	91.11 ± 2.01	91.22 ± 6.45	90.77 ± 5.02
	Graphormer	93.43 ± 4.03	93.03 ± 4.40	92.19 ± 2.02
	SSGT (Proposed)	96.60 ± 1.23	97.23 ± 2.34	95.11 ± 4.32

D. Comparison analysis with other classification models

To generalize our model performance, we compare the performance of the proposed model (SSGT) with classical ML and DL models. The result is shown in Table II. For ML models (SVM, RF, XGBoosting), we use hand-crafted EEG **connectivity features** as inputs and train the classifiers, yielding accuracies in the 80.34–86.14% range. For DL models (ViT, SwinT, Graphormer), we adopt the same training protocol described in the output analysis but add **self-supervised pretraining** on unlabeled trials, followed by a transformer encoder for supervised fine-tuning, achieving higher accuracies of 89.23–93.43%. The proposed **SSGT** model attains $96.60 \pm 1.23\%$ accuracy—an improvement of **3.17%** over the best DL baseline (Graphormer) and **10.46%** over the best ML baseline (XGBoosting). As all models are trained under the same hyperparameters and evaluation settings, these gains highlight that the proposed SSGT model extracts task-relevant representations from limited labels, thereby outperforming the ML and DL baselines.

IV. CONCLUSIONS

We propose the SSGT, which leverages masked-edge reconstruction and edge-biased attention to learn rich and structurally aware representations from EEG-derived connectivity graphs. In the RAS setting—where objective and real-time feedback is crucial for improving surgical skill and patient outcomes—SSGT achieves a remarkable **96.60%** accuracy, significantly outperforming existing EEG-based ML/DL models in classifying surgeon skill levels during RAS tasks. This demonstrates the model's strong potential to enhance surgical training and assessment, aligning with the growing demand within the RAS community for automated, high-fidelity skill evaluation tools that can complement traditional expert review. The limitation of the study lies in computational constraints, which led us to restrict data analysis to only five of the 27 available tasks and to rely solely on EEG data from this subset.

Future efforts will focus on (i) **evaluating model performance on additional RAS tasks** to assess the robustness of the model; (ii) **performing an ablation study** of SSGT with and without the SSL backbone; and (iii) **developing a multimodal fusion framework** that integrates EEG with other data sources—such as eye-gaze, kinematics, or video—to further refine real-time, automated skill evaluation in RAS.

REFERENCES

[1] Y. Wang, J. Dai, T. N. Morgan, M. Elsaied, A. Garbens, X. Qu, R. Steinberg, J. Gahan, and E. C. Larson, "Evaluating robotic-assisted surgery training videos with multi-task convolutional neural networks," *Journal of robotic surgery*, vol. 16, no. 4, pp. 917–925, 2022.

[2] S. B. Shafiei, S. Shadpour, J. L. Mohler, F. Sasangohar, C. Gutierrez, M. Seilanian Toussi, and A. Shafqat, "Surgical skill level classification model development using eeg and eye-gaze data and machine learning

algorithms," *Journal of robotic surgery*, vol. 17, no. 6, pp. 2963–2971, 2023.

[3] D. Lee, H. W. Yu, H. Kwon, H.-J. Kong, K. E. Lee, and H. C. Kim, "Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations," *Journal of clinical medicine*, vol. 9, no. 6, p. 1964, 2020.

[4] R. Soangra, R. Sivakumar, E. Anirudh, S. V. Reddy Y, and E. B. John, "Evaluation of surgical skill using machine learning with optimal wearable sensor locations," *PLoS One*, vol. 17, no. 6, p. e0267936, 2022.

[5] Z. Wang and A. Majewicz Fey, "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery," *International journal of computer assisted radiology and surgery*, vol. 13, no. 12, pp. 1959–1970, 2018.

[6] M. J. Fard, S. Ameri, R. Darin Ellis, R. B. Chinnam, A. K. Pandya, and M. D. Klein, "Automated robot-assisted surgical skill evaluation: Predictive analytics approach," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, no. 1, p. e1850, 2018.

[7] S. B. Shafiei, S. Shadpour, J. L. Mohler, E. C. Kauffman, M. Holden, and C. Gutierrez, "Classification of subtask types and skill levels in robot-assisted surgery using eeg, eye-tracking, and machine learning," *Surgical Endoscopy*, vol. 38, no. 9, pp. 5137–5147, 2024.

[8] J. Thomas, C. Abdallah, K. Jaber, M. Khweileh, O. Aron, I. Doležalová, V. Gnatkovsky, D. Mansilla, P. Nevalainen, R. Pana *et al.*, "Development of a stereo-eeg based seizure matching system for clinical decision making in epilepsy surgery," *Journal of neural engineering*, vol. 21, no. 5, p. 056025, 2024.

[9] D. D. Chakladar, S. Datta, P. P. Roy, and V. A. Prasad, "Cognitive workload estimation using variational autoencoder and attention-based deep model," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 2, pp. 581–590, 2022.

[10] D. D. Chakladar, P. P. Roy, and V. Chang, "Integrated spatio-temporal deep clustering (istdc) for cognitive workload assessment," *Biomedical Signal Processing and Control*, vol. 89, p. 105703, 2024.

[11] D. D. Chakladar, "Vision transformer & brain connectivity patterns for estimating cognitive states," *IEEE Access*, 2025.

[12] D. D. Chakladar, A. Shankar, F. Liwicki, S. Barma, and R. Saini, "Attention dynamics: Estimating attention levels of adhd using swin transformer," in *International Conference on Pattern Recognition*. Springer, 2025, pp. 270–283.

[13] A. Omurtag, C. Sunderland, N. J. Mansfield, and Z. Zakeri, "Eeg connectivity and bdnf correlates of fast motor learning in laparoscopic surgery," *Scientific reports*, vol. 15, no. 1, p. 7399, 2025.

[14] B. Fazli, S. S. Sajadi, A. H. Jafari, E. Garosi, S. Hosseinzadeh, S. A. Zakerian, and K. Azam, "Eeg-based evaluation of mental workload in a simulated industrial human-robot interaction task," *Health Scope*, vol. 14, no. 14, 2025.

[15] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[16] K. Kedia, A. Bhardwaj, P. Dan, and S. Choudhury, "Interact: Transformer models for human intent prediction conditioned on robot actions," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 621–628.

[17] H. J. Choi, S. Das, S. Peng, R. Bajcsy, and N. Figueroa, "On the feasibility of eeg-based motor intention detection for real-time robot assistive control," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5592–5599.

[18] C. Xu, C. Zhang, Y. Zhou, Z. Wang, P. Lu, and B. He, "Trust recognition in human-robot cooperation using eeg," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 7827–7833.

[19] J. Yang, I. C. Layadi, J. P. Wachs, and D. Yu, "Adaptive human-robotic interaction for robotic-assisted surgical settings," *Military Medicine*, vol. 188, no. Suppl 6, p. 480, 2023.

[20] D. D. Chakladar and S. Chakraborty, "Eeg based emotion classification using "correlation based subset selection"," *Biologically inspired cognitive architectures*, vol. 24, pp. 98–106, 2018.

[21] Z. A. A. Alyasseri, O. A. Alomari, S. N. Makhadmeh, S. Mirjalili, M. A. Al-Betar, S. Abdullah, N. S. Ali, J. P. Papa, D. Rodrigues, and A. K. Abasi, "Eeg channel selection for person identification using binary grey wolf optimizer," *Ieee Access*, vol. 10, pp. 10 500–10 513, 2022.

[22] M. Arif, F. U. Rehman, L. Sekanina, and A. S. Malik, "A comprehensive survey of evolutionary algorithms and metaheuristics in brain eeg-based applications," *Journal of Neural Engineering*, 2024.

[23] D. Panda, D. D. Chakladar, S. Rana, and M. N. Shamsudin, "Spatial attention-enhanced eeg analysis for profiling consumer choices," *Ieee Access*, vol. 12, pp. 13 477–13 487, 2024.

[24] D. D. Chakladar and N. R. Pal, "Brain connectivity analysis for eeg-based face perception task," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 4, pp. 1494–1506, 2024.

[25] S. B. Shafiei, A. A. Hussein, and K. A. Guru, "Relationship between surgeon's brain functional network reconfiguration and performance level during robot-assisted surgery," *IEEE Access*, vol. 6, pp. 33 472–33 479, 2018.

[26] T. Zhou, J. S. Cha, G. Gonzalez, J. P. Wachs, C. P. Sundaram, and D. Yu, "Multimodal physiological signals for workload prediction in robot-assisted surgery," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 2, pp. 1–26, 2020.

[27] S. B. Shafiei, S. Shadpour, and J. L. Mohler, "An integrated electroencephalography and eye-tracking analysis using extreme gradient boosting for mental workload evaluation in surgery," *Human Factors*, vol. 67, no. 5, pp. 464–484, 2025.

[28] Q. Zhang, Y. Li, J. Zhan, and S. Chen, "Improved harris hawks algorithm and its application in feature selection." *Computers, Materials & Continua*, vol. 81, no. 1, 2024.

[29] S. B. Shafiei, S. Shadpour, J. Mohler, M. S. Toussi, P. Doherty, and Z. Jing, "Electroencephalogram and eye-gaze datasets for robot-assisted surgery performance evaluation," *Physionet https://doi.org/10.13026/qj5m*, no. 649, 2023.

[30] E. Battaglia, B. Mueller, D. Hogg, R. Rege, D. Scott, and A. M. Fey, "Evaluation of pre-training with the da vinci skills simulator on motor skill acquisition in a surgical robotics curriculum," *Journal of Medical Robotics Research*, vol. 6, no. 03n04, p. 2150006, 2021.

[31] W. Robison, S. K. Patel, A. Mehta, T. Senkowski, J. Allen, E. Shaw, and C. K. Senkowski, "Can fatigue affect acquisition of new surgical skills? a prospective trial of pre-and post-call general surgery residents using the da vinci surgical skills simulator," *Surgical endoscopy*, vol. 32, no. 3, pp. 1389–1396, 2018.

[32] A. K. Dubin, R. Smith, D. Julian, A. Tanaka, and P. Mattingly, "A comparison of robotic simulation performance on basic virtual reality skills: simulator subjective versus objective assessment tools," *Journal of minimally invasive gynecology*, vol. 24, no. 7, pp. 1184–1189, 2017.

[33] T. Alzahrani, R. Haddad, A. Alkhayal, J. Delisle, L. Drudi, W. Gotlieb, S. Fraser, S. Bergman, F. Bladou, S. Andonian *et al.*, "Validation of the da vinci surgical skill simulator across three surgical disciplines: A pilot study," *Canadian Urological Association Journal*, vol. 7, no. 7-8, p. E520, 2013.

[34] N. Raison, K. Ahmed, N. Fossati, N. Buffi, A. Mottrie, P. Dasgupta, and H. Van Der Poel, "Competency based training in robotic surgery: benchmark scores for virtual reality robotic simulation," *Bju international*, vol. 119, no. 5, pp. 804–811, 2017.

[35] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," *Future generation computer systems*, vol. 97, pp. 849–872, 2019.

[36] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "A critical analysis of self-supervision, or what we can learn from a single image," *arXiv preprint arXiv:1904.13132*, 2019.

[37] W. Duan, X. Chen, Y.-J. Wang, W. Zhao, H. Yuan, and X. Lei, "Reproducibility of power spectrum, functional connectivity and network construction in resting-state eeg," *Journal of Neuroscience Methods*, vol. 348, p. 108985, 2021.

[38] D. S. Oh, M. Ershad, J. O. Wee, M. S. Sancheti, D. M. D'Souza, L. J. Herrera, L. Y. Schumacher, M. Shields, K. Brown, S. Yousaf *et al.*, "Comparison of global evaluative assessment of robotic surgery with objective performance indicators for the assessment of skill during robotic-assisted thoracic surgery," *Surgery*, vol. 174, no. 6, pp. 1349–1355, 2023.