Machine learning-based automatic evaluation of tissue handling skills in laparoscopic colorectal surgery: A retrospective experimental study

Shoma Sasaki[1,3], MD; Daichi Kitaguchi[1,2], MD; Shin Takenaka[2], MD, PhD; Kei Nakajima[2], MD; Kimimasa Sasaki[2], MD; Tateo Ogane[2]; Nobuyoshi Takeshita[1,2], MD, PhD; Naoto Gotohda[3], MD, PhD; Masaaki Ito[1,2], MD, PhD

[1]Department of Colorectal Surgery, National Cancer Center Hospital East, 6-5-1, Kashiwanoha, Kashiwa-City, Chiba 277-8577, Japan

[2]Surgical Device Innovation Office, National Cancer Center Hospital East, 6-5-1, Kashiwanoha, Kashiwa-City, Chiba 277-8577, Japan

[3]Course of Advanced Clinical Research of Cancer, Juntendo University Graduate School of Medicine, 2-1-1, Hongo, Bunkyo-ku, Tokyo 113-8421, Japan

Corresponding Author: Masaaki Ito, MD, PhD; Surgical Device Innovation Office, National Cancer Center Hospital East, 6-5-1, Kashiwanoha, Kashiwa, Chiba 277-8577, Japan (matio@east.ncc.go.jp).

Requests for reprints should be addressed to Masaaki Ito, MD, PhD; Surgical Device Innovation Office, National Cancer Center Hospital East, 6-5-1, Kashiwanoha, Kashiwa, Chiba 277-8577, Japan (+81-4-7133-1111).

Conflicts of Interest and Source of Funding: This study was funded by the Japan Agency for Medical Research and Development (20vk0124001h0001). All authors have no conflicts of interest or financial ties to disclose.

Running Head: Evaluation of tissue handling skills

INTRODUCTION

Laparoscopic surgery has become increasingly prevalent in many areas of abdominal surgery. While laparoscopic surgery provides patients with early postoperative recovery[1,2], it also requires surgeons to have advanced surgical skills. Disparities have been reported in surgical outcomes between skilled and unskilled surgeons, especially in the field of laparoscopic gastrointestinal surgery[3-8]. Accordingly, it is of paramount importance to provide surgical education based on objective and quantitative evaluation of surgeons' skills to enhance patient safety and prevent harmful intraoperative problems, such as bleeding and tissue damage.

Tissue handling, or respect for tissue, is one of the elements of surgical skills. Rough tissue handling must be avoided because it can cause tissue impairment and bleeding. Further, the resulting blood can hide critical structures or obscure the correct dissection layer, leading to secondary adverse events. Taken together, surgeons should receive appropriate feedback by which to improve their tissue handling skills if they perform procedures that result in bleeding or oozing. The commonly used existing surgical skill assessment tools, such as the Objective Structured Assessment of Technical Skills (OSATS)[9] and Global Operative Assessment of Laparoscopic Skills (GOALS)[10], include tissue handling as an item in the assessment.

However, the problems with these tools are that subjectivity, or rater bias, is inevitable because an expert evaluates surgeons’ tissue handling skills and that they are not purely quantitative as evaluations are made on a scale of one to five. Observational clinical human reliability analysis

(OCHRA), which can reduce subjectivity by counting errors, has a time consuming and labor-intensive nature, involving experts to identify errors throughout unedited operation videos[11]. In this context, automated evaluations through the application of artificial intelligence and machine learning could be highly beneficial.

According to a previous analysis of OCHRA, most tissue handling errors are reflected in the number of bleeding errors[12], which suggests that tissue handling can be evaluated by bleeding events; however, assessments are difficult using the conventionally chosen estimated blood loss because the assessment of tissue handling should consider not only major hemorrhage, as reflected in the amount of blood loss, but also minor hemorrhage, such as oozing. Therefore, a system that can automatically detect even minor hemorrhage is necessary for automatic tissue handling skill evaluation. Additionally, to our knowledge, our study provides the first report of an automated assessment of tissue handling skill using videos of actual operations. This study aimed to develop a machine learning model that automatically quantifies the spread of blood in the surgical field using intraoperative videos of laparoscopic colorectal surgery and evaluated whether indices measured with the developed model could be used to automatically assess tissue handling skill.

MATERIALS AND METHODS

Study design

This was a retrospective experimental study using intraoperative videos of laparoscopic colorectal surgery collected from multiple institutions in Japan. The protocol for this study was reviewed and approved by the Ethics Committee of the National Cancer Center Hospital East (Registration No.: 2020-315). This study conforms to the provisions of the Declaration of Helsinki 1964 (as revised in Brazil in 2013).

Machine learning-based blood pixel classification

A total of 504 blood areas and 504 non-blood areas were extracted from 28 intraoperative videos of laparoscopic colorectal surgery as small images and used to train the machine learning-based classification model. These 28 intraoperative videos were randomly selected from the video data set described below. A representative image is shown in Figure 1. The area of the individual small images was 10 to 150 pixels. In general, the colors of pixels on a screen can be expressed in terms of RGB values, which represent the ratio of mixing red (R), green (G), and blue (B) colors. These three primary colors each can be set to one of 256 levels of brightness, and thus can represent a total of 16,777,126 different hues. The RGB values of all pixels in the extracted areas were obtained.

All pixels were divided into three datasets: training, validation, and test of the algorithm. The validation dataset was utilized to adjust the parameters to ensure that the best model performance was achieved and to prevent the model from overfitting. The data were split at a per-case level rather than at a per-pixel level. Therefore, the pixels from a video in the training dataset were not included in the other datasets; that is, the assessment was not circular. The hyper-parameters of the machine learning model to classify pixel RGB values into blood and non-blood were optimized to maximize the classification accuracy. Logistic regression analysis was chosen as the machine learning method in this study.

The evaluation metrics for the machine learning model's classification of blood or non-blood were overall accuracy, sensitivity, and specificity. The calculation formula for each metric is shown below. TP represents the number of blood pixels correctly classified, TN represents the number of non-blood pixels correctly classified, FN represents the number of blood pixels incorrectly classified as non-blood pixels, and FP represents the number of non-blood pixels incorrectly classified as blood pixels.

$$Overall\ accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{FP+TN}$$

Video dataset for surgical skill assessment

To apply the results of the machine learning-based blood pixel measurements to the assessment of surgical skill, especially tissue handling skill, a video dataset was constructed. The videos used for surgical skill assessment were surgical videos of laparoscopic sigmoidectomy (Lap-S) submitted by hospitals throughout Japan to the Endoscopic Surgical Skill Qualification System (ESSQS) established by the Japan Society for Endoscopic Surgery between 2016 and 2017. ESSQS uses a scoring system in which the evaluation criteria are divided into two categories: common criteria, which correspond to basic endoscopic techniques that are commonly used for all procedures, and organ-specific criteria, which correspond to special endoscopic surgical techniques for individual organs[13]. The details of ESSQS are shown in the Supplementary table 1.

In ESSQS, two experts score tissue handling skill, which is one of the elements of common criteria, in each video from 0 to 3 points. The “high tissue handling score group” was defined as surgical videos in which the sum of the tissue handling scores given by the two experts was 5 or more out of 6, while the "low tissue handling score group" was defined as surgical videos in which either one of the two experts assigned a score of 0.

ESSQS evaluates not only tissue handling but also other technical factors, such as the development of the surgical field, efficiency of the procedure, and the surgeon's autonomy, for a total score of 100 points. Videos with an ESSQS overall score in the +2SD (standard deviation) range were defined as the “overall score +2SD group” and videos with an overall score in the -2SD range as the “overall score -2SD group.”

Since applicants for the ESSQS must be board certified surgeons of the Japan Surgical Society and must have performed at least 20 highly difficult surgeries, such as esophageal surgery, gastrectomy, and colorectal resection within the last 3 years, the skills of these applicants are considered to have reached at least an intermediate level. To include groups with lower skills in the analysis, Lap-S videos performed by beginner surgeons who had conducted fewer than five Lap-S surgeries were collected from multiple institutions and defined as the “novice surgeon group.”

Blood pixel count-based surgical skill assessments

All video data were divided into static frames every 1/30 s (30 fps), and the number of pixels classified as blood by the machine learning model developed in this study was counted in all pictures. A conceptual diagram from the blood/non-blood pixel classification to the automatic counting of blood pixels for tissue handling skill evaluation is shown in Figure 2.

To eliminate the influence of the surgery duration, the blood pixel count per frame, which is the number of blood pixels in the entire video divided by the number of frames, was used in the tissue handling skill evaluation.

To test whether the blood pixel count per frame indicated good or poor tissue handling skills, the index was compared among three groups: the high tissue handling score group, the low tissue handling score group, and the novice surgeon group. We hypothesized that there would be differences in the blood pixel count per frame among these three groups, reflecting differences in tissue handling skills.

In addition, to examine whether the index reflected overall surgical skill, comparisons were made among the overall score +2SD group, the overall score -2SD group, and the novice surgeon group. A supplemental analysis verified whether conventionally estimated blood loss correlated with tissue handling skills. Our hypotheses were that there would be no difference in the blood pixel count per frame among the overall score +2SD group, the overall score -2SD group, and the novice surgeon group because overall surgical skill consists of several elements; not only tissue handling, but also the development of the surgical field, the surgeon’s autonomy, and efficiency during the procedure.

Estimated blood loss was not anticipated to correlate with tissue handling skills because blood loss in laparoscopic colorectal surgery is small in absolute terms, regardless of surgical skill.

Statistical analysis

Quantitative data are reported as mean [standard deviation]. Two-group comparisons used t-tests and three-group comparisons used one-way ANOVA. When one-way ANOVA indicated significant differences, each between-group post-hoc comparison was conducted using the Tukey method. All tests performed were two-sided, with the level of significance set at $p < 0.05$. All statistical analyses were performed using EZR (Saitama Medical Center, Jichi Medical University, Saitama, Japan), a graphical user interface for R (R Foundation for Statistical Computing, Vienna, Austria). More precisely, EZR is a modified version of R Commander designed to add functions frequently used in biostatistics[14].

RESULTS

Machine learning-based blood pixel classification

In total, 1008 images were extracted from 28 videos of laparoscopic colorectal surgery. The images contained 23736 blood pixels and 20994 non-blood pixels. The training data consisted of 34988 pixels extracted from 20 cases, the validation data 4606 pixels from 4 cases, and the test data 5136 pixels from 4 cases. Figure 3 shows a three-dimensional mapping of blood pixels and non-blood pixels in RGB space. The inference results from the machine learning model applied to the test dataset are shown in the Supplementary table 2. The overall accuracy of the machine learning model was 85.7%. The sensitivity and specificity were 99.9% and 72.7%, respectively.

Automatic tissue handling skill assessment

The number of videos used in the analysis was 60 from the high tissue handling score group, 55 from the low tissue handling score group, and 36 from the novice surgeon group. The results of using the present machine learning model on these videos to obtain the blood pixel count per frame are shown in Figure 4. One-way ANOVA revealed significant differences among the three groups for blood pixel count per frame (high tissue handling score group 20972.23 [19287.05] vs. low tissue handling score group 34473.42 [28144.29] vs. novice surgeon group 50630.04 [42427.76], $p < 0.01$). Pairwise comparisons revealed that the high tissue handling score group had significantly fewer blood pixels per frame than did the low tissue handling score group and the novice surgeon group. The low tissue handling score group had significantly fewer such pixels than the novice surgeon group.

Correlation between blood pixel count per frame and overall surgical skill score

The number of videos used in the analysis was 17 from the overall score +2SD group and 20 from the overall score -2SD group. Comparison of these groups with the novice surgeon group indicated no significant difference in the blood pixel count per frame (overall score +2SD group 33341.44 [25450.99] vs. overall score -2SD group 39203.92 [39364.17] vs. novice surgeon group 50630.04 [42427.76], $p = 0.267$; Figure 5). That is, there was no correlation between the blood pixel count per frame and the overall surgical skill score.

Correlation between blood loss and tissue handling score

The estimated blood loss did not differ significantly among the high and low tissue handling score groups and the novice surgeon group (high score 7.10 [11.3] ml vs. low score group 12.8 [19.2] ml vs. novice 12.9 [26.4] ml, $p = 0.461$; Figure 6A). In addition, no significant difference was revealed in blood loss among the overall score +2SD and -2SD groups and the novice group. (+2SD 19.8 [28.0] ml vs. -2SD 15.0 [24.1] ml vs. novice 12.9 [26.4] ml, $p = 0.793$; Figure 6B). Based on the above results, conventionally estimated blood loss did not correlate with either tissue handling skill alone or overall skill.

DISCUSSION

In this study, we successfully developed a machine learning model that evaluates the number of blood pixels using RGB information and showed that the blood pixel count per frame calculated by the proposed model significantly differed among groups with different tissue handling skills. These results suggest the possibility of objective, quantitative evaluation of tissue handling without requiring the time and effort of experts.

The blood pixel count per frame, which is the total blood pixel count for the entire surgery divided by the number of frames, was significantly lower at higher levels of tissue handling. This suggests that the model may have correctly determined that tissue damage and oozing due to inappropriate grasping or traction are less common in those who are skilled in tissue handling. Surprisingly, pairwise comparisons showed that this model could distinguish poor and good tissue handling, not only among extremely unskilled surgeons, but also among surgeons who had reached a certain level of general surgical skill. Thus, the blood pixel count per frame may be a highly reliable indicator for automatically determining the level of tissue handling for surgeons at any level.

Comparison of the +2SD group and -2SD group for overall scores and the novice surgeon group showed no significant difference in the blood pixel count per frame between any two groups. Overall surgical skill is determined by a combination of factors, including development of the surgical field, the surgeon's autonomy and efficiency during the procedure, as well as tissue handling. Therefore, the difference in tissue handling skill was likely obscured by grouping the videos according to overall surgical skill. This result suggests that the present model can evaluate specifically the level of tissue handling, as intended.

Blood loss did not correlate with either tissue handling skill or overall skill. It is likely that the surgeon's skill makes no difference to the amount of blood loss because laparoscopic surgery involves less blood loss compared to open surgery. Chen et al. reported that, for laparoscopic surgery for colorectal cancer, surgery time decreased with experience, but there was no significant difference in blood loss[15], a finding compatible with the results of the present study. Our study confirmed that it is difficult to assess surgeons' tissue handling skill using existing parameters and that new measures are needed.

Regarding the reliability of the scores used in this study, the video test was based on an anonymized, unedited, random video review, and was scored by two to three experts in laparoscopic surgery designated by the Japan Society for Endoscopic Surgery; the scores were, therefore, likely highly reliable. Many skilled surgeons review multiple surgical videos each year for ESSQS, which is a time- and labor-demanding process. To mitigate this burden, a feasible automated review system for surgical procedures is highly worthwhile.

Kitaguchi et al. reported the use of three-dimensional convolutional neural network model for automatic surgical skill assessment (ASSA)[16]. In their report, they mentioned that classification by surgical skill assessment items may lead to further development of ASSA. To the best of our knowledge, the current study is the first report of an automated assessment of one component of laparoscopic surgical techniques using videos of actual operations. This model is simultaneously objective and quantitative. Subdividing surgical skills into subcategories for each assessment item, as in this study, may allow surgeons to obtain grades for each assessment item, which could lead to constructive feedback.

This study has several limitations. First, the overall accuracy of the machine learning model was not markedly high. Further improvement of the performance of the machine learning model would increase the reliability of the automatic tissue handling skill assessment. Second, this was a retrospective study, and hence, subject to selection bias. In addition, definitions of groups, such as novice surgeons, are arbitrary. In the future, further validation with a more diverse cohort of skill levels is necessary. Third, the number of cases in this study was relatively small, although far larger than previous studies that reported laparoscopic surgical skill assessment methods, most of which involved fewer than 30 participating trainees[17]. Further investigations with more subjects are warranted. Moreover, in the current study, we chose ESSQS as the reference source for the surgeon's tissue handling skill due to the accessibility of data. Correlations with OSATS or GOALS scores should also be examined in the future. We acknowledge that the present study is a proof-of-concept and does not report any data directly linked to clinical outcomes. However, multiple studies have shown that better clinical outcomes are achieved when laparoscopic surgeries are performed by skilled surgeons[3-8]. Therefore, our study represents progress toward automatic surgical skill evaluations, although the present index can only evaluate tissue handling, which is one element of surgical skills. We plan to conduct a prospective study to verify the relationship with clinical data when a comprehensive automated skill evaluation system is established.

In conclusion, we developed a machine learning model that can automatically count blood pixels in laparoscopic colorectal surgeries. The blood pixel count per frame correlated with the level of tissue handling skill. Automatic evaluation of surgical skills has significant potential to inform education for laparoscopic surgery.

ACKNOWLEDGMENTS

We would like to thank Editage (www.editage.com) for English language editing.

REFERENCES

1. McKay GD, Morgan MJ, Wong SK, et al. Improved short-term outcomes of laparoscopic versus open resection for colon and rectal cancer in an area health service: a multicenter study. *Dis Colon Rectum*. January 2012;55:42–50. doi: 10.1097/DCR.0b013e318239341f.

2. van der Pas MH, Haglind E, Cuesta MA, et al. Laparoscopic versus open surgery for rectal cancer

(COLOR II): short-term outcomes of a randomised, phase 3 trial. *Lancet Oncol*. March 2013;14:210–218. doi: 10.1016/S1470-2045(13)70016-0.

3. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med 2013;369(15):1434–1442. doi: 10.1056/NEJMsa1300625.

4. Stulberg JJ, Huang R, Kreutzer L, et al. Association between surgeon technical skills and patient outcomes. JAMA Surg 2020;155(10):960–968. doi: 10.1001/jamasurg.2020.3007.

5. Ichikawa N, Homma S, Funakoshi T, et al. Impact of technically qualified surgeons on laparoscopic colorectal resection outcomes: results of a propensity score-matching analysis. BJS Open 2020;4(3):486–498. doi: 10.1002/bjs5.50263.

6. Aoyama S, Inoue Y, Ohki T, et al. Usefulness of the endoscopic surgical skill qualification system in laparoscopic colorectal surgery: short-term outcomes: a single-center and retrospective analysis. BMC Surg 2019;19(1):90. doi: 10.1186/s12893-019-0528-2.

7. Kazama K, Numata M, Aoyama T, et al. Does the Encoscopic Surgical Skill Qualification System improve patients' outcome following laparoscopic surgery for colon cancer? A multicentre, retrospective anaylsis with propensity score matching. World J Surg Oncol. Feburary 2021;19(1):53. doi: 10.1186/s12957-021-02155-z.5.

8. Akagi T, Endo H, Inomata M, et al. Clinical impact of Endoscopic Surgical Skill Qualification System (ESSQS) by Japan Society for Endoscopic Sugery (JSES) for laparoscopic distal gasterectomy and low anterior resection based on the National Clinical Database (NCD) registry. Ann

Gastroenterol Surg. August 2020;4(6):721-734. doi: 10.1002/ags3.12384.

9. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. February 1997;84:273–278. doi: 10.1046/j.1365-2168.1997.02502.x.

10. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. July 2005;190:107–113. doi: 10.1016/j.amjsurg.2005.04.004.

11. Tang B, Cuschieri A. Objective assessment of surgical operative performance by observational clinical human reliability analysis (OCHRA): a systematic review. *Surg Endosc*. April 2020;34:1492–1508. doi: 10.1007/s00464-019-07365-x.

12. Miskovic D, Ni M, Wyles SM et al. Observational clinical human reliability analysis (OCHRA) for competency assessment in laparoscopic colorectal surgery at the specialist level. *Surg Endosc*. March 2012;26:796–803. doi: 10.1007/s00464-011-1955-z.

13. Mori T, Kimura T, Kitajima M. Skill accreditation system for laparoscopic gastroenterologic surgeons in Japan. *Minim Invasive Ther Allied Technol*. 2010;19:18–23. doi: 10.3109/13645700903492969.

14. Kanda Y. Investigation of the freely available easy-to-use software "EZR" for medical statistics. *Bone Marrow Transplant*. March 2013;48:452–458. doi: 10.1038/bmt.2012.244.

15. Chen G, Liu Z, Han P et al. The learning curve for the laparoscopic approach for colorectal cancer: a single institution's experience. *J Laparoendosc Adv Surg Tech A*. January 2013;23:17–21.

doi: 10.1089/lap.2011.0540.

16. Kitaguchi D, Takeshita N, Matsuzaki H et al. Development and validation of a 3-dimensional convolutional neural network for automatic surgical skill assessment based on spatiotemporal video analysis. *JAMA Netw Open*. August 2 2021;4:e2120786. doi: 10.1001/jamanetworkopen.2021.20786.

17. Haug TR, Ørntoft MW, Miskovic D et al. How can surgical skills in laparoscopic colon surgery be objectively assessed?-a scoping review. *Surg Endosc*. 2022;36:1761–1774. doi: 10.1007/s00464-021-08914-z.

Figure titles and legends

Figure 1.

Figure title: How images were created for the extraction of blood/non-blood pixels.

Figure legend: The areas that the surgeon determined to be blood/nonblood are shown, with image sections 4 through 9 being nonblood areas and the remainder being blood areas. Several still images were taken from each case, and blood/non-blood images were created from different scenes.

Figure 2.

Figure title: Flow from the blood/non-blood pixel classification to the automatic counting of blood pixels.

Figure legend: a) Blood pixels and non-blood pixels were extracted from videos of colorectal surgeries and were input as training data. b) The machine learning model was fine-tuned using the validation dataset and evaluated for accuracy using the test dataset. c) This model was applied to the evaluation of the tissue handling skill by counting the number of blood pixels in surgical videos of laparoscopic sigmoidectomy in another data set.

Figure 3.

Figure title: Plot in RGB space of extracted pixels.

Figure legend: A space where three axes consist of the brightness of red (R), green (G), and blue (B). Each of these three primary colors could be set to one of 256 levels of luminance, although they are linearly reduced on the axes of the figure and expressed from 0 to 1. Each point in the space corresponds to an individual pixel. The red points represent pixels extracted from blood areas and the blue pixels those extracted from non-blood areas. The black points had RGB values that could not be distinguished as blood/non-blood in this model. The surface formed by the black points serves as a boundary surface that distinguishes blood from non-blood.

Figure 4.

Figure title: Blood pixel count per frame: High and low tissue handling score and novice surgeon groups.

Figure legend: Significant differences were found in all two-group comparisons, with the group considered to have a higher level of tissue handling having a lower blood pixel count.

Figure 5.

Figure title: Blood pixel count per frame in the overall score: +2SD, -2SD, and novice surgeon groups.

Figure legend: One-way ANOVA comparison of the three groups revealed no significant differences.

Figure 6.

Figure title: The amount of blood loss among groups.

Figure legend: A**:** A comparison of blood loss between the high and low tissue handling score groups revealed no significant difference. B: There was no significant difference in the amount of blood loss between the overall score +2SD and -2SD groups.

Figure 1

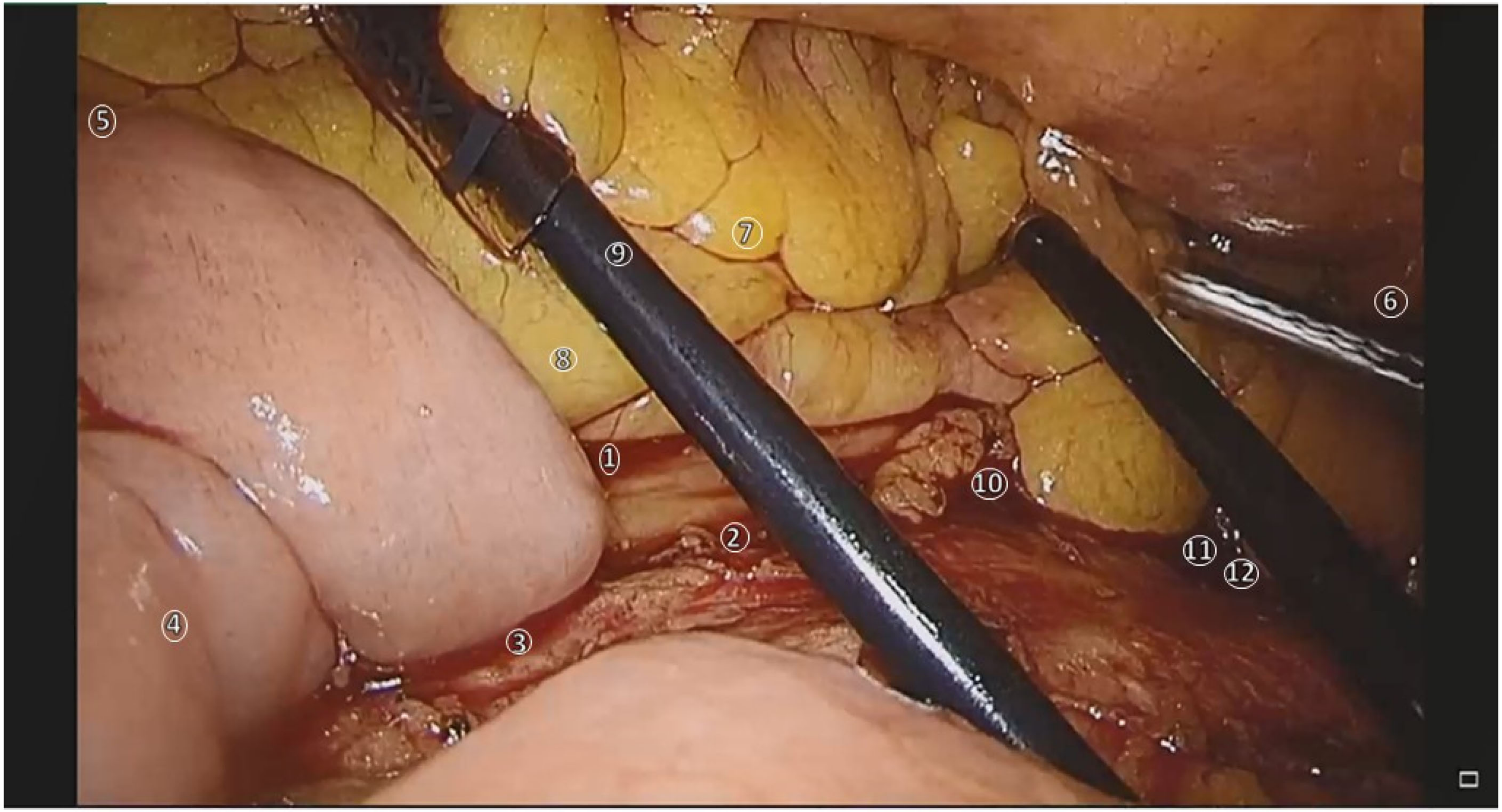


Figure 2

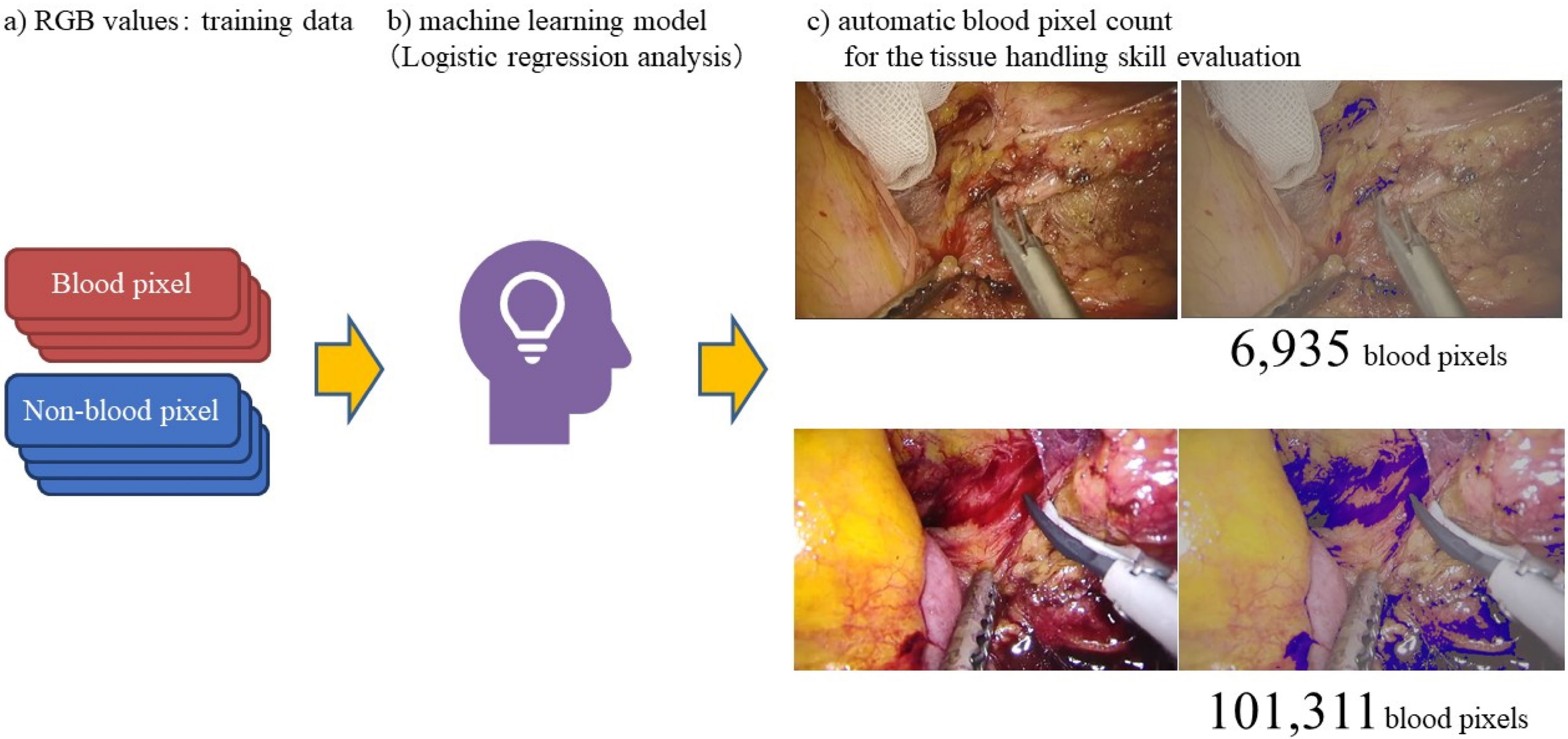
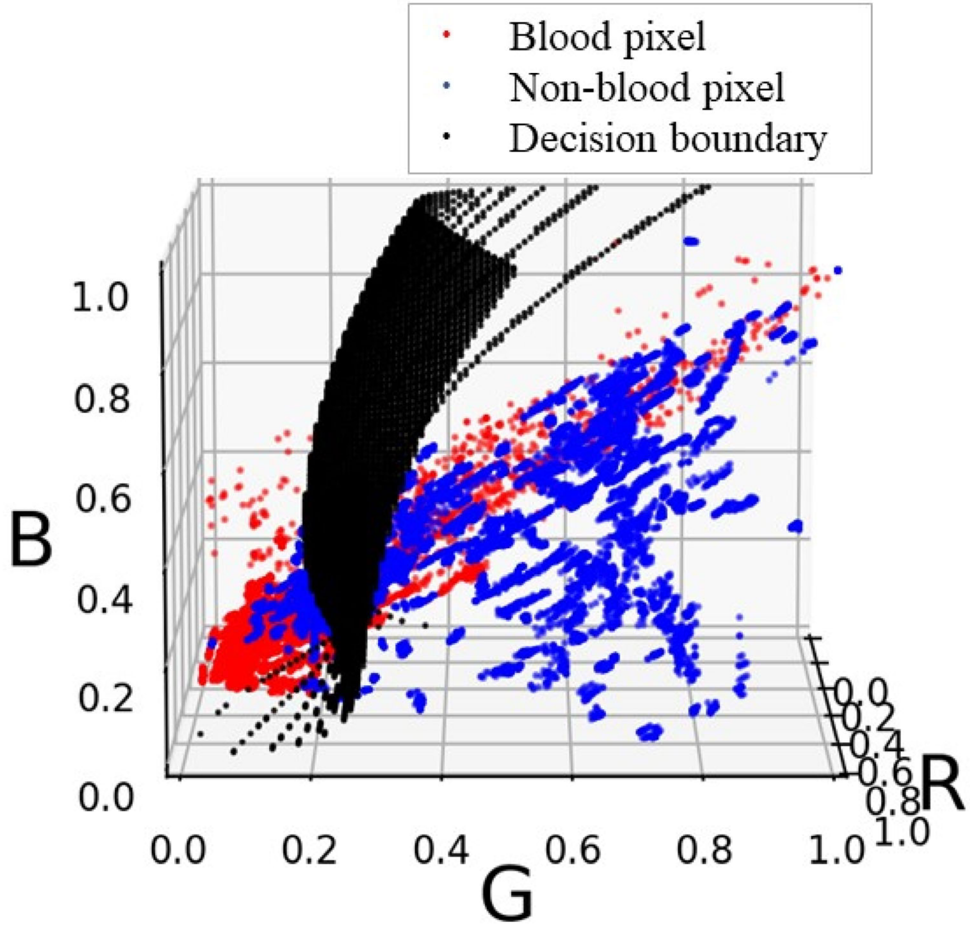

a) RGB values: training data
b) machine learning model
(Logistic regression analysis)
c) automatic blood pixel count
for the tissue handling skill evaluation
Blood pixel
Non-blood pixel
6,935 blood pixels
101,311 blood pixels

Figure 3



Blood pixel
Non-blood pixel
Decision boundary
B
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
G
0.0
0.2
0.4
0.6
0.8
1.0
R

Figure 4

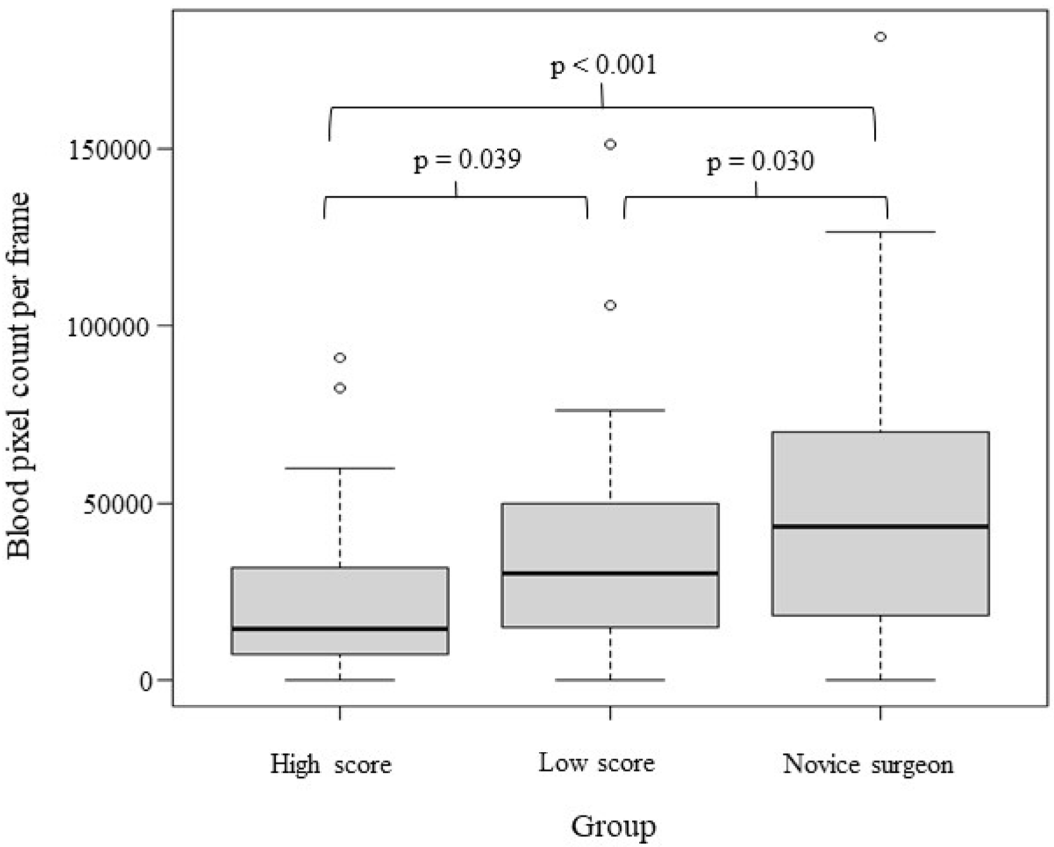
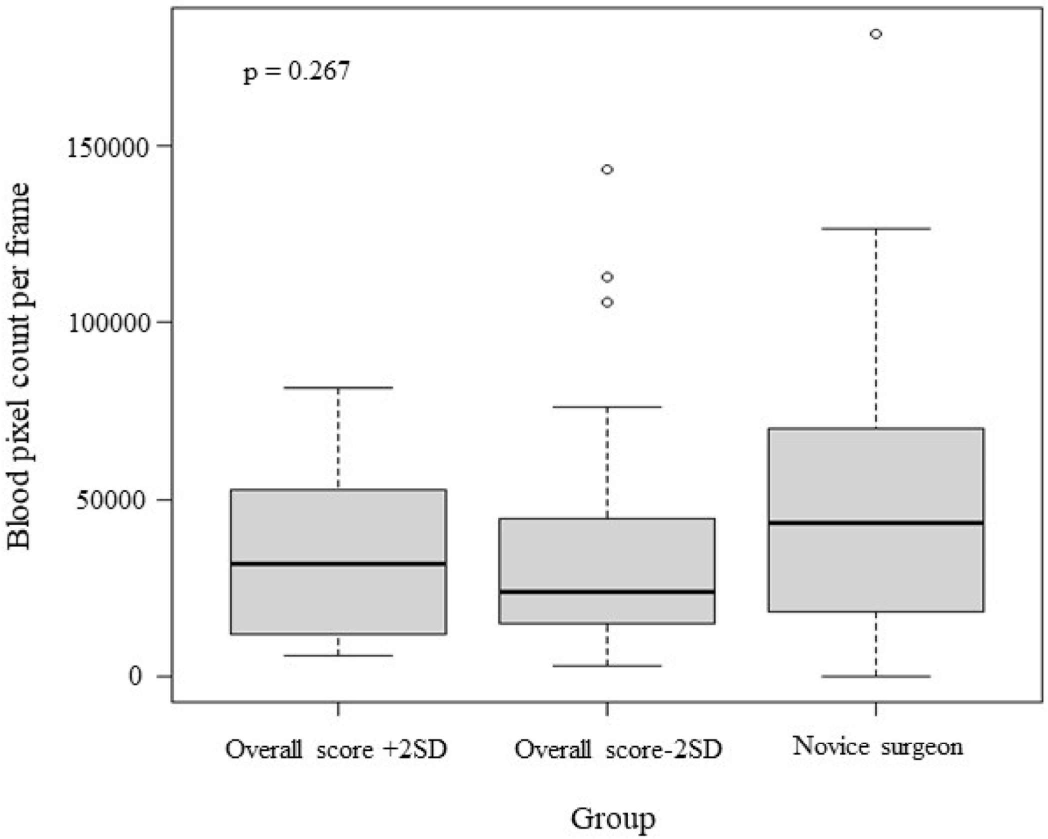
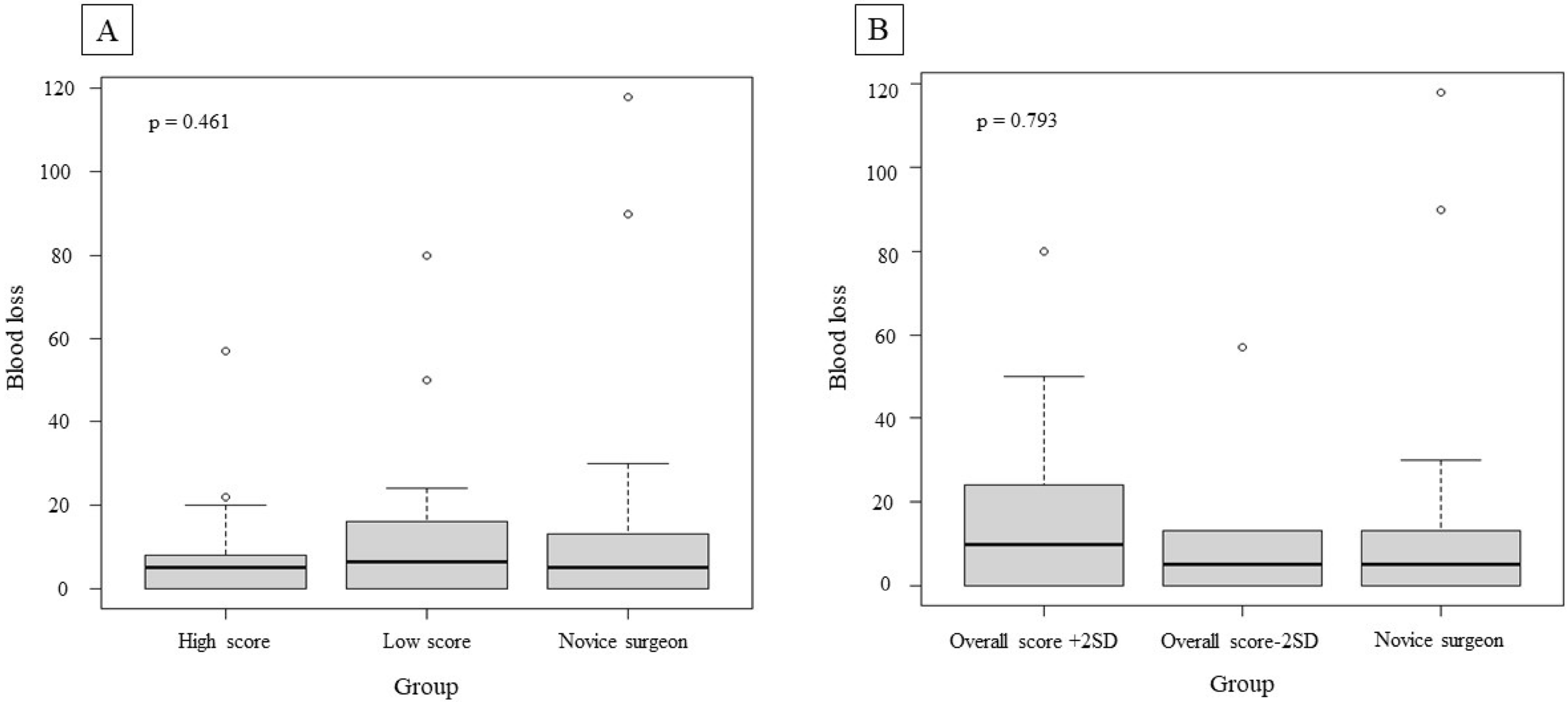

p < 0.001
p = 0.039
p = 0.030
150000
100000
50000
0
Blood pixel count per frame
High score
Low score
Novice surgeon
Group

Figure 5



p = 0.267
Blood pixel count per frame
150000
100000
50000
0
Overall score +2SD
Overall score-2SD
Novice surgeon
Group

Figure 6


A
p = 0.461
Blood loss
120
100
80
60
40
20
0
High score
Low score
Novice surgeon
Group
B
p = 0.793
Blood loss
120
100
80
60
40
20
0
Overall score +2SD
Overall score-2SD
Novice surgeon
Group