Surgical Endoscopy (2025) 39:6560–6568
https://doi.org/10.1007/s00464-025-12015-6

Using artificial intelligence to model expert panel diagnosis of cholecystitis severity

Griffin H. Olsen[1] · Emmett D. Goodman[2,3] · Josiah G. Aklilu[2] · Sebastiano Bartoletti[4] · Kay S. Hung[4] · Janice H. Yang[2] · Eric C. Sorenson[5] · Jeffrey K. Jopling[6] · Serena Y. Yeung[2,3,7,8] · Dan E. Azagury[4]

Received: 11 March 2025 / Accepted: 13 July 2025 / Published online: 18 August 2025

Abstract

Background Determining cholecystitis severity via the clinically validated Parkland Grading Scale (PGS) is useful for predicting case difficulty and likelihood of postoperative complications. A panel assessment by multiple surgeons can reduce variation in PGS due to subjectivity, but is time-consuming. An artificial intelligence (AI) model trained on the assessments of an expert clinician panel may improve efficiency and reduce variability in diagnosis in image-based assessments.

Methods Laparoscopic cholecystectomy videos were obtained from one public and two private data sources. Representative frames were chosen for PGS grading and manually labeled. Three surgical experts independently assigned PGS scores to the selected frames. They then convened as a panel to decide on the score if those were discrepant at individual scoring. Weighted Cohen's kappa statistic was measured for inter-rater variability. Two AI models were developed for automated PGS grading and their accuracy and interpretability evaluated.

Results 319 videos were compiled. Three surgical experts independently assigned identical PGS grades for 51% of cases, and weighted Cohen's kappa statistics ranged between 0.76 and 0.83. The accuracy of Model A using absolute agreement with the expert panel's consensus was 69%, and weighted Cohen's kappa statistic was 0.62. The accuracy of Model B using absolute agreement with the panel's consensus was 72%, and weighted Cohen's kappa statistic was 0.77. Interpretability analysis was conducted. Three anatomical structures played a key role in Model B's grading of cholecystitis severity: the appearance of the gallbladder, liver, and omentum had notable impact on performance.

Conclusions A transformer-based AI model can be trained on consensus from an expert panel to predict ratings of cholecystitis severity (Parking Grading Scale), performing competitively with some individual experts at predicting PGS when compared to the panel-based ground truth. However, variance and subjectivity of PGS remain, thus presenting its limitations as a ground truth for computer vision-based models.

Keywords Artificial intelligence · Computer vision · Cholecystitis · Diagnosis prediction · Inter-rater variability

Gallstone disease is common worldwide, and many patients require surgical treatment [1, 2]. Laparoscopic

Griffin H. Olsen and Emmett D. Goodman are co-first authors.

Serena Y. Yeung and Dan E. Azagury are co-senior authors.

✉ Kay S. Hung
kayhung@stanford.edu

[1] Intermountain Healthcare Delivery Institute, Intermountain Health, Salt Lake City, UT, USA

[2] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

[3] Department of Computer Science, Stanford University, Stanford, CA, USA

[4] Department of Surgery, Stanford University School of Medicine, 300 Pasteur Drive, Room H3591, Stanford, CA 94305-5641, USA

[5] Department of Surgery, Intermountain Medical Center, Murray, UT, USA

[6] Department of Surgery, Johns Hopkins School of Medicine, Baltimore, MD, USA

[7] Department of Electrical Engineering, Stanford University, Stanford, CA, USA

[8] Clinical Excellence Research Center, Stanford University School of Medicine, Stanford, CA, USA

cholecystectomy is the recommended therapy for patients with symptomatic gallstones and one of the most commonly performed surgical procedures in the United States [3–5]. Clinical symptoms, physical examination findings, and diagnostic studies can suggest the severity of cholecystitis. However, the true degree of inflammation is fully ascertained during surgery [6, 7]. Determining cholecystitis severity is useful for intraoperative planning and retrospective reporting of clinical outcomes. The Parkland Grading Scale is a five-tiered scale for assessing cholecystitis severity based on visual indications of inflammation during surgery [8]. The scale ranges from one, indicating normal anatomy, to five, indicating necrosis and perforation. The Parkland Grading Scale has been clinically validated and shown to outperform rival cholecystitis classification systems for predicting case difficulty and the likelihood of postoperative complications [9, 10].

The creators of the Parkland Grading Scale envisioned a way for multiple surgeons to assess cholecystitis severity using intraoperative images stored within a patient's electronic medical record [8]. Combining expertise to produce a panel diagnosis has been shown to reduce variation due to subjectivity for image-based disease assessments [11–14]. However, clinician case review can be time-consuming. Thus, it is pertinent to automate assessment to reduce postoperative clinician workload and potentially reduce variation in diagnosis due to subjectivity in image-based assessment. There is growing interest in using artificial intelligence (AI) to rapidly analyze videos of laparoscopic surgery [15, 16]. Previous research has established the feasibility of building AI models to predict cholecystitis severity based on the assessments of a single individual [17, 18]. Naturally, we posited that the diagnostic distribution of a panel of experts could also be modeled by computer vision-based AI. Thus, in this study, we measured inter-expert agreement for the assessment of cholecystitis severity and tested the capabilities of an AI model to model an expert diagnostic panel. Then, we aimed to evaluate AI model performance for predicting the expert panel's consensus and the distribution of experts' independent assessments. Finally, we aimed to determine the relative contribution of specific visual objects for making accurate predictions.

Materials and methods

This study was approved by the Institutional Review Boards of the Stanford University School of Medicine (protocol number 57677) and Intermountain Health (protocol number 2020440). A waiver of informed consent was provided due to minimal risk to subjects. We followed the Standards for the Reporting of Diagnostic Accuracy Studies (STARD) 2015 and the Minimum Information about Clinical Artificial Intelligence Modeling (MI-CLAIM) checklists [19, 20].

Data collection

We assembled a case series of laparoscopic cholecystectomy videos from Intermountain Health and two publicly available sources. Intermountain Health is an integrated network of 33 hospitals located in the western United States. Twenty-six Intermountain surgeons at five facilities recorded cases using standard laparoscopy equipment between July and November of 2021. *Cholec80* is a publicly available dataset of 80 laparoscopic cholecystectomy videos recorded by 13 surgeons at the University Hospital of Strasbourg in France [21]. YouTube is an open video sharing platform and a common source of medical videos for building AI models [22, 23]. We collected laparoscopic cholecystectomy videos performed by over 54 surgeons and 22 institutions across the globe on YouTube. All videos were deidentified using FFmpeg software and stored on a secure central server [24]. Members of the research team followed the Parkland Grading Scale protocol, extracting a single frame from each video. The chosen frame depicted the right upper abdominal structures after placement of all four laparoscopic ports. If the gallbladder was visualized easily, we selected a frame after grasping and cephalad retraction of the gallbladder but prior to dissection of the cystic pedicle, in line with the PGS methodology. In cases of severe inflammation that impeded mobilization or gallbladder visibility, a frame of the inflamed area was chosen [8]. All videos from the three sources were manually reviewed by the research team, and any cases that did not contain a gradable view for PGS-based severity assessment were excluded (main reason was due to the video starting after the time of PGS definition).

Cholecystitis severity grading

Three clinical experts (DA, ES, and JJ) assessed cholecystitis severity for all cases using the initial view image. Their annotations were used for training the computer vision model. One surgical resident (GO) and three computer scientists (EG, JA, and JY) provided additional PGS labels for all cases in the same manner, and these labels were used to benchmark the AI model against those without extensive clinical backgrounds. Only the three clinical experts labels were used to create a final consensus PGS label for each case. The clinical experts were practicing surgeons with fellowship training in minimally invasive and bariatric surgery (DA), general surgical oncology (ES), and acute care and trauma surgery (JJ). All graders were trained to assess cholecystitis severity using the

Parkland Grading Scale [8]. Then, a diagnostic panel comprised of the three clinical experts (DA, ED, and JJ) collectively determined a final consensus grade for every case during a plenary discussion [11].

Measuring inter-expert agreement

We measured inter-expert agreement for cholecystitis severity grading among the three surgeons on the diagnostic panel. We determined the proportion of cases for which all three experts independently agreed, two experts agreed, or none of the experts agreed. We also assessed inter-expert agreement for clinical expert pairs using absolute agreement and the weighted Cohen's kappa statistic [25]. The weighted Cohen's kappa statistic is a measurement of inter-rater reliability that considers both absolute agreement and partial agreement with an increasing penalty as the distance between assessments widens. All statistical calculations were performed in R version 1.3.1073 using base R functions and the *irr* package [26, 27].

Artificial intelligence model development

We created two AI models for the automated grading of cholecystitis severity using the Parkland Grading Scale. Model A was trained to predict the expert diagnostic panel's consensus grade. Model B was trained to additionally predict the distribution of the clinical experts' independent assessments. Each model assigned grades to match the predicted probability distribution. For Model A, the distribution was based on the consensus prediction task. For Model B, the distribution was based on the secondary task of predicting the distribution of the clinical experts' independent assessments.

We used a transformer-based neural network to construct our models [28]. Transformers are state-of-the-art computer algorithms that were originally used for natural language processing but have recently been adapted for image processing [29]. Transformers constitute the latest generation of AI algorithms and have been shown to outperform convolutional neural networks for analyzing medical images [30]. Both transformer models were pre-trained using the large-scale *ImageNet* visual database [31, 32]. The laparoscopic cholecystectomy cases were divided into training, validation, and test subsets using a 60:20:20 distribution (191/64/64 images). We used clinical experts' independent grades and the panel consensus grades as the reference standard for model training. The vision transformers were trained using an Adam optimizer with a learning rate of 0.0001 and a batch size of 64 images. RandAugment was used to create simulated training data [33]. RandAugment is a technique that applies randomly selected image transformations, such as rotation, re-coloring, sharpening, or brightening, enabling a model to learn robustness to lighting changes and other spurious visual artifacts unrelated to the primary area of interest (e.g., the color of the gallbladder body). This helps ensure our model is focused on salient features of gallbladder inflammation. The number of sequential augmentations (N) was five, and the magnitude of transforms (M) was five. Each model was trained with early stopping monitoring the total model loss and stopped when total loss had not decreased for 20 epochs.

The best model throughout training, as evaluated by accuracy on the validation set, was chosen. Two objective functions were used for model optimization. A cross-entropy loss was used for panel consensus prediction, and a Kullback–Leibler divergence loss was used for predicting the distribution of the clinical experts' independent grades. Cross-entropy loss is a measure of how well the model's predictions match the ground-truth distribution. In practical terms, for each case, the model produces a probability for every possible diagnosis or category (e.g., inflammation grades 1–5). Cross-entropy loss increases when the model assigns low probability to the correct rating and decreases when the model assigns high probability to the correct rating. A lower cross-entropy loss indicates that the model is making predictions more confidently and accurately. During training, the model adjusts itself to minimize this loss, gradually learning to assign higher probability to the correct diagnosis or label for each example. The KL divergence loss is also one of the most commonly used objectives in deep learning and we employ this loss to compare the predicted distribution of a panel diagnosis to the ground-truth distribution given by our expert annotators. For Model A, only the consensus cross-entropy loss was optimized. For Model B, the weighted sum of both losses was optimized, with a 0.8 coefficient on the Kullback–Leibler divergence loss and a 0.2 coefficient on the panel consensus prediction loss. These weights were chosen to sum to one and chosen via hyperparameter optimization. The models were trained using PyTorch and python on RTX 3090 graphic processing units [34]. Figure 1 displays the model architecture and losses.

Performance evaluation

We measured individual grader accuracy by comparing independent assessments with the panel consensus reference standard for all 319 cases using absolute agreement and the weighted Cohen's kappa statistic. We measured model accuracy by comparing the predicted cholecystitis severity grade with the panel consensus using absolute agreement and the weighted Cohen's kappa statistic for the test cases unseen by the model throughout training. We also measured model accuracy by comparing the predicted distribution of experts' independent assessments with the actual distribution of assessments using absolute agreement.

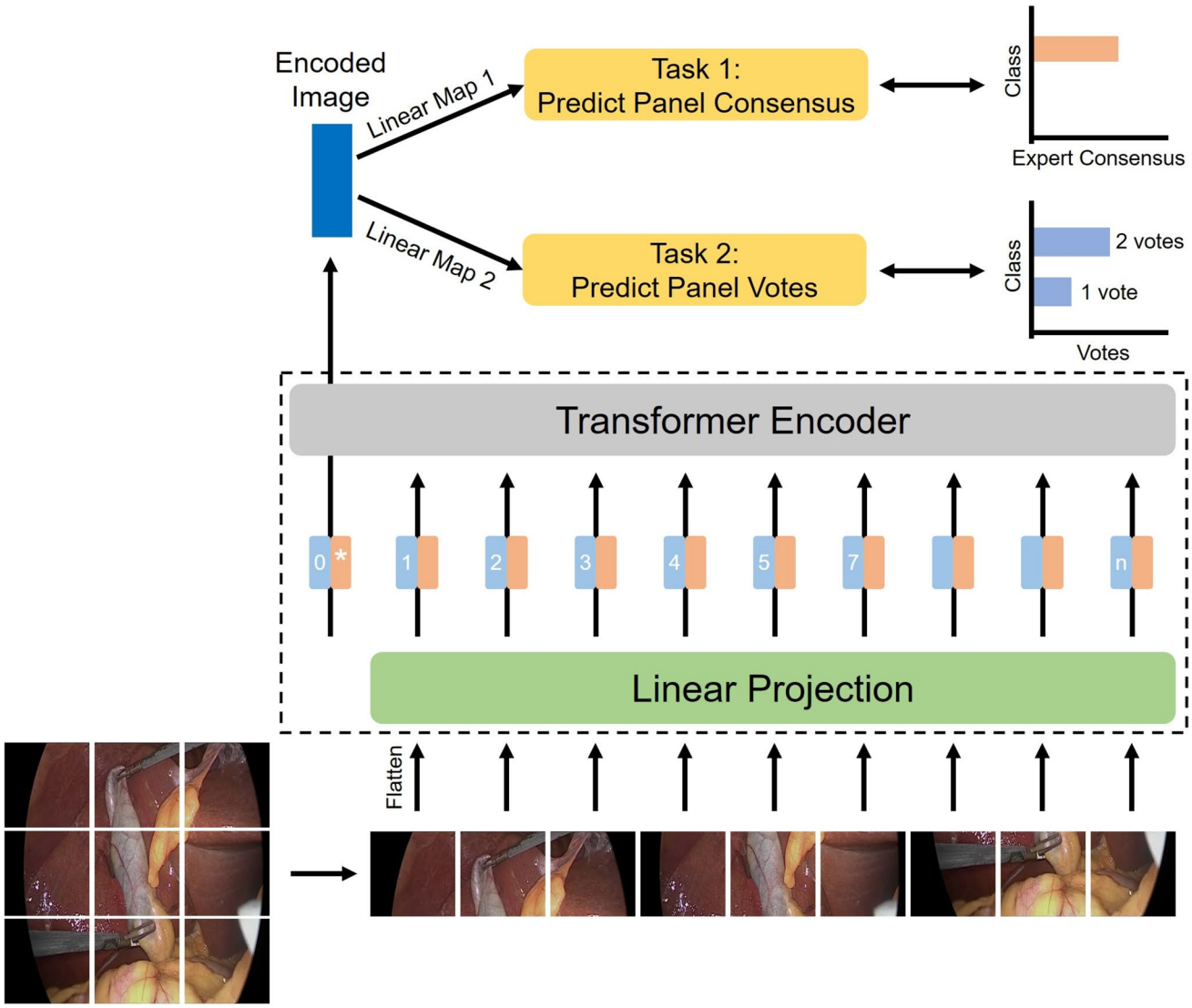

Fig. 1 Artificial intelligence models for grading of cholecystitis severity using intraoperative images

Model interpretability

We evaluated the interpretability of the model using occlusion experiments. Occlusion experiments use segmentation masks to obscure objects with black pixels to show which visual elements within the images are most critical for making correct predictions. Members of the research team created segmentation masks for key anatomic structures and instruments within the 64-image test set, including the gallbladder, the liver, the abdominal fat/omentum, the abdominal wall, the gastrointestinal tract, graspers, and the dark image background. We then calculated the proportion of cases for which masking an object changed the model's prediction from correct to incorrect. Annotations were performed using hasty.ai, an online annotation platform [35].

Results

We compiled a dataset of 319 laparoscopic cholecystectomy videos: 76 from *Cholec80,* 86 from Intermountain Healthcare, and 160 from YouTube. In this study, YouTube videos were used to supplement training of our PGS classifier, employing a weighted sampler to ensure balanced exposure to all PGS levels during model development. This approach enables our models to generalize across the full spectrum of disease severity and learn salient features of inflamed gallbladders, rather than being limited by the frequency distribution seen in typical clinical practice. The distribution of the graders' independent assessments of cholecystitis severity and the expert diagnostic panel's consensus are shown in Fig. 2.

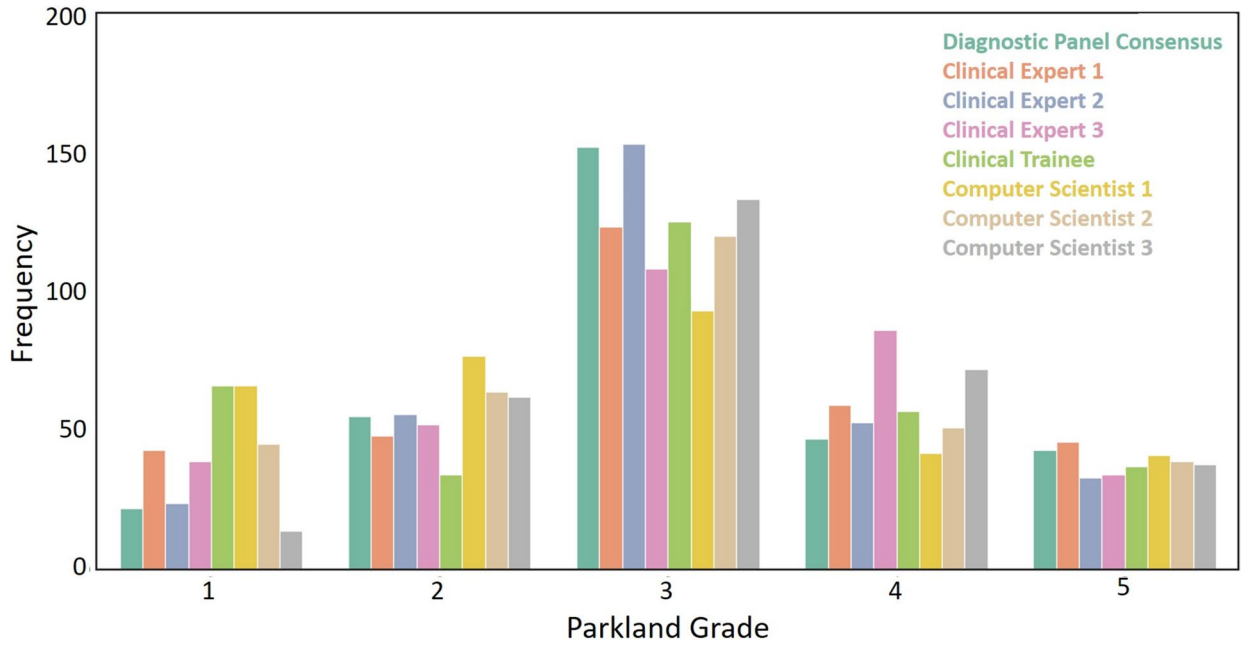

Fig. 2 The distribution of graders' independent cholecystitis severity assessments and the expert diagnostic panel's consensus

Table 1 Measurements of inter-expert agreement for the assessment of cholecystitis severity for clinical expert pairs

Expert pair	Absolute agreement no. (%)	Weighted Cohen's kappa
Clinical expert 1 vs. Clinical expert 2	210/319 (66%)	0.78
Clinical expert 1 vs. Clinical expert 3	215/319 (67%)	0.76
Clinical expert 2 vs. Clinical expert 3	206/319 (65%)	0.83

Inter-expert agreement

All three clinical experts independently assigned identical cholecystitis severity grades for 163 of 319 cases (51%). Two of the three experts agreed for 142 of 319 cases (45%), and all three experts disagreed for 14 of 319 cases (4%). Absolute agreement and weighted Cohen's kappa statistics for clinical expert pairs are shown in Table 1.

Performance evaluation

The median accuracy of the individual graders using absolute agreement with the expert panel's consensus was 72% (range: 60–79%). The median grader accuracy using the weighted Cohen's kappa statistic considering proximity to the panel's consensus was 0.83 (range: 0.73–0.90). Individual grader accuracy stratified according to the level of clinical training with clinicians performing slightly better than the computer scientists.

The accuracy of Model A using absolute agreement with the expert panel's consensus was 69%, and the weighted Cohen's kappa statistic was 0.62. On average, Model A, which was not explicitly trained to predict the distribution of the clinical experts' independent assessments, correctly predicted 1 of 3 independent assessments compared with the reference standard distribution. The accuracy of Model B using absolute agreement with the panel's consensus was 72%, and the weighted Cohen's kappa statistic was 0.77. Model B, which was trained to predict the distribution of the clinical experts' independent assessments, correctly predicted 2 of 3 independent assessments compared with the reference standard distribution. Performance data for members of the research team and the AI models are displayed in Table 2.

Model interpretability

Masking the gallbladder in the test set images resulted in a change in Model B's prediction for 14 of 64 cases (22%). Masking the liver or abdominal fat/omentum resulted in a change in Model B's prediction for 11 of 64 cases (17%) and 10 of 64 cases (15%), respectively. In contrast, masking the gastrointestinal tract, graspers, or the black border around the image led to misclassification in only 2 of 64 cases (4%). Results of the object occlusion experiments are shown in Fig. 3.

Table 2 Performance evaluation for individual graders and the artificial intelligence models

Rater	No. (%) absolute agreement with diagnostic panel consensus	Weighted Cohen's kappa statistic for agreement with diagnostic panel consensus
Clinical expert 2	253/319 (79%)	0.83
Clinical expert 1	251/319 (79%)	0.90
Clinical expert 3	232/319 (73%)	0.85
Clinical trainee	230/319 (72%)	0.84
Computer scientist 3	209/319 (68%)	0.73
Computer scientist 2	216/319 (65%)	0.78
Computer scientist 1	209/319 (60%)	0.79
Model A[a]	44/64 (69%)	0.62
Model B[b]	46/64 (72%)	0.77

[a]Predicts diagnostic panel consensus only

[b]Predicts diagnostic panel consensus and independent expert ratings

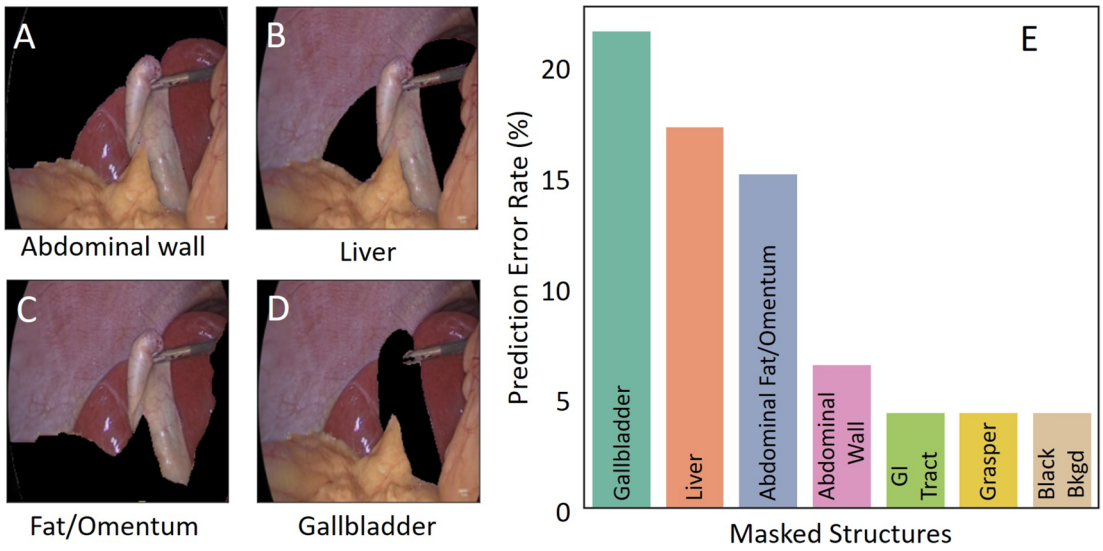

Fig. 3 Results of occlusion experiments for model interpretability. Examples of masking **A** the abdominal wall, **B** the liver, **C** the abdominal fat/omentum, and **D** the gallbladder. **E** Percentage of miss-classified cholecystitis severity grades by masked object

Discussion

In this study, we found inter-expert agreement for the diagnosis of cholecystitis severity using the Parkland Grading Scale, and a single intraoperative image was imperfect. We developed two transformer-based AI models to predict the consensus of an expert diagnostic panel, and one of the models also predicted the distribution of the panel members' independent assessments. The accuracy of our best model was comparable to experienced clinicians when comparing individual cholecystitis severity assessments with the expert panel reference standard. With the occlusion experiments, we also showed the model relied on the appearance of key anatomic structures included in the Parkland Grading Scale criteria to make accurate predictions. To our knowledge, this is the first attempt to use AI to model an expert diagnostic panel to automate the assessment of cholecystitis severity using the Parkland Grading Scale. In addition, this is the first use of a state-of-the-art transformer-based neural network architecture for this task.

The creators of the Parkland Grading Scale reported excellent inter-rater reliability when considering the proximity of ratings along the five-tiered scale using the intra-class correlation coefficient [8]. We also observed reasonable inter-expert agreement using the weighted Cohen's kappa statistic. However, we found all three clinical experts independently agreed for only half of the cases. This underscores the inherent subjectivity of assessing cholecystitis severity

using the Parkland Grading Scale. During the expert panel's plenary discussion, we noted the surgeons tended to agree on the appearance of visual indications of inflammation within an image. However, they did not always agree on how those indicators fit within the Parkland Grading Scale. For example, the level of omental adhesions to the gallbladder is one criterion included in the scale. The definition for grade three is "adhesions to the body," while the definition for grade 4 is "adhesions obscuring the majority of the gallbladder." [8] Traditionally, the gallbladder is anatomically divided into thirds: the neck, the body, and the fundus. Experts debated whether the threshold for the majority of the gallbladder surface area should be greater than one half, which would include the body, or greater than two thirds, which would require adhesions to reach the fundus. As a result, experts felt many cases fell somewhere between the two grades and had to clarify the definition to reach a consensus.

The subjectivity of assessing cholecystitis severity using the Parkland Grading Scale presents a challenge when trying to establish a reference standard for training and evaluating AI models. Previous work to develop AI models for predicting cholecystitis severity using the Parkland Grading Scale relied on the assessments of a single individual [17]. Combining expertise to produce an expert panel diagnosis may reduce variation due to subjectivity [11–14]. In our study, using the ratings of a single expert to train the AI models would have inherently limited model performance when comparing predictions with the panel's consensus. If we consider inter-expert consensus to achieve Bayes error rate, defined as the lowest conceivable error rate given the distribution of the data, any single expert would provide reference standard grades with only 79% accuracy at best.

Despite the inherent challenges of training an AI model to grade cholecystitis severity using qualitative criteria based on clinical judgment, we were able to build a model with prediction accuracy comparable to trained clinicians. It is important to note that the AI model described in this study is not intended for direct clinical decision support but rather serves as an initial demonstration of technical feasibility and as a foundation for future development of clinically robust systems that explicitly account for inter-expert variability in grading disease severity. Previous research used convolutional neural networks to model the prediction of cholecystitis severity using the Parkland Grading Scale [17, 18, 36]. We used a state-of-the-art transformer architecture, which has been shown to outperform convolutional neural networks for medical imaging analysis [30]. We also found the use of individual experts' independent assessments to train the model slightly improved the accuracy of Model B over Model A. The difference in these models was that Model B was given access to the expert consensus as well as the individual pre-consensus expert grades. This latter information likely provides useful auxiliary information for understanding the nuance between different cholecystitis severity grades. This was reflected in the weighted Cohen's kappa statistics for the models, which consider both absolute agreement and the closeness of agreement between predictions and reference labels. Moving forward, this predicted distribution of individual experts' independent assessments could provide an estimate of the model's degree of confidence.

Finally, we demonstrated that anatomy, and most importantly the appearance of the gallbladder itself, played a key role in Model B's grading of cholecystitis severity. In addition, two other key structures, the liver and the omentum, had a notable impact on performance. These three anatomic structures feature prominently in the Parkland Grading Scale criteria. Interestingly, even when the gallbladder was masked, the model incorrectly predicted cholecystitis severity only 22% of the time. This suggests multiple visual elements and the spatial relationships between them work together to provide information about cholecystitis severity. Previous research has demonstrated the interpretability of AI models by incorporating qualitative criteria from the Parkland Grading Scale within the prediction tasks [36]. We offer an additional method for evaluating interpretability based on segmentation masking that could shed new light on the relationships between specific anatomic structures and surgical instruments when assessing cholecystitis severity.

Limitations to the study include reliance of the Parkland Grading Scale on individual clinical judgment, which affects the reliability of the scale. We assembled an expert diagnostic panel to minimize bias related to variations in individual tendencies. However, we found that limitations inherent in the scale still presented difficulties for the panel members. While our dataset included a variety of cases performed by multiple surgeons from institutions across the globe, the true distribution of cholecystitis severity among the general population of patients who undergo laparoscopic cholecystectomy remains unknown. If the distribution of case severity for a particular institution varies dramatically from that used in our study, model retraining may be needed prior to direct application. Finally, although the Parkland Scale has been clinically validated, clinical indicators of disease severity were not available for graders to review and were not included in the model [9]. Clinical data could be useful for strengthening the validity of disease severity assessments and model predictions.

The creators of the Parkland Grading Scale aimed to develop a simple, reliable system for classifying disease severity and operative difficulty during laparoscopic cholecystectomy [8–10]. Ease of use is one strength of the Parkland Scale, and we observed that non-clinician computer scientists could be trained to perform disease severity assessment with accuracy not much below medical professionals. However, we also noted inherent limitations in the

qualitative criteria of the scale. AI can perform rapid calculations beyond human ability. Quantification could allow for a more nuanced characterization of cholecystitis severity. For example, AI could specify the density of omental adhesions in relation to gallbladder surface area along a continuous scale. We imagine similar possibilities for other aspects of disease severity, including the degree of hyperemia and edema. Future work on the automated assessment of cholecystitis severity should focus on developing a grading system that leverages the computational strength of AI while encompassing the needs of practicing surgeons. While our study examined how modeling consensus among experts can improve the robustness of AI models for disease severity assessment, future work should incorporate strategies from active learning and uncertainty quantification to better characterize clinician-provided ground-truth labels with confidence scores. For example, follow-up studies could explore weighting labels from experts according to their level of certainty, particularly for subjective grading systems like PGS, enabling the model to calibrate its reliance on human annotations during training.

Distinctively, in comparison to prior work, we relied on consensus of a panel of expert clinicians to build an accurate transformer-based AI model to predict ratings of cholecystitis severity derived from the Parking Grading scale. This model also predicted the individual assessments of the clinical experts who provided this consensus. Interestingly, our work showed the variance and subjectivity of PGS even among experienced clinicians and illustrates the limitations of the Parking Grading Scale as a ground-truth for computer-vision-based models. Our findings highlight the potential of AI as a robust evaluative tool; however, for it to serve effectively in clinical decision-making, there is a need for rating scales tailored for AI comprehension. Directions for future research should shift to developing methods that take advantage of the computational strengths of AI to produce a more nuanced characterization of disease severity that transcends human capabilities.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00464-025-12015-6.

Acknowledgements Dr. Olsen and Dr. Goodman contributed equally as the co-first authors. Dr. Azagury and Dr. Yeung contributed equally as co-senior authors. All the authors (GO, EG, JA, SB, KH, JY, ES, SY, JJ, and DA) made substantial contributions to the design of the study, data acquisition, data analysis, and drafting and revision of the manuscript. All the authors approve this manuscript and agree to be accountable for all aspects of the work.

Funding This work was supported by Stanford Medicine and Intermountain Health through a collaboration grant; the National Library of Medicine of the National Institutes of Health (grant T15LM007033); and the National Science Foundation Future of Work Investment (award number 2026498). Dr. Olsen wishes to thank the Stanford-Intermountain Fellowship in Population Health, Delivery Science, and Primary Care. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Declarations

Disclosures Dr. Kay Hung was an intern at Intuitive Surgical, Inc during the submission of this study. Drs. Griffin Olsen, Emmett Goodman, Sebastiano Bartoletti, Eric Sorenson, Jeffrey Jopling, Serena Yeung, Dan Azagury, Mr. Josiah Aklilu and Ms. Janice Yang have no conflicts of interest or financial ties to disclose. This work was supported by Stanford Medicine and Intermountain Health through a collaboration grant; the National Library of Medicine of the National Institutes of Health (grant T15LM007033); and the National Science Foundation Future of Work Investment (award number 2026498). Dr. Olsen wishes to thank the Stanford-Intermountain Fellowship in Population Health, Delivery Science, and Primary Care. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

References

1. Friedman GD (1993) Natural history of asymptomatic and symptomatic gallstones. Am J Surg 165(4):399–404. https://doi.org/10.1016/s0002-9610(05)80930-4
2. Stinton LM, Shaffer EA (2012) Epidemiology of gallbladder disease: cholelithiasis and cancer. Gut Liver 6(2):172–187. https://doi.org/10.5009/gnl.2012.6.2.172
3. Ansaloni L, Pisano M, Coccolini F et al (2016) 2016 WSES guidelines on acute calculous cholecystitis. World J Emerg Surg WJES 11:25. https://doi.org/10.1186/s13017-016-0082-5
4. Wakabayashi G, Iwashita Y, Hibi T et al (2018) Tokyo Guidelines 2018: surgical management of acute cholecystitis: safe steps in laparoscopic cholecystectomy for acute cholecystitis (with videos). J Hepato-Biliary-Pancreat Sci 25(1):73–86. https://doi.org/10.1002/jhbp.517
5. McDermott KW, Liang L (2018) Overview of operating room procedures during inpatient stays in US hospitals
6. Tominaga GT, Staudenmayer KL, Shafi S et al (2016) The American association for the surgery of trauma grading scale for 16 emergency general surgery conditions: disease-specific criteria characterizing anatomic severity grading. J Trauma Acute Care Surg 81(3):593–602. https://doi.org/10.1097/TA.0000000000001127

7. Yokoe M, Hata J, Takada T et al (2018) Tokyo Guidelines 2018: diagnostic criteria and severity grading of acute cholecystitis (with videos). J Hepato-Biliary-Pancreat Sci 25(1):41–54. https://doi.org/10.1002/jhbp.515
8. Madni TD, Leshikar DE, Minshall CT et al (2018) The Parkland grading scale for cholecystitis. Am J Surg 215(4):625–630. https://doi.org/10.1016/j.amjsurg.2017.05.017
9. Madni TD, Nakonezny PA, Barrios E et al (2019) Prospective validation of the parkland grading scale for cholecystitis. Am J Surg 217(1):90–97
10. Madni TD, Nakonezny PA, Imran JB et al (2019) A comparison of cholecystitis grading scales. J Trauma Acute Care Surg 86(3):471–478
11. Bertens LCM, Broekhuizen BDL, Naaktgeboren CA et al (2013) Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLOS Med 10(10):e1001531. https://doi.org/10.1371/journal.pmed.1001531
12. Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT (2007) Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. Arch Ophthalmol 125(7):875–880. https://doi.org/10.1001/archopht.125.7.875
13. Ryan MC, Ostmo S, Jonas K et al (2014) Development and evaluation of reference standards for image-based telemedicine diagnosis and clinical research studies in ophthalmology. AMIA Annu Symp Proc 2014:1902–1910
14. Brown JM, Campbell JP, Beers A et al (2018) Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA Ophthalmol 136(7):803–810. https://doi.org/10.1001/jamaophthalmol.2018.1934
15. Hashimoto DA, Rosman G, Rus D, Meireles OR (2018) Artificial intelligence in surgery: promises and perils. Ann Surg 268(1):70–76. https://doi.org/10.1097/SLA.0000000000002693
16. Ward TM, Mascagni P, Ban Y et al (2021) Computer vision in surgery. Surgery 169(5):1253–1256. https://doi.org/10.1016/j.surg.2020.10.039
17. Ward TM, Hashimoto DA, Ban Y, Rosman G, Meireles OR (2022) Artificial intelligence prediction of cholecystectomy operative course from automated identification of gallbladder inflammation. Surg Endosc. https://doi.org/10.1007/s00464-022-09009-z
18. Korndorffer JR, Hawn MT, Spain DA et al (2020) Situating artificial intelligence in surgery: a focus on disease severity. Ann Surg 272(3):523–528. https://doi.org/10.1097/SLA.0000000000004207
19. Bossuyt PM, Reitsma JB, Bruns DE et al (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 351(h5527):5
20. Norgeot B, Quer G, Beaulieu-Jones BK et al (2020) Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 26(9):1320–1324. https://doi.org/10.1038/s41591-020-1041-y
21. Twinanda AP, Shehata S, Mutter D, Marescaux J, De Mathelin M, Padoy N (2016) EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imag 36(1):86–97
22. Goodman ED, Patel KK, Zhang Y, Locke W, Kennedy CJ, Mehrotra R, Ren S, Guan MY, Downing M, Chen HW, Clark JZ (2021) A real-time spatiotemporal AI model analyzes skill in open surgical videos. arXiv preprint arXiv:2112.07219
23. Google LLC (2021) YouTube. Accessed November 1, https://www.youtube.com
24. Bellard F (2022) About FFmpeg. FFmpeg. Accessed January 29, https://ffmpeg.org/about.html
25. Cohen J (1968) Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70(4):213–220. https://doi.org/10.1037/h0026256
26. R Core Team (2020). R: A language and environment for statistical computing. Published online 2020. https://www.R-project.org/
27. Gamer M, Lemon J, Fellows I, Singh P (2019) irr: Various coefficients of interrater reliability and agreement. https://www.r-project.org
28. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. arXiv preprint. https://doi.org/10.48550/arXiv.1706.03762
29. Dosovitskiy A, Beyer L, Kolesnikov A et al (2021) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint. https://doi.org/10.48550/arXiv.2010.11929
30. Matsoukas C, Haslum JF, Söderberg M, Smith K (2021) Is it time to replace CNNs with transformers for medical images? arXiv preprint. http://arxiv.org/abs/2108.09038. Accessed January 29, 2022
31. Bao H, Dong L, Wei F (2021) BEiT: BERT pre-training of image transformers. arXiv preprint. http://arxiv.org/abs/2106.08254. Accessed February 4, 2022
32. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
33. Cubuk ED, Zoph B, Shlens J, Le QV (2020) Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops
34. Paszke A, Gross S, Massa F, et al (2019) PyTorch: an imperative style, high-performance deep learning library. Adv Neural Inf Proc Syst
35. Hasty GmbH (2022) About us. Hasty.ai. https://hasty.ai/about-us Accessed June 8, 2022
36. Ban Y, Eckhoff JA, Ward TM, Hashimoto DA, Meireles OR, Rus D, Rosman G (2023) Concept graph neural networks for surgical video understanding. IEEE Trans Med Imag 43(1):264–274