Transformer-based Automated Skill Assessment and Interpretation in Robot-Assisted Surgery

Yi Zheng[1] and Ann Majewicz-Fey[1,2]

***Abstract*— Different artificial intelligence approaches have been made to automatically assess skills during robotic surgical training. However, limitations still exist in these studies, including issues related to feature engineering, cross-validation methods, complex model architectures, and interpretability. In response to these limitations, this study introduces a Transformer-based model that processes kinematic data and identifies surgical skill levels. The model performance was rigorously evaluated under the Leave-One-User-Out cross-validation method, resulting in a classification accuracy of 80%. Beyond skill level classification, this study also explores deeper into the interpretability aspect. It includes the extraction of global-attention from the model, providing insights into the significance of each part or gesture within a task during the classification decision-making process. This interpretability holds the potential to help surgeon improve their skill by offering a comprehensive and detailed understanding of their performance.**

I. INTRODUCTION

Traditional surgical skill assessment methods primarily depend on the subjective observations of experienced surgeons, including intra-operative observation and post-operative video analysis, which are inherently less objective, unstructured, and can be influenced by personal biases. Besides, the manual review of surgical procedures by expert surgeons is time-consuming, and costly [1]–[3].

Numerous approaches have been made to mitigate the problem of reviewer bias in surgical skill assessment. One is using crowd-sourcing assessment [3], [4]. However, it still requires reviewers to watch surgical videos. Furthermore, these public reviewers typically lack medical training, which can lead to concerns about the reliability and validity of the assessment results.

Artificial intelligence (AI) techniques, including machine learning (ML) and deep learning (DL), have proven to be effective tools for surgical data processing and interpretation. Leveraging AI algorithms allows us to extract meaningful information from surgical procedures, enabling affordable, objective, more accurate, and consistent technical skill assessment [5].

The adoption of robotic surgical platforms has experienced rapid growth since its introduction. Besides enhanced dexterity and 3D vision systems, another advantage is the wealth of data sources provided by the robotic platforms. These platforms provide a variety of data sources for analysis, including high-resolution video feed and kinematic data from robotic sensors. The combination of these benefits and the exponential growth positions robotic platforms as an ideal environment for AI applications. As the adoption of robotic technology continues to expand, the opportunities for AI in this context will be continuously growing.

Numerous studies have implemented AI models for automated surgical skill assessment on robotic surgical platforms, including:

- Surgical videos to generate global rating scores [6], [7].
- Surgical videos to identify expertise levels [8]–[11].
- Kinematic data to generate global rating scores [12], [13].
- Kinematic data to identify expertise levels [12], [14]–[23].

While these studies have shown innovative approaches and promising results, several limitations still exist, such as cross-validation method selection, feature engineering, and complex deep learning (black-box) models.

In this study, we take a step towards addressing these challenges by working toward the development of a generalizable, computationally efficient, and interpretable surgical skill assessment model. We employ a Transformer-based model to process kinematic data and generate expertise level predictions and attention vectors for interpretation. Additionally, we utilize a more robust cross-validation method to enhance the reliability of model evaluation.

II. BACKGROUND

A. Expertise Level Assessment Using Kinematic Data

Robotic surgical platforms provide kinematic data from robotic manipulators. Analyzing kinematic data is more computationally efficient than processing video data, especially in real-time settings. Furthermore, kinematic data is potentially less invasive in terms of privacy compared to video data. Although video data could provide more information about the surgical context, the advantages of kinematic data make it a good choice for skill assessment.

Fard et al. extracted movement features from kinematic data to capture movement characteristics, including time to completion, path length, depth perception, speed, motion smoothness, turning angle, and more. They further utilized machine learning algorithms such as k-nearest neighbors, logistic regression, and support vector machines for the classification of surgical expertise levels [14]. While feature engineering based on raw kinematic data can be time-consuming and requires extensive calculation, one advantage

*This work was supported in part by NIH #1R01EB030125, and in part by NSF 2109635.

[1]Walker Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX 78712, USA

[2]Department of Surgery, UT Southwestern Medical Center, Dallas TX 75390, USA

Email: yi.zheng@austin.utexas.edu

of feature engineering is interpretability, for example, Zia and Essa took a similar approach by calculating four types of features from robotic kinematic data, including sequential motion texture, discrete Fourier transform, discrete cosine transform, and approximate entropy. After feature extraction, they used the nearest neighbor classifier for skill level classification and support vector regression models for rating score prediction. Notably, they also introduced the "impact scores" derived from the calculated features. The approach could indicate which parts of the input sequence had positive or negative effects on the final rating score prediction. By overlapping the impact scores with gestures, it could help surgeons understand the parts within a task that they need to improve on [12].

With the advancement of deep learning models, several studies have utilized CNN and its variations to process raw kinematic data in surgical skill assessment, including the combination with GRU and LSTM [15], [18], [19], [21], [22]. These approaches uncovered the underlying patterns within data sequences, potentially enhancing skill level classification accuracy. However, despite the advantages, deep learning models often present challenges in terms of interpretability due to their black-box nature. Similar to Zia and Essa's work, Fawaz et al. employed CNN in combination with the class activation map (CAM) to localize which parts of the trial impacted the model's decision when evaluating the skill level [19]. However, it's noteworthy that this study, along with the other deep learning-based studies mentioned earlier in this section, adopted Leave-One-Supertrial-Out (LOSO) cross-validation to evaluate the model performance on the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [24], [25]. Under LOSO cross-validation, each iteration involves leaving one of the five trials performed by one subject as the testing set, while using the remaining four trials from the same subject, combined with trials from other subjects, as the training set. However, this approach can introduce data leakage since the model "sees" data from a specific subject in both the training and testing process, potentially leading to an overestimated model performance. In the context of Fawaz's study, under LOSO cross-validation, besides an overestimated model performance, the highlighted parts generated by CAM within a trial might not be generalized enough to newly coming subjects.

Therefore, the discussion above demonstrates a research opportunity in the development of more robust, computationally efficient, and interpretable models for skill level classification in surgical skill assessment.

B. Transformer Models

Given that identifying skill levels in surgical skill assessment can be formulated as a sequence classification problem, our attention is drawn to an emerging and popular model - the Transformer model. Although the initial idea of Transformer was to improve the Natural Language Processing (NLP) tasks, such as machine translation, as proposed by Vaswani et al., it gained a wide range of applications in different fields, to name a few, image processing [26], speech recognition [27], and more. In sequence classification tasks, Yan et al. utilized the Transformer model to handle arrhythmia heartbeat classification [28]. Nambiar et al. adapted the Transformer model for protein family classification and protein interaction prediction [29], and Sun et al. constructed multiple Transformer-based models for motor imaginary EEG classification [30]. These diverse applications underscore Transformer's ability in sequence classification tasks across various domains, due to the key factor of the attention mechanism that could capture long-range dependencies in the sequence more effectively.

C. Transformer Models in Robotic Surgery

Transformer models have also found compelling applications in the domain of robotic surgery. Kiyasseh et al. proposed a unified surgical AI system (SAIS) based on the vision Transformer architecture that could decode elements of intra-operative surgical activities from videos collected during surgery at three different hospitals, such as surgical steps, surgical gestures, and quality during surgery. During skill assessment on surgical gestures, SAIS could also place attention on each video frame. By inspecting the attention, they were able to quantify the gesture relevance to the skill being assessed [31]. Anastasiou et al. introduced a novel video-based, contrastive regression architecture, Contra-Sformer. The model could capture the differences in the surgical performance between a test video and a reference video by calculating the similarity and deviation between test and reference videos, and generate the rating score for the test video [6]. Shi et al. utilized a Transformer-based model to recognize and predict surgical activities, such as surgical gestures and end-effector trajectories [32]. Furthermore, building upon their research, Shi et al. used the predicted future trajectory by the Transformer-based model for assistive and resistive haptic cues during robotic surgical training tasks on a da Vinci Research Kit. They observed task performance improvement and a large decrease in user difficulty ratings in washout trials that followed the resistive conditions, indicating the resistive haptic cues during surgical training could potentially result in lasting after-effects on performance once the cues are removed [33].

Therefore, as inspired by these studies and their limitations, we are taking a progressive step in our research. Our aim is to employ a more robust cross-validation method to assess the performance of our proposed Transformer-based model for processing kinematic data and identify expertise levels. Additionally, we seek to extract and analyze the specific movements within the kinematic data that have more impact on the model's decision regarding skill level classification. In other words, we aim to find the movements that best characterize distinct skill levels, thus advancing the interpretability of skill assessment.

III. METHODS

A. Dataset

To train and validate our proposed model, we used the JHU-ISI Gesture and Skill Assessment Working Set (JIG-

TABLE I: Gesture Descriptions in JIGSAWS Suturing Task

Gesture ID	Description
G0	Unannotated
G1	Reaching for needle with right hand
G2	Positioning needle
G3	Pushing needle through tissue
G4	Transferring needle from left to right
G5	Moving to center with needle in grip
G6	Pulling suture with left hand
G8	Orienting needle
G9	Using right hand to help tighten suture
G10	Loosening more suture
G11	Dropping suture at end and moving to endpoints

SAWS) in the study [25]. JIGSAWS is a well-known dataset in the field of robotic surgical skill assessment. It includes three common surgical training tasks: Suturing, Needle Passing, and Knot Tying. Eight subjects with different robotic surgical experiences were recruited. The expertise level of each subject was determined by self-reported experience - Expert (EX): more than 100 hours, Intermediates (IN): between 10 and 100 hours, Novices (NO): less than 10 hours. The subjects repeated each task five times using a da Vinci Surgical System (dVSS), and kinematic data from the dVSS was captured.

Kinematic data of both left and right master tool manipulators (MTM), and the first and second patient side manipulators (PSM) were recorded. It includes 19 variables for each manipulator - Cartesian Positions (3): xyz, Rotation Matrix (9): R, Linear Velocities (3): $x'y'z'$, Angular Velocities (3): $\alpha'\beta'\gamma'$, and a gripper angle (1): θ.

Besides the kinematic data, another feature of the JIGSAWS is the manually annotated surgical gestures which are synchronized with the kinematic data. The dataset specified 15 gestures for all three tasks, as shown in Table. I.

For model training and evaluation, we only used 39 Suturing trials and 19 kinematic data of both MTMs (38 in total). The Suturing task includes 10 gestures out of a total of 15 gestures in JIGSAWS. For each trial, we labeled the unannotated time step as "0".

B. Transformer Model

We formulate the task in this study as a time-sequence classification task. Our proposed model is a derivative of the original Transformer architecture as described in [34]. The original Transformer model was designed for sequence-to-sequence tasks, therefore, we have adapted and refined the Transformer model to suit our classification objectives. For example, we removed the parts of Embedding layers and Decoder Input layer from the original Transformer model, and we added an Average layer to map the encoder output dimension to decoder dimension, in order to calculate the attention between encoder sequence and decoder (final) output. In this study, the task is sequence classification, where the input comprises kinematic data from a trial in JIGSAWS and the output is a classification of the subject's expertise level - NO, IN, and EX (Fig.1).

The input data initially undergoes Positional Encoding.

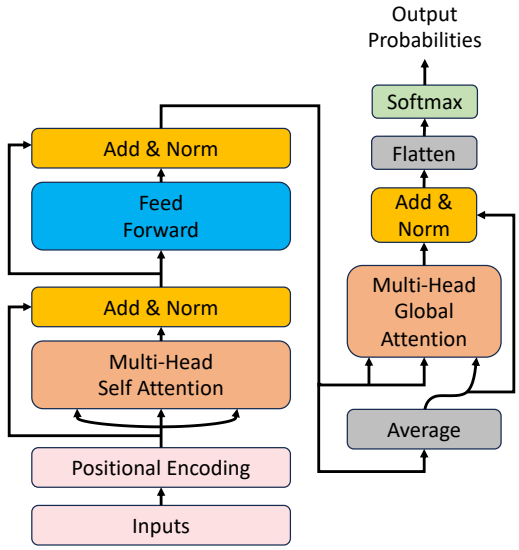

Fig. 1: Architecture of the proposed Transformer model. The diagram illustrates the model's structure, where the input consists of the kinematic data with dimension [max_length, 38]. The Multi-Head Attention on the left calculates the self-attention within the input sequence. The Multi-Head Attention on the right calculates the global-attention between the final output and processed sequence. The model processes the input to yield output probabilities for 3 classes.

Unlike traditional sequential models, the Transformer architecture does not inherently process sequential information. Positional Encoding provides the model with necessary sequence order information.

Subsequently, the encoded data is fed into a Multi-head Attention mechanism. For model simplicity, our implementation utilizes a single head. This layer captures self-attention within the input sequence, enabling the model to discern dependencies between various time steps and providing more information on data representation for final classification.

Following this, the data progresses through a Feedforward layer, which further processes the information.

To understand and interpret the model's decision-making process during classification, the data is then processed through another Multi-head Attention layer. This layer is distinct from the self-attention mechanism, as it focuses on global attention. The global attention could describe the importance of each time step for final classification.

Finally, the data is directed through a fully-connected layer to make the final classification.

C. Model Training and Testing

To evaluate the model performance, we used the data of Suturing in JIGSAWS and adopted Leave-One-User-Out (LOUO) cross-validation. LOUO ensures that the model is tested on unseen users, providing a robust estimate of how well the model generalizes to new users, and closely simulates real-world scenarios where models encounter data from new users not present in the training set. More specifi-

cally, for each iteration of the validation process, the dataset comprising all five trials from the i^{th} subject was left as the testing set, while the datasets from the remaining subjects formed the training set. The procedure was repeated for each of the eight subjects, therefore, ensuring comprehensive assessment and validation across all individual users.

Prior to the training and testing, the data underwent standardization using a StandardScaler. To ensure consistent scaling, the StandardScaler was fitted exclusively to the training set for each iteration. Subsequently, the scaling transformation was applied to both the training and testing sets, maintaining data integrity by preventing data leakage from the training set.

After obtaining the standardized data, we addressed the variance in trial lengths by implementing zero-padding. The process was essential to equalize the length of each trial to the same dimension, max_length. The value of max_length was determined as the length of the longest trial in terms of time steps. For trials shorter than max_length, zeros were appended to the end of the data sequence, ensuring that all trials conformed to this standardized length.

D. Model Evaluation

After training and obtaining testing results, we used standard metrics to evaluate our model performance, Precision, Recall, and F1-score, as well as the Macro- and Micro Averages.

Precision is the ratio of correctly predicted positive (True Positive, TP) labels to the total predicted positive labels (True Positive and False Positive, TP + FP). High precision indicates a low rate of false positives for that class.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall is the ratio of correctly predicted positive (True Positive, TP) labels to the total number of predictions in actual labels (True Positive and False Negative, TP + FN). High recall indicates that the model is good at capturing positives for that class.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F1-score is the weighted average of Precision and Recall. It takes both false positives and false negatives into account. A high F1-score indicates a good balance between precision and recall for that class.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

In multi-class tasks, a common practice is to calculate these metrics for each class and average them - Macro and Micro Averaging. Macro Averaging treats all classes equally, computes the metric independently for each class, and then takes the average. Micro Averaging aggregates the contributions of all classes to compute the average metric by aggregating the total true positives, false positives, and false negatives. Micro Averaging is more reflective of the model's performance on the data as a whole, especially in the presence of class imbalance.

IV. RESULTS

A. Expertise Level Classification

As discussed in earlier sections, we employed LOUO cross-validation to train and evaluate our model. The dataset consisted of kinematic data from the MTMs (38 features in total), collected during Suturing task from the JIGSAWS. The dataset includes eight subjects, labeled into one of three expertise levels: NO, IN, and EX.

In each iteration of LOUO cross-validation, all five trials from the i^{th} subject were left as the testing set, while data from the remaining subjects were used to train the model. Our model takes zero-padded kinematic data of each entire trial as input and outputs the corresponding expertise level.

To assess the model performance, we calculated metrics including precision, recall, and F1-score, and their Macro and Micro averages to provide a comprehensive model evaluation.

Table. II summarizes the LOUO cross-validation results, listing the classification accuracy for each subject when used as the testing set. The model demonstrated high accuracy in classifying NO and EX subjects, correctly classifying all trials for these expertise levels. However, its performance in classifying IN trials was deficient. For subject C, only 2 trials were classified correctly, while for subject F, all trials were misclassified.

To gain deeper insights into the model performance, we constructed a confusion matrix, which is presented in Fig. 2. It becomes clear that the model is accurate in identifying NO and EX levels. However, it encountered difficulties when classifying IN. As shown in Table. III, the Macro averages, treating all classes equally, yield reasonably good results (Precision: 0.85, Recall: 0.73, F1-Score: 0.68), although these metrics are adversely affected by the misclassifications of IN as EX.

On the other hand, the Micro averages also exhibit strong performance (Precision: 0.79, Recall: 0.79, F1-Score: 0.79), suggesting an overall good model performance. However, it is worth noting that the performance in NO and EX classes may be masking the model's weakness in identifying IN. The misclassifications of IN as EX presents a key area for potential model improvement.

B. Attention Interpretations

In addition to the final classifications, we also generated a global attention vector using the model. Global attention indicates the importance of each time step in influencing the final classification results. Given the synchronized gesture sequence with the kinematic data in JIGSAWS (Table. I), for each trial, we overlayed the global attention vector onto the corresponding gesture sequence. This approach allowed us to gain insights into the relative importance of different gestures in generating the final classifications.

As our model demonstrated a good performance in discriminating between NO and EX trials, we focused our analysis solely on this subset of the data. For each trial within NO and EX, we computed the attention devoted to

TABLE II: Summary of classification results using LOUO cross-validation and Suturing task.

Testing Subject	Sub_B	Sub_C	Sub_D	Sub_E	Sub_F	Sub_G	Sub_H	Sub_I
Expertise Levels	NO	IN	EX	EX	IN	NO	NO	NO
Accuracy	100%	40%	100%	100%	0%	100%	100%	100%

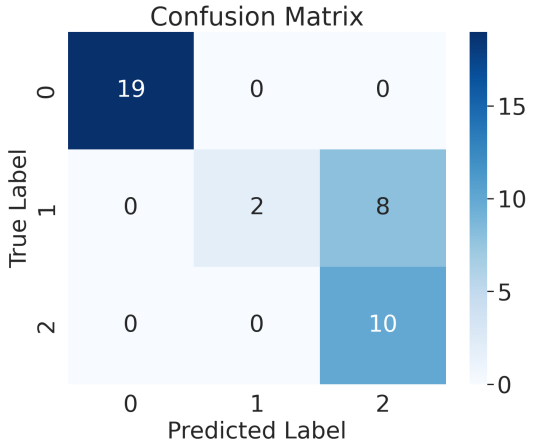

Fig. 2: Confusion matrix of the final classification results indicating the number of instances in each class. Class 0: Novices, Class 1: Intermediates, Class 2: Experts.

TABLE III: Summary of model evaluation metrics.

Metrics		
Overall Accuracy		80%
Macro	Precision	0.85
	Recall	0.73
	F1-Score	0.68
Micro	Precision	0.79
	Recall	0.79
	F1-Score	0.79

each gesture, or gesture attention, by averaging the attention scores across individual gestures.

First, we conducted a comparison of gesture attention across all NO and EX trials, as illustrated in Fig. 3 and Table. IV, G1 received the highest attention and it is significantly higher than G4, G6, G8, G11, and G9. In contrast, G6 received the lowest attention and it is significantly lower than the attentions received by G0, G1 G2, G3, G5, and G11. In summary, when the model predicts the expertise level of a given trial, it will focus more on G0, G1, G2, G3, G5, and G11 than the other gestures.

TABLE IV: Comparing Gestures Attentions of All Subjects. Based on the data distribution property, we used the Kruskall-Wallis test for statistical analysis.

Significance		
G0	>	G4 (p = 0.0002), G6 (p = 0.0000)
G1	>	G4 (p = 0.0000), G6 (p = 0.0000), G8 (p = 0.0157), G11 (p = 0.0135), G9 (p = 0.0061)
G2	>	G4 (p = 0.0002), G6 (p = 0.0000)
G3	>	G6 (p = 0.0041)
G5	>	G4 (p = 0.0001), G6 (p = 0.0000)
G11	>	G6 (p = 0.0370)

TABLE V: Comparing Gestures Attentions of EX Subjects. Based on the data distribution property, we used the Kruskall-Wallis test for statistical analysis.

Significance		
G1	>	G6 (p = 0.0027), G9 (p = 0.0479), G11 (p = 0.0175)
G2	>	G6 (p = 0.0146)
G5	>	G6 (p = 0.0001), G9 (p = 0.0052), G11 (p = 0.0011)

TABLE VI: Comparing Gestures Attentions of NO subjects. Based on the data distribution property, we used the Kruskall-Wallis test for statistical analysis.

Significance		
G0	>	G4 (p = 0.0000), G6 (p = 0.0001)
G1	>	G4 (p = 0.0000), G6 (p = 0.0000), G8 (p = 0.0349)
G2	>	G4 (p = 0.0000), G6 (p = 0.0005)
G3	>	G4 (p = 0.0152)
G5	>	G4 (p = 0.0003), G6 (p = 0.0054)
G11	>	G4 (p = 0.0002), G6 (p = 0.0038)

Next, we proceeded to compare gesture attentions specifically among EX trials. As depicted in Fig. 4 (Red), G5 received the highest attention, whereas G11 received the lowest attention. Among the gestures, G1 and G5 received significantly higher attentions than G6, G9, and G11 (Table. V). In summary, when the model predicts a certain trial as EX, it will focus more on G1, G2, and G5 than the other gestures. In other words, the user's performance of G1, G2, and G5 has more importance in characterizing an EX trial.

Similarly, we then compared the gesture attention across NO subjects. As shown in Fig. 4 (Blue), G1 received the highest attention, whereas G4 received the lowest attention and it received significantly lower attention than G0, G1, G2, G3, G5, and G11 (Table VI). In summary, when the model predicts a certain trial as NO, it will focus more on G0, G1, G2, G3, G5, and G11 than the other gestures. In other words, the user's performance of G0, G1, G2, G3, G5, and G11 has more importance in characterizing a NO trial.

To comprehend the distinct roles of different gestures in characterizing NO and EX trials, we conducted a comparison

TABLE VII: Comparing Each Gesture Attention between EX and NO. *: t-test. **: Mann-Whitney U test.

Gesture	Significance	p-value
G0	na*	0.7225
G1	na*	0.8903
G2	**EX >NO***	**0.0496**
G3	na*	0.2253
G4	**EX >NO***	**0.0000**
G5	**EX >NO****	**0.0026**
G6	na*	0.5037
G8	na**	0.1265
G9	na*	0.2463
G11	**NO >EX***	**0.0138**

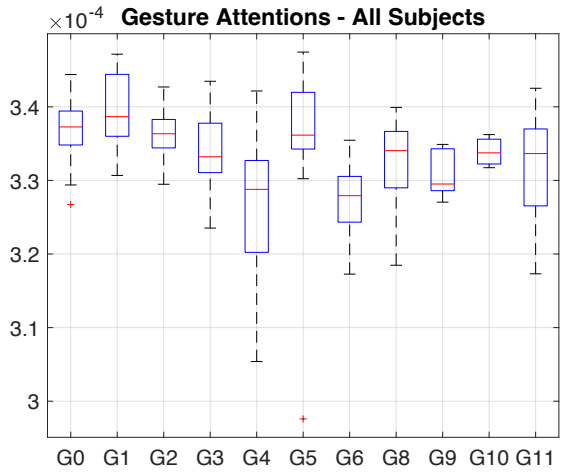

Fig. 3: Comparing the gesture attentions across all subjects.

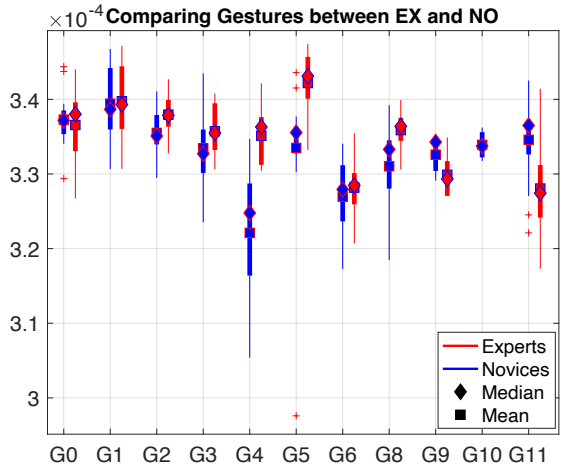

Fig. 4: Comparing the gesture attention between EX and NO.

of the attention each gesture received in these two categories. In Fig. 4 and Table. VII, G2, G4, G5 received significantly higher attention when characterizing EX trials compared to NO trials, while G11 received significantly higher attention when characterizing NO trials compared to EX trials.

V. DISCUSSION

According to a review paper by Yanik et al., LOSO is the most common cross-validation method used for surgical skill assessment. The models in the literature could achieve very high accuracies up to 98.7%. The limitation of LOSO lies in the model's exposure to data from all subjects, including the one under evaluation. The main reason that researchers tend to use LOSO over LOUO is the data imbalance within the JIGSAWS dataset, hurting the generalization of the models. However, the authors still believe that more work should be done to improve the model performance under LOUO to fill the gap in the literature, recognizing it as a more robust method for evaluating the model's generalizability [35].

Therefore, in this study, we used LOUO cross-validation to evaluate our proposed model. Based on our results, we observed that the model tends to confuse IN with EX (Fig. 2). Similar findings have been observed in related studies, for example, Funke et al. used a 3D CNN for skill level classification and the model misclassified all IN trials to either EX or NO in Knot-Tying using LOUO cross-validation [9]. Yanik et al. also mentioned that IN trials are the most difficult to classify, and the reason includes the categorization methods in JIGSAWS which is based on self-reported hours of experience, and the unbalanced dataset may bias models toward NO data (4 NO, 2 IN, and 2 EX). However, in contrast, our results suggested that 2 IN trials were correctly classified and 8 IN trials were misclassified as EX. This suggests that our model, while still struggling with IN classification, may offer an improved ability to handle the bias associated with dataset imbalance. Therefore, our future work will improve the classification accuracy of IN and finding the key differentiators between IN and EX/NO. This could involve collecting a larger dataset that contains more diverse subject population, incorporating multiple data sources - kinematic data and video from the robot, physiological sensor data to precisely measure the human operator's performance (e.g., EMG, GSR).

Additionally, we also examined the global-attention exerted by the proposed model when classification results are being made. Unlike self-attention, global-attention provides insights into the significance of each time step when the model is making a classification decision. By overlapping the attention onto the gesture sequence, we could identify specific gestures that could better characterize a NO or an EX performance. For example, G1, G2, and G5 received significantly higher attention in both EX and NO trials. This indicates that these gestures play important roles in characterizing skill levels, serving as focus points for surgeons. When comparing gesture attention between EX and NO, as shown in Table. VII, G2, G4, and G5 can better characterize an EX performance. In contrast, G11 emerged as a key identifier for a NO performance. This approach offers an opportunity for the interpretation of skill level classification. Surgeons being assessed can gain a deeper understanding of which parts of the performance require improvement. Such interpretability can serve as a tool for skill enhancement during surgical training.

VI. CONCLUSIONS

This study addresses the current limitations of automated skill assessment within robotic surgical training. The proposed Transformer-based model achieved a classification accuracy of 80% under LOUO cross-validation. The global-attention exerted from the model could improve the interpretability of classification results. The identification of key gestures and their significance in characterizing skill levels offer insights for surgeons to refine their performance. This work contributes to the ongoing efforts to enhance surgical training through AI techniques. Future research may further refine the proposed approach by collecting an extensive real-world dataset for model training.

REFERENCES

[1] K. Lam, J. Chen, Z. Wang, F. M. Iqbal, A. Darzi, B. Lo, S. Purkayastha, and J. M. Kinross, "Machine learning for technical skill assessment in surgery: a systematic review," *NPJ digital medicine*, vol. 5, no. 1, p. 24, 2022.

[2] R. Aggarwal, T. Grantcharov, K. Moorthy, T. Milland, and A. Darzi, "Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room," *Annals of surgery*, vol. 247, no. 2, pp. 372–379, 2008.

[3] T. S. Lendvay, L. White, and T. Kowalewski, "Crowdsourcing to assess surgical skill," *JAMA surgery*, vol. 150, no. 11, pp. 1086–1087, 2015.

[4] L. W. White, T. M. Kowalewski, R. L. Dockter, B. Comstock, B. Hannaford, and T. S. Lendvay, "Crowd-sourced assessment of technical skill: a valid method for discriminating basic robotic surgery skills," *Journal of endourology*, vol. 29, no. 11, pp. 1295–1301, 2015.

[5] R. Pedrett, P. Mascagni, G. Beldi, N. Padoy, and J. L. Lavanchy, "Technical skill assessment in minimally invasive surgery using artificial intelligence: a systematic review," *Surgical endoscopy*, vol. 37, no. 10, pp. 7412–7424, 2023.

[6] D. Anastasiou, Y. Jin, D. Stoyanov, and E. Mazomenos, "Keep your eye on the best: Contrastive regression transformer for skill assessment in robotic surgery," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1755–1762, 2023.

[7] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li, "Towards unified surgical skill assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9522–9531.

[8] M. Pan, S. Wang, J. Li, J. Li, X. Yang, and K. Liang, "An automated skill assessment framework based on visual motion signals and a deep neural network in robot-assisted minimally invasive surgery," *Sensors*, vol. 23, no. 9, p. 4496, 2023.

[9] I. Funke, S. T. Mees, J. Weitz, and S. Speidel, "Video-based surgical skill assessment using 3d convolutional neural networks," *International journal of computer assisted radiology and surgery*, vol. 14, pp. 1217–1225, 2019.

[10] S. Khalid, M. Goldenberg, T. Grantcharov, B. Taati, and F. Rudzicz, "Evaluation of deep learning models for identifying surgical actions and measuring performance," *JAMA network open*, vol. 3, no. 3, pp. e201 664–e201 664, 2020.

[11] A. Soleymani, A. A. S. Asl, M. Yeganejou, S. Dick, M. Tavakoli, and X. Li, "Surgical skill evaluation from robot-assisted surgery recordings," in *2021 International Symposium on Medical Robotics (ISMR)*. IEEE, 2021, pp. 1–6.

[12] A. Zia and I. Essa, "Automated surgical skill assessment in rmis training," *International journal of computer assisted radiology and surgery*, vol. 13, pp. 731–739, 2018.

[13] M. Benmansour, A. Malti, and P. Jannin, "Deep neural network architecture for automated soft surgical skills evaluation using objective structured assessment of technical skills criteria," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–9, 2023.

[14] 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.

[15] 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, pp. 1959–1970, 2018.

[16] G. Lajkó, R. Nagyne Elek, and T. Haidegger, "Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery," *Sensors*, vol. 21, no. 16, p. 5412, 2021.

[17] B. B. Oğul, M. Gilgien, and S. Özdemir, "Ranking surgical skills using an attention-enhanced siamese network with piecewise aggregated kinematic data," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 6, pp. 1039–1048, 2022.

[18] D. Castro, D. Pereira, C. Zanchettin, D. Macêdo, and B. L. Bezerra, "Towards optimizing convolutional neural networks for robotic surgery skill evaluation," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[19] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks," *Interna-*

tional journal of computer assisted radiology and surgery, vol. 14, pp. 1611–1617, 2019.

[20] L. Juarez-Villalobos, N. Hevia-Montiel, and J. Pérez-Gonzalez, "Machine learning based classification of local robotic surgical skills in a training tasks set," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 4596–4599.

[21] X. A. Nguyen, D. Ljuhar, M. Pacilli, R. M. Nataraja, and S. Chauhan, "Surgical skill levels: Classification and analysis using deep neural network model and motion signals," *Computer methods and programs in biomedicine*, vol. 177, pp. 1–8, 2019.

[22] Z. Wang and A. M. Fey, "Satr-dl: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 1793–1796.

[23] G. Forestier, F. Petitjean, P. Senin, F. Despinoy, A. Huaulmé, H. I. Fawaz, J. Weber, L. Idoumghar, P.-A. Muller, and P. Jannin, "Surgical motion analysis using discriminative interpretable patterns," *Artificial intelligence in medicine*, vol. 91, pp. 3–11, 2018.

[24] N. Ahmidi, L. Tao, S. Sefati, Y. Gao, C. Lea, B. B. Haro, L. Zappella, S. Khudanpur, R. Vidal, and G. D. Hager, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2025–2041, 2017.

[25] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI workshop: M2cai*, vol. 3, no. 3, 2014.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[27] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[28] G. Yan, S. Liang, Y. Zhang, and F. Liu, "Fusing transformer model with temporal features for ecg heartbeat classification," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 898–905.

[29] A. Nambiar, M. Heflin, S. Liu, S. Maslov, M. Hopkins, and A. Ritz, "Transforming the language of life: transformer neural networks for protein prediction tasks," in *Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics*, 2020, pp. 1–8.

[30] J. Sun, J. Xie, and H. Zhou, "Eeg classification with transformer-based models," in *2021 ieee 3rd global conference on life sciences and technologies (lifetech)*. IEEE, 2021, pp. 92–93.

[31] D. Kiyasseh, R. Ma, T. F. Haque, B. J. Miles, C. Wagner, D. A. Donoho, A. Anandkumar, and A. J. Hung, "A vision transformer for decoding surgeon activity from surgical videos," *Nature Biomedical Engineering*, pp. 1–17, 2023.

[32] C. Shi, Y. Zheng, and A. M. Fey, "Recognition and prediction of surgical gestures and trajectories using transformer models in robot-assisted surgery," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8017–8024.

[33] C. Shi, J. Madera, H. Boyea, and A. M. Fey, "Haptic guidance using a transformer-based surgeon-side trajectory prediction algorithm for robot-assisted surgical training," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2023, pp. 1942–1949.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] E. Yanik, X. Intes, U. Kruger, P. Yan, D. Diller, B. Van Voorst, B. Makled, J. Norfleet, and S. De, "Deep neural networks for the assessment of surgical skills: A systematic review," *The Journal of Defense Modeling and Simulation*, vol. 19, no. 2, pp. 159–171, 2022.