2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Surgical skill analysis using explainable AI in endoscopic sinus surgery

Kaito Yamada[1], Masanobu Suzuki[2], Kou Miyaji[1], Koki Ebina[1], Kazuya Sase[3], Teppei Tsujita[4], Xiaoshuai Chen[5], Takashige Abe[6], Shunsuke Komizunai[7], Yuji Nakamaru[2], Taku Senoo[1], Akihiro Homma[2] and Atsushi Konno[1]

***Abstract*— Endoscopic sinus surgery (ESS) is a standard procedure performed worldwide for diseases of the nose and sinuses, but it is highly technical and requires effective training. A system for evaluating the skill level of surgeons using machine learning based on measurement data from endoscopic sinus surgeries performed with a 3D sinus model has been developed. In this study, an analysis using SHapley Additive exPlanations (SHAP) values was conducted to investigate behaviors that significantly reflect surgeons' skill proficiency. Additionally, feature reduction was performed using a variable reduction method, eliminating features in ascending order of their contributions as calculated by SHAP. This approach aimed not only to mitigate the risk of overfitting due to a decrease in explanatory variables but also to improve the accuracy of classifying surgeons based on skill differences. The results demonstrated that reducing features based on SHAP contributions led to an improvement in classification accuracy.**

I. INTRODUCTION

Endoscopic sinus surgery (ESS) is a type of endoscopic surgery in which surgical instruments are inserted through the nostrils. ESS is the first choice for the treatment of inflammatory and neoplastic diseases of the paranasal sinuses because it is less demanding on the patient than the traditional canine fossa approach. A high level of skill is required for ESS, as the procedure relies on two-dimensional information from the endoscope and requires the surgeon to operate the forceps and endoscope with each hand. Furthermore, the paranasal sinuses are adjacent to critical structures such as the orbit and brain parenchyma, meaning that inadequate surgical skills could lead to severe complications, including blindness or brain injury. Thus, efficient and reliable methods for skill acquisition are essential.

*This work was supported by JSPS Grants-in-Aid for Scientific Research (A)(JP23H00480), Grant-in-Aid for Challenging Research (Exploratory)(JP23K18486), Fund for the Promotion of Joint International Research (JP18KK0444) and Grant-in-Aid for Early-Career Scientists (JP22K16923).

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan. {yamada@scc, konno@ssi}.ist.hokudai.ac.jp

[2]Department of Otolaryngology-Head and Neck Surgery, Faculty of Medicine and Graduate School of Medicine, Hokkaido University, Sapporo, Japan.

[3]Department of Mechanical Engineering and Intelligent Systems, Tohoku Gakuin University, Sendai, Japan

[4]Department of Mechanical Engineering, National Defense Academy of Japan, Yokosuka, Japan

[5]Graduate School of Science and Technology, Hirosaki University, Hirosaki, Japan

[6]Department of Urology, Hokkaido University Graduate School of Medicine, Sapporo, Japan

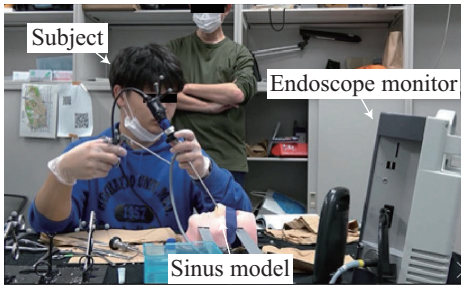
[7]Faculty of Engineering and Design, Kagawa University, Takamatsu, Japan

Current methods of skill acquisition include On the Job Training (OJT) and Off the Job Training (Off-JT). OJT is considered the most practical form of training, in which surgery is performed on a real human body under the supervision of an experienced surgeon [1]. However, there are patient safety issues with OJT, such as the risk of complications, as well as concerns about the burden on the supervising surgeon. Off-JT is training on simple models or cadavers. The issues are that simple models are too simple to be applied clinically and that cadaver training has few opportunities. In addition, restrictions on working hours due to work style reforms limit the time available for experienced surgeons to teach and conduct training. Therefore, there is a need for efficient methods of skill transfer and the development of systems that can provide appropriate evaluations even in training conducted solely by beginners.

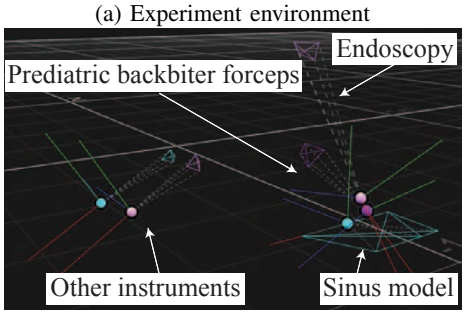
There has been several research on efficient skill acquisition in ESS. Majority of them focused on the analysis and evaluation of surgical instruments' movements. Sugino et al. conducted a quantitative analysis of endoscope manipulation, revealing differences in endoscope handling among the different skill level [2]. However, their study was limited to the analysis of endoscope manipulation alone, and different anatomies among participants posed challenges. Similarly, Ahmidi et al. developed a system to identify surgical skill levels using gaze information and instrument tracking data during surgery [3]. However, their study included a task that required touching a designated location, which differs from the movements observed in clinical practice. To address these issues, we standardized measurement conditions by using a 3D sinus model and employed a system capable of simultaneously measuring multiple surgical instruments for skill evaluation [4]. However, this evaluation was based on manually defined features (e.g., velocity, acceleration, jerk), which limited the range of extractable features. Therefore, a quantitative evaluation of skills was performed by automatically generating a large number of features using tsfresh [5], a feature generation library based on Python [6].

In this study, class classification was performed using the features calculated by tsfresh to reveal differences in surgical technique due to previously unidentified skill differences. Subsequently, analysis was conducted using SHapley Additive exPlanations (SHAP), a type of explainable AI, to identify the features deemed important by the skill assessment model [7]. Additionally, feature reduction based on the calculated contributions was performed to prevent overfitting and improve classification accuracy.

979-8-3315-3161-4/25/$31.00 ©2025 IEEE



(a) Experiment environment

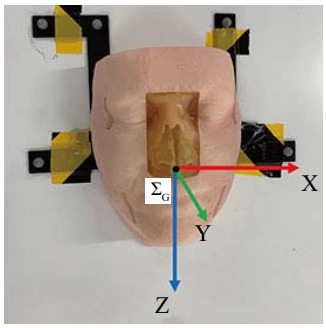


(b) Measured surgical instruments movement

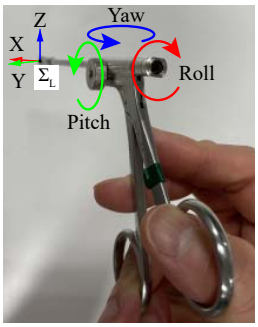
Fig. 1: Overview of the experiment

TABLE I: Subjects grouping

Group	ESS-OSATS score	Number of subjects
Expert	65 – 90 (highest)	13
Intermediate	54 – 64	9
Novice	18 (lowest) – 53	14



(a) World coordinate frame



(b) Local coordinate frame

Fig. 2: Coordinate frame

II. Procedure Measurement Experiment

A. Measurement System

A technique measurement experiment was conducted to quantify the surgeon's movements during surgery (Fig. 1(a)). In this study, a multi-instrument measurement system developed by the authors, which can measure multiple surgical instruments simultaneously, was used [4]. An optical motion capture system was employed to measure the procedure, and 11 OptiTrack Prime41 cameras were used. Markers were attached in different configurations to four types of surgical instruments (An endoscope, Pediatric backbiter forceps, Straight-cutting forceps, Upturned-cutting forceps) to allow simultaneous measurement of the movements of each instrument. Using the Mocap system control software (OptiTrack Motive), the marker sets were registered as rigid bodies and the tip position and orientation of each surgical instrument was determined. When the offset was taken into account, the measurement error of the tip position of the surgical instruments was less than 1 (mm). The surgical instruments measured simultaneously by Mocap are shown in Fig. 1(b). The measurement experiment used a 3D sinus model produced by FuseTec, which replicates the elasticity and texture of skin and cartilage, enabling surgeries under conditions close to clinical reality [8]. The targeted procedure was a Full-house ESS [9], aimed at opening the paranasal sinuses, with a time limit of 45 minutes. The subjects for the procedure measurement experiment were 47 otolaryngologists. Data from 36 of the 47 subjects who used the endoscope and all three types of forceps without significant data loss were used in the skill analysis of this study.

Scoring was performed based on the ESS-specific evaluation metric (ESS-OSATS) [10] based on endoscopic records. The scoring was on a 5-point scale from 1 to 5 for 18 categories, excluding anaesthesia-related items, with a maximum score of 90. The scoring was conducted by one board-certified otolaryngologist from the Japan Society of Otorhinolaryngology-Head and Neck Surgery. For 24 randomly selected cases, an additional otolaryngologist also conducted the scoring, and the intraclass correlation coefficient (ICC) was calculated. The correlation between the two raters was $r = 0.966$, demonstrating high inter-rater reliability. Based on the scores, participants were classified into three groups: Expert, Intermediate, and Novice. The threshold scores for classification were set based on discussions with the board-certified otolaryngologist, and the composition of the subjects is shown in Table I.

B. Feature Calculation

In previous ESS skill analyses, features defined based on the analyst's hypothesis were used to characterize the movement of the instruments. However, the number of features that can be hypothesized by humans is limited, and it was possible that unintentionally useful features were overlooked. This study used tsfresh [5], a Python library that automatically calculates features for time-series data. Because tsfresh can calculate countless features by changing the parameters of 76 predefined features, it can calculate various features without being bound by preconceived notions. The measurement data of each surgical instrument obtained by the measurement system consist of time-series data of the instrument's tip position in the world coordinate system (px, py, pz) and rotation angles in the local coordinate system (rx, ry, rz), as shown in Fig. 2. Twenty-four types of data (6 time series $\times$ 4 types of surgical instruments) were input into tsfresh along with timestamps at 100 Hz. However, since some of the calculated features are not suitable for this study due to their definitions, preprocessing was performed. In this study, since surgical instruments are exchanged during surgery, there are times when certain instruments are not in use, resulting in discontinuous measurement data. Therefore,

TABLE II: Summary of tsfresh features selected in preprocessing

Instruments	Features (18 types)
En (Endoscope)	rx_LoM (Location of minimum), py_RO1.5σ (Ratio of values beyond 1.5 standard deviations), pz_0.1Q (Quantile), pz_V (Variance)
Ba (Pediatric backbiter forceps)	ry_LSBM (Length of consecutive subsequence below mean)
St (Straight-cutting forceps)	rx_RMS (Root mean square), rz_PSD (Cross power spectral density), rz_V (Variance), px_K (Kurtosis), py_VC (Variation coefficient), py_0.1Q (Quantile), py_V (Variance), pz_PSD (Cross power spectral density)
Up (Upturned-cutting forceps)	rz_FE (Fourier entropy), rz_AM (Absolute maximum), rz_S (Skewness) pz_PSD (Cross power spectral density), pz_0.1Q (Quantile)

px, py, pz, rx, ry, rz: position and rotation of the tip

features dependent on data length and those that took the same value across all subjects were excluded from the analysis as part of the preprocessing.

C. Skill Analysis

For the preprocessed features, a verification was conducted to check for significant differences among the three skill-level groups shown in Table I. The Kruskal-Wallis test was used with a significance level of 1%. Since tsfresh-generated features include similar features differing only in threshold values, only the feature with the smallest p-value was selected from among these. Additionally, to prevent multicollinearity due to features with similar meanings, the correlation coefficients between features were calculated, In combinations where the absolute value of the correlation coefficient exceeded 0.8, the feature with the smallest p-value was used. As a result of the test, 18 features were identified with significant differences among the three skill-level groups, as shown in Table II. For the feature names in Table II, the two letters before the underscore indicate a time series, and after the underscore are the feature abbreviations and their meanings. These 18 features can be broadly classified into four categories.

(a) Motion variability:AM, V, VC, RMS, K, RB1.5σ
(b) Time series bias: LSBM, 0.1Q, S
(c) Oscillatory features: PSD, FE
(d) Efficient features: LoM

The features classified in (a) represent the variability of the distribution, with higher values indicating greater movement of the surgical instruments. The features included in (b) represent the skewness of the data. The features included in (c) represent the oscillations of the movement. The feature in classification (d) represents the time at which the minimum value of the time series data first appeared, and the distribution was biased towards the second half of the surgery in skilled surgeons.

III. Model Interpretation Using SHAP

In this study, a learning model was built to classify surgeon skill levels, and the contribution of each feature in the model was visualised and analysed using SHapley Additive exPlanations (SHAP), a type of explainable AI.

A. Model Generation Details

A Support Vector Machine (SVM), a method that has shown high classification accuracy in previous studies by the authors [4], was used to generate the training model used to classify skill levels. The three classes classified based on the ESS-OSATS scores shown in Table I were used as Ground Truth for training the models. The features used for generating the learning model were the 18 features with significant differences among the three classes (Table II). The model with the best generalisation performance was constructed using 10-fold cross-validation and grid search. The metric used to evaluate the models was accuracy, the percentage of correct predictions divided by the number of correct predictions.

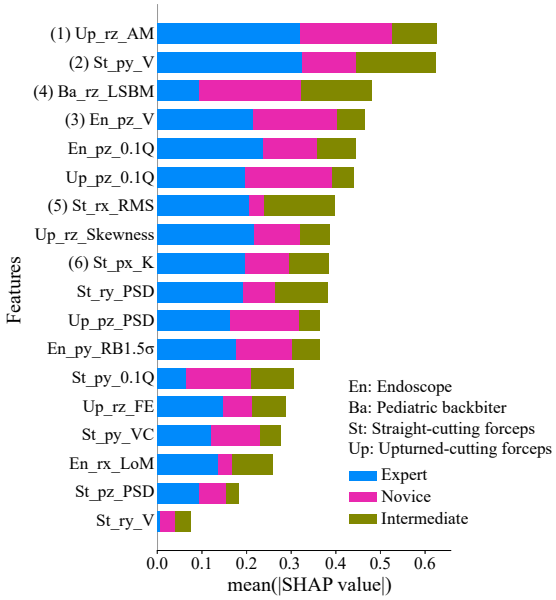


Fig. 3: Result of SHAP (summary plot). See Table II for feature abbreviations.

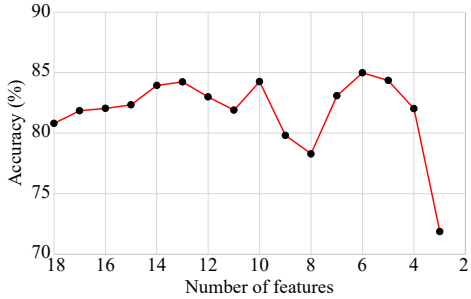


Fig. 4: The relationship between feature count and accuracy

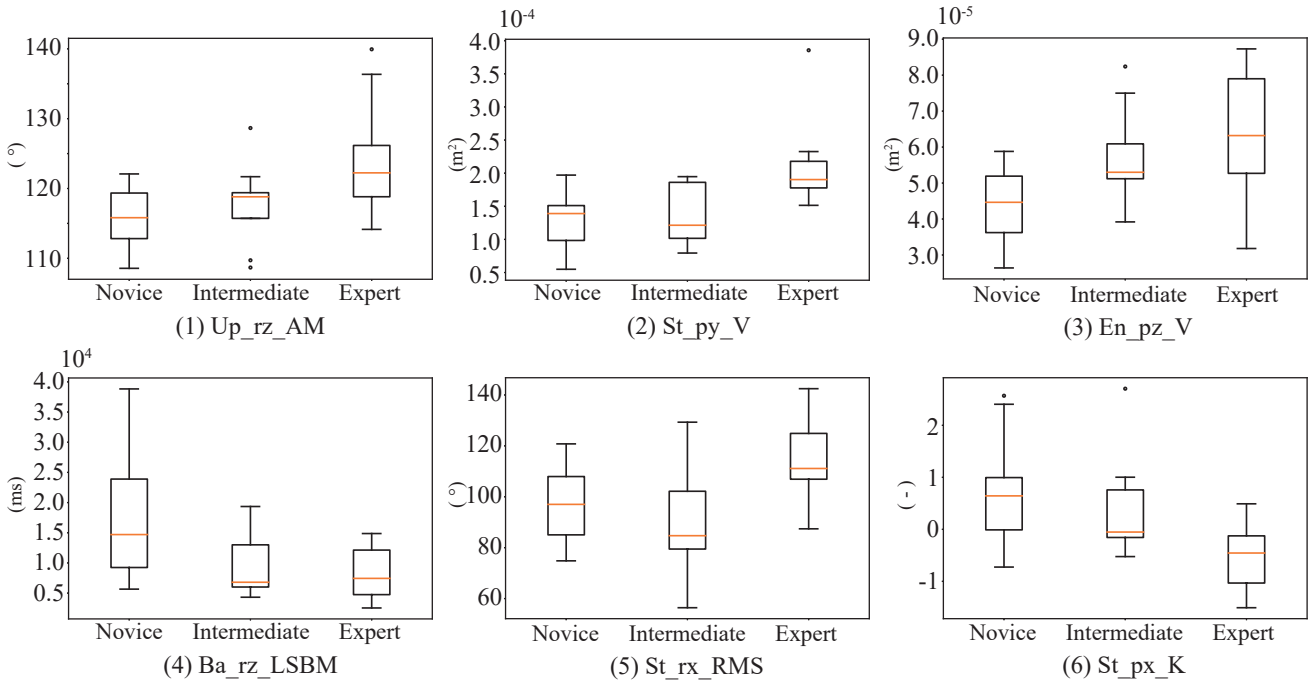


Fig. 5: Boxplots of the selected 6 features. See Table II for feature abbreviations.

B. Model Interpretation Using SHAP

Many machine learning models have the problem of not being able to clearly present the reasoning behind their processing results. However, to efficiently improve surgical skills, it is necessary to present problems in surgical operations. For this purpose, SHAP was used to analyze the skill of the subjects. SHAP is a game theoretically driven AI which computes Shapley scores. In classification models, SHAP is a method to approximate the Shaply value, which is the contribution of the features in determining the probability of belonging to each class. SHAP not only reveals the importance of an indicator in the overall machine learning model, but also provides the basis for individual prediction results. The model contribution for each feature calculated using SHAP for a model generated from 18 features is shown in Fig. 3. The contribution to each class is colour-coded, and the features are ordered in descending order of the total contribution. As shown in Fig. 3, features related to the variability and bias of movements occupy a high proportion of the contributions, indicating that the ability to move surgical instruments smoothly and evenly is essential for improving a surgeon's skill. On the other hand, features indicating movement vibration and surgical efficiency show relatively low contributions. This suggests that focusing on smooth manipulation of forceps, rather than shortening surgery time or performing surgery with small, frequent forceps manipulations, may lead to skill improvement.

IV. Feature Reduction

In this study, feature reduction was performed to prevent overfitting and improve classification accuracy. The variable reduction method was used as the approach for feature reduction. This method involves creating a model using all variables and then sequentially removing those with low contributions. The features to be reduced were determined based on the sum of the contributions per class calculated using SHAP. The process was repeated: one feature was deleted, the model was regenerated and the features to be reduced were determined based on the contributions calculated by SHAP. This process continued until the accuracy rate of the model dropped by more than 10% compared to the previous model. The relationship between the number of features and classification accuracy is shown in Fig. 4. The highest accuracy was achieved when the number of features was six. The differences between the skill level groups for the six features after feature reduction are shown in a box plot (Fig. 5). Note that (1) - (6) in Fig. 5 correspond to features (1) - (6) in Fig. 3. Variability in movement was greater in experts, while bias in data is more pronounced in novices. The frequent stagnation in movements observed in novices could have been attributed to a lack of understanding of the surgical flow or insufficient decision-making ability.

Next, to investigate whether feature reduction actually improved the model, the generalisation performance of the model with 18 features and the model with 6 features were evaluated. To validate the generalisation performance, nested 10-fold cross-validation was performed (Fig. 6). The optimal hyperparameters were determined by performing a grid search in the inner cross validation. It was also repeated 100 times to eliminate bias due to chance. The evaluation metric was accuracy, and the results of 100 trials were plotted as a box plot, shown in Fig. 7. The median accuracy of

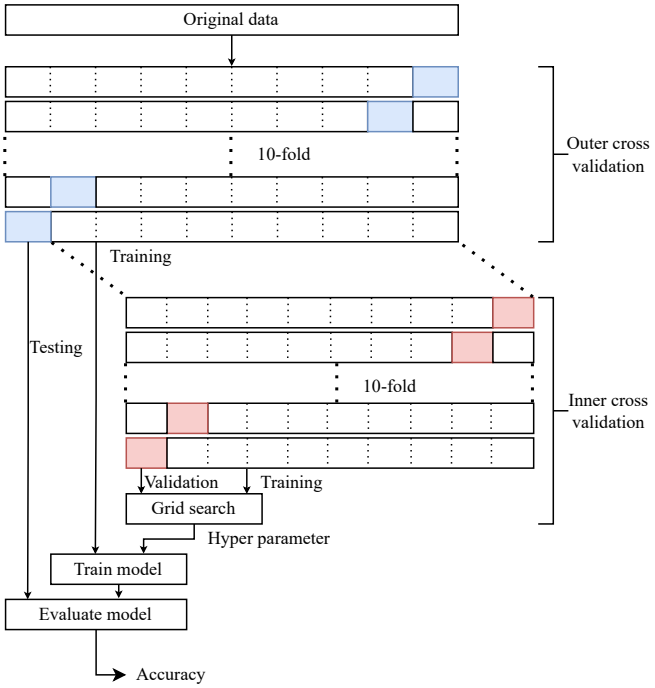


Fig. 6: Nested 10-fold cross validation

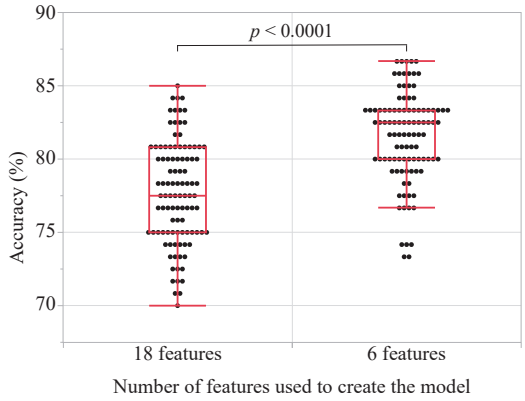


Fig. 7: Result of the accuracy evaluation test

the model using 18 features was 77.5 %, while the median accuracy of the model using 6 features was 82.5 %. A Friedman test was conducted to compare these models, and the result was $p < 0.0001$, indicating that the model using 6 features had significantly higher classification accuracy than the model using 18 features.

In conclusion, reducing the number of features based on the contributions calculated using SHAP effectively prevented overfitting and improved classification accuracy.

V. CONCLUSIONS

In this study, the analysis using SHAP, which calculates the contributions of features to the learning model, was conducted, and classification accuracy was improved by reducing features based on their contributions. The analysis based on SHAP revealed that features related to movement variability and bias had high contributions, while features related to forceps manipulation vibrations and surgical efficiency had low contributions. Additionally, the variable reduction method achieved the highest accuracy when the number of features was reduced to six, demonstrating that feature reduction based on contributions calculated by SHAP is an effective method for both preventing overfitting and improving classification accuracy.

Future prospects include conducting detailed analyses and evaluations for each specific surgical task and verifying the results in clinical trials.

REFERENCES

[1] K. Laeeq, et al.: "Achievement of competency in endoscopic sinus surgery of otolaryngology residents," The Laryngoscope, vol. 123, no. 12, pp. 2932–2934, 2013.

[2] T. Sugino, et al.: "Quantitative Analysis of a Camera Operation for Endoscopic Sinus Surgery Using a Navigation Information: Clinical Study," Clinical Study, Journal of Japan Society of Computer Aided Surgery, vol. 19, no. 1, pp. l7-26, 2017.

[3] N. Ahmidi, et al.: "An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data," International Forum of Allergy & Rhinology, vol. 2, no. 6, pp. 507–515, 2012.

[4] K. Yamada, et al.: "Development of the Classification System for Surgical Skills in Endoscopic Sinus Surgery", Abstract Booklet of the 16th World Congress of the International Federation for the Promotion of Mechanism and Machine Science, pp. 3-4, 2023.

[5] M. Christ, et al.: "Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)," Neurocomputing, vol. 307, pp. 72–77, 2018.

[6] K. Yamada, et al.: "Development of a Quantitative Evaluation System for Surgical Skill in Endoscopic Sinus Surgery", The 7th Jc-IFToMM International Symposium, vol. 7, pp. 55–62, 2024.

[7] S. M. Lundberg and Su-In Lee: "A Unified Approach to Interpreting Model Predictions, Proceedings of the 31st International Conference on Neural Information Processing Systems," pp. 4768-4777, 2017.

[8] M. Suzuki, et al.: "Repetitive simulation training with novel 3D-printed sinus models for functional endoscopic sinus surgeries," Laryngoscope Investigative Otolaryngology, vol. 7, no. 4, pp. 943–954, 2022.

[9] P. H. Shen, et al.:"Retrospective study of full-house functional endoscopic sinus surgery for revision endoscopic sinus surgery," International Forum of Allergy & Rhinology, vol. 1, no. 6, pp. 498–503, 2011.

[10] S.Y. Lin, et al.: "Development and Pilot-Testing of a Feasible, Reliable, and Valid Operative Competency Assessment Tool for Endoscopic Sinus Surgery," American Journal of Rhinology & Allergy, vol. 23, no. 3, pp. 354–359, 2009.