Medical Image Analysis 115 (2027) 104290

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Federated Learning for surgical vision in appendicitis classification: Results of the FedSurg EndoVis 2024 challenge

Max Kirchner [a,*], Hanna Hoffmann [a], Alexander C. Jenke [a], Oliver L. Saldanha [d,f], Kevin Pfeiffer [f], Weam Kanjo [s], Julia Alekseenko [g,h], Claas de Boer [a], Santhi Raj Kolamuri [i], Lorenzo Mazza [a], Nicolas Padoy [g,h], Sophia Bano [l], Annika Reinke [m,n], Lena Maier-Hein [m,n,o,p,q,r], Danail Stoyanov [k,l], Jakob N. Kather [c,d,f], Fiona R. Kolbinger [e,j], Sebastian Bodenstedt [a,b], Stefanie Speidel [a,b]

[a] Department of Translational Surgical Oncology, National Center for Tumor Diseases (NCT), NCT/UCC Dresden, a partnership between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, and Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany
[b] Centre for Tactile Internet with Human-in-the-Loop (CeTI), TUD Dresden University of Technology, Dresden, Germany
[c] Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany
[d] Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany
[e] Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA
[f] Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany
[g] IHU Strasbourg, Institute of Image-Guided Surgery, Strasbourg, France
[h] University of Strasbourg, CNRS, INSERM, ICube, UMR7357, Strasbourg, France
[i] Dr. NTR University of Health Sciences, Vijayawada, India
[j] Department of Visceral, Thoracic and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany
[k] Digital Technologies, Medtronic, London, United Kingdom
[l] UCL Hawkes Institute and Department of Computer Science, University College London, London, UK
[m] German Cancer Research Center (DKFZ) Heidelberg, Division of Intelligent Medical Systems, Heidelberg, Germany
[n] German Cancer Research Center (DKFZ) Heidelberg, Helmholtz Imaging, Heidelberg, Germany
[o] National Center for Tumor Diseases (NCT) Heidelberg, a partnership between DKFZ and Heidelberg University Hospital, Heidelberg, Germany
[p] Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany
[q] Helmholtz Information and Data Science School for Health, Heidelberg, Karlsruhe, Germany
[r] Heidelberg University Hospital, Surgical Clinic, Surgical AI Research Group, Heidelberg, Germany
[s] Krankenhaus St. Joseph-Stift Dresden GmbH, Dresden, Germany

ARTICLE INFO

Dataset link: https://doi.org/10.25532/OPARA-1173, https://doi.org/10.25532/OPARA-1534

Keywords:
Federated Learning
EndoVis challenge
Appendectomy
Video classification
Surgical data science

ABSTRACT

Developing generalizable surgical AI requires multi-institutional data, yet privacy constraints preclude direct data sharing, making Federated Learning (FL) a natural candidate. Its application to complex, spatiotemporal surgical video remains largely unbenchmarked. We present the FedSurg Challenge, the first international initiative dedicated to FL in surgical vision, as a proof-of-concept evaluation using a multi-center dataset of laparoscopic appendectomies (preliminary subset of Appendix300). Three participant submissions were evaluated on generalization to an unseen clinical center and center-specific local adaptation, alongside centralized, Swarm Learning and parameter-efficient fine-tuning baselines, and constant and chance-level reference classifiers. Our analysis identifies temporal modeling as the architectural factor most consistently associated with generalization to the unseen center, although the effect is not uniform across metrics in the controlled baseline comparison, and both parameter-efficient baselines remain more costly than a constant predictor under the ordinal metric. Classifier collapse arises both from failure of the global model to transfer under domain shift and from unconstrained fine-tuning on small, imbalanced local datasets, motivating structured personalized FL with parameter-efficient fine-tuning for center-specific adaptation. Absolute performance remains far from clinical viability: even with all data pooled centrally the task reached 26.31% F1-score on the unseen center, and no clinically acceptable threshold for intraoperative appendicitis grading has been established. Paired

* Corresponding author.
E-mail address: max.kirchner@nct-dresden.de (M. Kirchner).

https://doi.org/10.1016/j.media.2026.104290
Received 6 October 2025; Received in revised form 17 August 2026; Accepted 22 August 2026
Available online 7 September 2026
1361-8415/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

permutation tests over the test cases resolve only large differences, and no adaptation comparison reaches significance at this sample size. By characterizing these limitations, and the limits of what the evaluation can resolve at this scale, this work establishes a methodological reference point for privacy-preserving surgical video AI.

1. Introduction

The combination of early successes in AI algorithms and the emerging field of Surgical Data Science (SDS) holds strong potential to transform surgery (Maier-Hein et al., 2017). Recent work has demonstrated that AI can reliably analyze surgical video, comprehending anatomy, tool usage, and procedural events in real time (Brandenburg et al., 2023). Such capabilities are clinically relevant given the established relationship between video-derived quality indicators and postoperative complications. However, the development of robust AI systems critically depends on access to large volumes of high-quality, diverse surgical data. This requirement remains one of the field's most pressing challenges (Maier-Hein et al., 2022).

Surgical datasets are typically sourced from either a single institution or a small group of collaborators (Maier-Hein et al., 2022; Carstens et al., 2025). While these datasets enable preliminary developments, they inherently suffer from limited diversity and scale if applied in real-world scenarios (Kirtac et al., 2022). Single-center datasets often lack the heterogeneity required for generalizable AI models and involve prolonged data acquisition cycles (Lavanchy et al., 2023). Multi-institutional data aggregation, on the other hand, can significantly enrich the dataset, but is frequently obstructed by stringent regulatory frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) (Rights (OCR), 2021) or the General Data Protection Regulation (GDPR) (GDPR, 2016) that prohibit unrestricted data sharing.

A viable solution is Federated Learning (FL). FL enables decentralized model training across multiple clients, such as hospitals, without the need to share raw image or video data (McMahan et al., 2017). Instead, the model is brought to the data, preserving privacy and complying with regulations like GDPR or HIPAA. In a standard FL workflow, each client trains a model locally on its private dataset. Only model updates are sent to a central server, where they are aggregated using strategies like Federated Averaging (FedAvg) (McMahan et al., 2017) or more robust alternatives such as FedMedian (Yin et al., 2018). The updated global model is then redistributed to clients, and this cycle repeats until convergence.

FL is especially well-suited for medical applications, where data privacy, security, and the lack of standardized datasets pose significant challenges (Rieke et al., 2020). It enables large-scale collaborative model development without the need to centralize sensitive, multi-institutional data, thereby preserving patient confidentiality and regulatory compliance (Rieke et al., 2020). However, FL also introduces new challenges. These include handling data and system heterogeneity across institutions (Li et al., 2020a; Kairouz et al., 2021; Rauniyar et al., 2023), balancing personalization with generalization (Tan et al., 2023), managing communication overhead due to frequent model updates (McMahan et al., 2017; Kairouz et al., 2021), addressing fairness and potential model biases (Kairouz et al., 2021; Li et al., 2020b), and coping with the increased complexity of evaluation and debugging in distributed settings (Kairouz et al., 2021). In the context of SDS, recent studies have begun to explore FL for tasks such as surgical phase recognition, scene segmentation, and tool detection (Kassem et al., 2022; Kirchner et al., 2026; Li et al., 2025).

To address challenges in SDS, initiatives such as the Endoscopic Vision (EndoVis) Challenge (Speidel et al., 2024) have become critical accelerators for progress. The Federated Learning for Surgical Vision (FedSurg) challenge, introduced as part of the 2024 edition of EndoVis, represents the first FL challenge in the field of SDS. The challenge aimed to establish a new classification task, dataset and evaluation protocol for applying FL to surgical AI, enabling comparison of privacy-preserving model development strategies across institutions. We use the term benchmark in the sense of a shared task definition with a fixed dataset split and a specified evaluation and ranking protocol, rather than in the sense of a mature comparison among methods already close to state-of-the-art performance. With three complete submissions, this first edition establishes the former and does not claim the latter; the evaluation accordingly characterizes the limits of what the protocol can currently resolve as well as the relative performance of the submitted methods.

In particular, the challenge focused on evaluating model performance in terms of generalization to unseen centers versus adaptation to individual centers, reflecting two core challenges of FL in real-world surgical applications. The Appendix300 dataset is a multi-institutional appendectomy video dataset (Kolbinger et al., 2026). As the first FL challenge of its kind in SDS, FedSurg pioneers the use of a preliminary version of the Appendix300 dataset as a foundational platform for algorithm development and validation. A separate study using the complete dataset is published (Saldanha et al., 2026). In this paper, we report the challenge design, results, and findings of FedSurg according to the transparent reporting of biomedical image analysis challenges (BIAS) guidelines (Maier-Hein et al., 2020).

The main contributions of this paper are:

- The design and results of FedSurg, the first international challenge in Surgical Data Science dedicated to Federated Learning, which pioneers the use of the preliminary, multi-center Appendix300 dataset.
- The first shared task definition and evaluation protocol for FL on patient-level surgical video classification, with the three submitted methods and four organizer baselines evaluated against constant and chance-level references.
- A rigorous, data-driven analysis of the critical trade-off between model generalization to unseen institutions and personalization to local client data.
- A characterization of the practical limitations and failure modes of FL in surgical AI, contextualized against constant and chance-level reference classifiers and assessed by bootstrap rank stability and paired permutation testing.

2. Challenge design

This section outlines the design of the FedSurg challenge, detailing its organizational structure, the core mission guiding the competition, and the datasets employed for evaluation. Additionally, it describes the assessment methods used to fairly and rigorously evaluate submitted models and FL settings, ensuring robust comparison across the multi-centric surgical video dataset.

2.1. Challenge organization

The FedSurg challenge was a one-time event with a fixed submission deadline, held as part of the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2024 conference in Marrakech, Morocco (October 6–10, 2024). Organized jointly by research groups from the Dresden University of Technology (TUD), Purdue University, and the National Center for Tumor Diseases (NCT) Dresden (see Appendix C), the challenge offered a prize pool of €500, provided by the Horizon Europe NearData project and distributed equally between the two challenge tasks.

The primary mission of the FedSurg challenge was to benchmark FL approaches for surgical video classification using a new multi-institutional dataset. The dataset was collected from four German hospitals under institutional ethical approval (see Appendix F). Only training data was released to participants, while test data remained private and was accessible only to the organizers. The dataset used in the FedSurg challenge is a preliminary subset of the Appendix300 dataset (Kolbinger et al., 2026) (see Section 2.3).

All relevant information, including registration, data access, submission progress tracking, guidelines, and a discussion forum, was made available through the official challenge website on Synapse (see Appendix D). This platform served as the central hub for participant onboarding, communication, and submission management. Participants were required to sign challenge rules before they could participate and get access to the data (see Appendix B). The challenge timeline included registration in April 2024, a feedback-based testing phase during August–September, and the final submission deadline on September 18, 2024 (Fig. 1).

Participants were restricted to the released training data and publicly available resources, including pre-trained models. Challenge-provided data could not be used for pre-training in FL submissions. Members of the organizing institutes were permitted to participate but were not eligible for awards. More information about the challenge is available in the challenge design document (see Appendix E).

Submissions were required in a containerized format (Docker) and were evaluated on a dedicated server equipped with up to 8 NVIDIA V100 GPUs, 56 CPUs, and 756 GB RAM. This standardized hardware environment and containerized execution ensured reproducibility and fair benchmarking under identical conditions (Fig. 1). To support participants, example FL code and evaluation scripts were made publicly available (see Appendix D).

Results were first announced during the MICCAI 2024 Satellite Events and subsequently published on the challenge website, alongside the evaluation framework, rankings, and key performance metrics. All results and analyses from teams with a complete working submission are included in this joint publication, with contributing team members listed as co-authors. Authors were not permitted to publish individual challenge results prior to the release of this paper, ensuring coordinated dissemination and preserving the novelty of the outcomes.

2.2. Challenge mission

The FedSurg challenge focuses on classifying the inflammatory stage of acute appendicitis using laparoscopic appendectomy videos, an important and commonly performed surgical procedure. The primary objective is to provide a comparative benchmark of existing solutions, with a special emphasis on exploring various FL strategies. In particular, the challenge investigates the balance between personalization and generalization under inherent data heterogeneity. For this purpose, the newly created Appendix300 dataset was introduced, with the challenge of utilizing a partial subset of its data. This makes FedSurg not only the first FL challenge in SDS but also the first shared task definition and evaluation protocol for comparing FL approaches on the patient-level task of appendicitis classification.

The challenge cohort comprises human patients, both adults and children of diverse ages and biological sex, who underwent appendectomy for suspected appendicitis at four German medical centers. In general, the challenge cohort is a representative sample of patients undergoing appendectomy for suspected appendicitis. The dataset consists of laparoscopic video recordings captured during these interventions. While no additional patient data was provided for the challenge, the final Appendix300 dataset includes supplementary clinical information (Kolbinger et al., 2026).

This challenge's main contribution is its patient-level prediction task for appendicitis staging, a transferable innovation for AI in surgery. This challenge serves as a proof-of-concept for applying FL to patient-level surgical video analysis, with methodological insights intended to transfer to future tasks with broader clinical relevance, such as intra-operative decision support or quality control across diverse surgical procedures.

Participants were tasked with developing FL algorithms to classify appendicitis stages in laparoscopic videos, structured around two core tasks:

- Task 1, Generalization: Evaluate the model's ability to generalize to unseen centers. Participants train their models on a subset of centers and are evaluated on a held-out center that was not involved in training.
- Task 2, Adaptation: Assess the model's ability to personalize to each center's test data. Here, the same trained model is fine-tuned (or adapted) for each single center of the federated setup and then evaluated independently on each center's test set.

In the context of FL, generalization is the ability of a collaborative global model to perform well on data from new clients, while personalization involves adapting the model to achieve optimal performance for a specific client's unique data. The challenge promotes the creation of privacy-preserving algorithms that can tackle both of these issues, creating models that are both robust and adaptable to diverse surgical settings.

2.3. Challenge dataset

The challenge dataset is a preliminary subset of the Appendix300 collection and comprises frames from 223 full-length recordings of laparoscopic appendectomies. Beyond image frames and laparoscopic grading, the finalized Appendix300 dataset is enriched with detailed histopathological findings and patient anamnesis data (Kolbinger et al., 2026). For this challenge, up to 200 frames were extracted per video using FFmpeg software (Tomar, 2006), capturing a 100-second window sampled at two frames per second around an annotated timestamp (Kolbinger et al., 2026). Where the available footage around the timestamp was shorter than this window, fewer frames were obtained.

Frames were selected at the timestamp identified by the operating surgeon when the Appendix was fully visible prior to dissection (Kolbinger et al., 2026). Inflammation severity, graded from level 0 to 5, was annotated by the operating surgeon (surgery residents with varying years of experience). The annotation protocol (available at Kolbinger et al. (2026)) defines class descriptions and includes illustrative examples based on the definition by Gomes et al. (2012). Furthermore, fine-grained classes 3 A/3B were merged into class 3, and classes 4 A/4B into class 4 in this challenge dataset. An example overview of the data is visible in Fig. 1. By using this annotation protocol, a verbal explanation to the participating surgeons, and a custom graphical user interface of the annotation software, we minimize misclassification. Nonetheless, the reliability of this grading scheme has since been characterized on the complete Appendix300 dataset, of which the challenge data forms a subset. An independent second reading of all recordings yielded a weighted Cohen's kappa of 0.614 against the original annotation, with exact agreement varying from 0.618 at grade 0 to no exact agreement at the most severe (grade 5), and nine percent of cases required a third annotation to resolve disagreements of three or more grades (Kolbinger et al., 2026). These figures postdate the challenge and do not describe the labels used here, which derive from the original annotation alone, but they indicate that ambiguity in this grading task is not only confined to adjacent grades.

The challenge dataset comprises data collected from multiple hospitals across Germany, including both university and community hospitals with varied surgical profiles. While the challenge simulates an FL setup, it reflects a realistic scenario since the data originates from four distinct, real-world centers that have been anonymized to protect institutional privacy. The contributing institutions include:

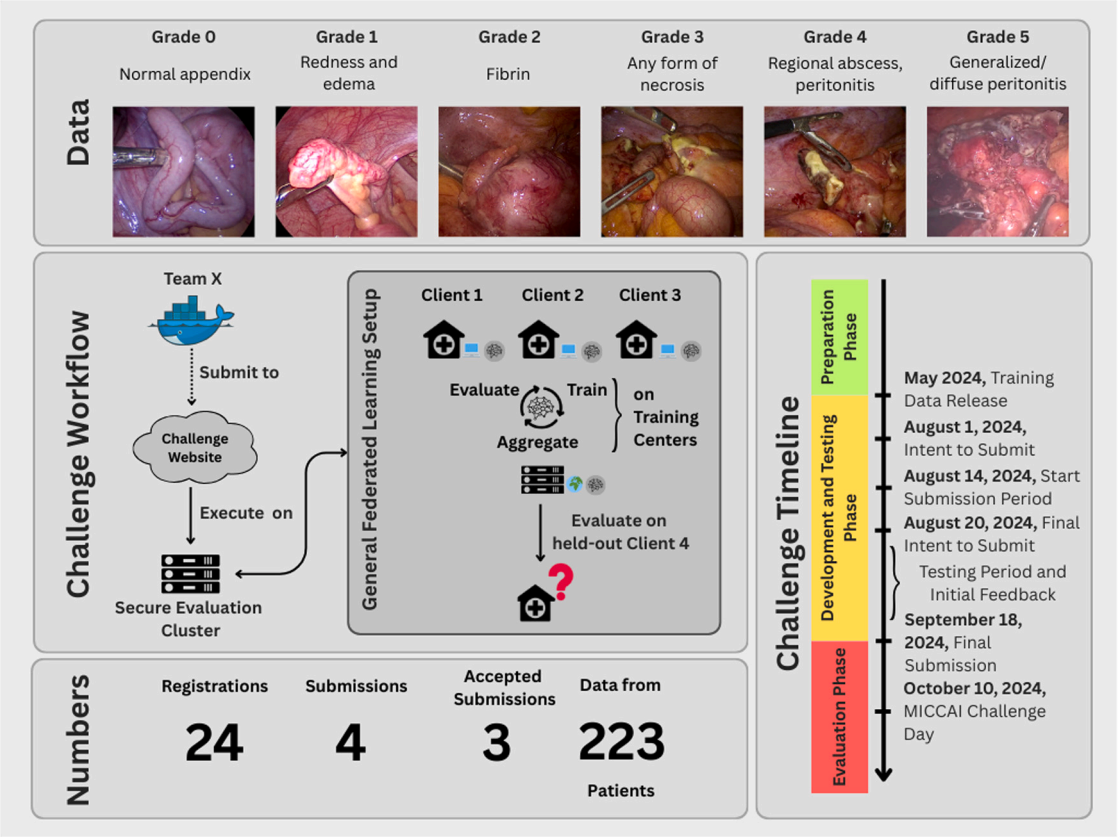


Fig. 1. FedSurg24 challenge highlights: The top panel shows example images of intraoperative appendicitis grades, defined according to Gomes et al. (2012), which were used for video annotation. The lower panel illustrates the FedSurg Challenge workflow: teams submitted Docker containers via Synapse, which were executed on a secure cluster simulating FL across three centers with local training and centralized aggregation. Final performance was assessed by testing each center's best local model on its own test set, while the global model was evaluated on the unseen hold-out center to measure generalization. The challenge timeline with key dates is shown alongside.

Table 1
Data Distribution Across Participating Centers. Video snippets span up to 100 s, centered on the timestamp of full appendiceal visibility prior to surgical preparation; where the available footage around the timestamp is shorter, the snippet is correspondingly shorter.

Center	Training videos	Testing videos	Total
1	40	10	50
2	33	9	42
3	80	22	102
4	0	29	29
Total	153	70	223

- Asklepios-ASB Klinik Radeberg
- University Hospital Carl Gustav Carus Dresden, Department of Pediatric Surgery
- Krankenhaus St. Joseph-Stift, Dresden
- St. Elisabethen-Krankenhaus, Ravensburg

An overview of the video distribution across centers is shown in Table 1.

To discourage centralized model development and preserve the integrity of the federated setting, the training data was partitioned into public and private subsets. Participants were granted access only to the public subset, which constituted 25% of the total training data. The remaining 75% was used exclusively by the organizers during the final training phase, where both public and private data were jointly leveraged in a secure federated setup (Figs. 1 and 2). The class distribution within the dataset mirrors real-world clinical prevalence, featuring a predominance of mid-level cases and a lower representation of extreme or mild inflammation levels.

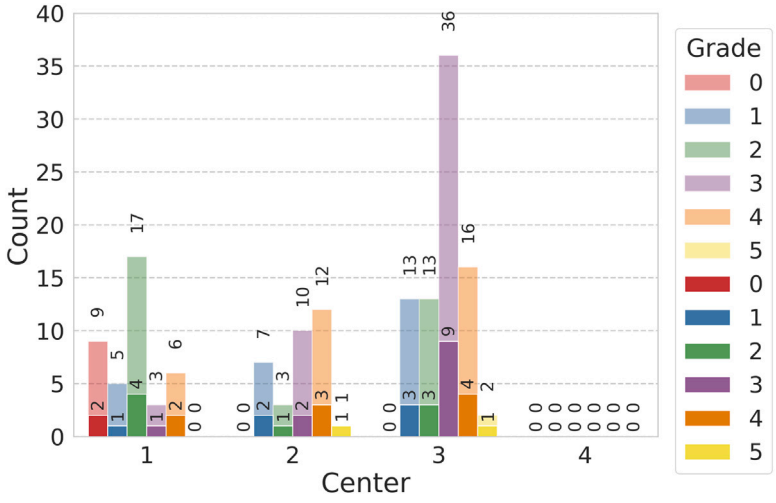
No predefined validation set was provided. Participants were expected to devise their own validation strategies. Final submissions were executed centrally by the organizers, who conducted full federated training and evaluation across all centers under secure and standardized conditions to ensure reproducibility and fairness.

Since the challenge dataset constitutes a preliminary subset of the Appendix300 dataset, we provide a CSV file specifying which samples were used in the challenge and their assignment to centers and to the training and test splits, together with two recordings that were used in the challenge but excluded from the final dataset because the available footage was shorter than the nominal 100 second window or the Appendix was not clearly visible at the annotated timestamp. This material is deposited under the Creative Commons Attribution (CC BY) license.[1] It does not contain the appendectomy recordings themselves, apart from the two excluded cases; the video data and accompanying clinical metadata are available separately as the Appendix300 dataset, also under CC BY.[2] Any use of the challenge dataset requires citing both this paper and the final Appendix300 publication.

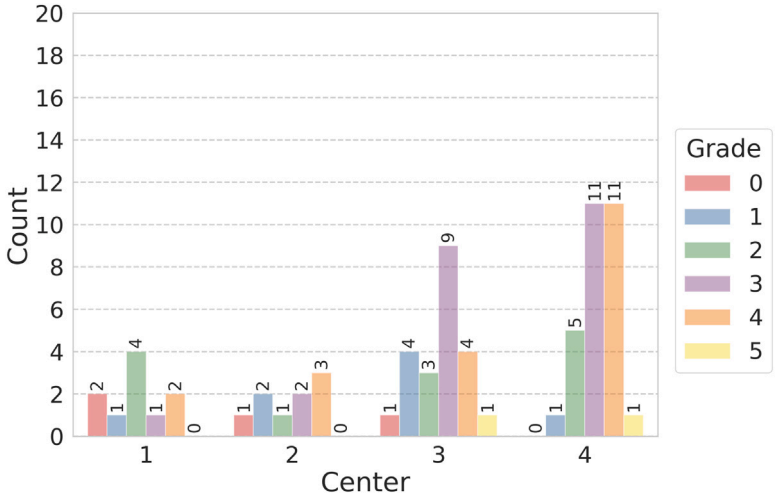
2.4. Assessment methods

The evaluation of this challenge is based on two complementary metrics: the macro-averaged F1-score, also known as the Dice coefficient, and the Expected Cost (EC) with linear weights. Together,

[1] Available at https://doi.org/10.25532/OPARA-1534.
[2] Available at https://doi.org/10.25532/OPARA-1173.



(a) Training dataset distribution.



(b) Testing dataset distribution.

Fig. 2. Label distribution across data subsets per center. Label distributions for (a) the training dataset and (b) the test dataset across the four centers. In the training set visualization, the darker segments represent the publicly available subset for participant development, while the lighter segments show the complete dataset used for final federated training. The plots highlight notable inter-center variability and class imbalance.

these metrics provide a robust assessment of model performance in a multi-class setting with ordinal structure.

F1-score (Dice coefficient)

The F1-score is widely used for balancing precision and recall, offering a harmonic mean that reflects both the accurate identification of relevant instances and the minimization of false positives (Dice, 1945; Maier-Hein et al., 2024). Given a confusion matrix $\mathbf{M} \in \mathbb{N}^{C\times C}$, where $M_{i,j}$ denotes the number of samples with ground-truth class i predicted as class j, the F1-score for class c is computed as:

$$\mathrm{F1}_c = \frac{2 \cdot \mathrm{TP}_c}{2 \cdot \mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c} \tag{1}$$

where:

$$\mathrm{TP}_c = M_{c,c} \quad \text{(true positives)}$$

$$\mathrm{FP}_c = \sum_{\substack{i=1 \\ i\neq c}}^{C} M_{i,c} \quad \text{(false positives)}$$

$$\mathrm{FN}_c = \sum_{\substack{j=1 \\ j\neq c}}^{C} M_{c,j} \quad \text{(false negatives)}$$

The overall F1-score is calculated as the macro-average across all classes:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{F1}_c \tag{2}$$

Expected Cost (EC)

To account for the ordinal nature of the classification task, we also report the Expected Cost (EC), which penalizes misclassifications based on their severity. This aligns with the principle of ordinal monotonicity, where predictions farther from the ground-truth class incur higher penalties (Hastie et al., 2009; Elkan, 2001; Maier-Hein et al., 2024). The EC is defined as:

$$\mathrm{EC} = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{C} M_{i,j} \cdot w_{i,j} \tag{3}$$

where:

$$N = \sum_{i=1}^{C}\sum_{j=1}^{C} M_{i,j} \quad \text{(total number of samples)}$$

$$w_{i,j} = \frac{|i-j|}{C-1} \quad \text{(cost of predicting class } j \text{ when the true class is } i\text{)}$$

with the linear weight function:

$$w_{i,j} = \frac{|i-j|}{C-1} \tag{4}$$

This assigns zero cost to correct predictions and a linearly increasing penalty for deviations, with a maximum cost of 1 for the farthest misclassifications.

2.4.1. Ranking

To ensure a fair evaluation of submissions across both generalization and adaptation scenarios, we implemented a ranking framework that combines multiple performance metrics with an assessment of ranking stability and statistical comparison. Only submissions that successfully completed both tasks were considered. Teams failing to meet this criterion or submitting non-executable code were disqualified. Our ranking methodology guaranteed that models were evaluated not just on average performance but also on their robustness and consistency across heterogeneous data settings. The details of our evaluation procedure are described below.

For Task 1 (generalization), we computed the F1-score and the EC metric for all test cases. Separate rankings were assigned based on each metric, and a team's overall rank for Task 1 was derived by averaging these two ranks. For Task 2 (adaptation), the metrics were first averaged across all three centers, and the same ranking scheme used in Task 1 was then applied.

The final leaderboard was determined by averaging each team's ranks across both tasks, providing a comprehensive assessment of overall performance throughout the challenge.

Additionally, to ensure the robustness and reliability of the rankings, bootstrapping (Efron, 1992) was applied. Bootstrapping, as emphasized by Maier-Hein et al. (2018), is a key method for assessing the variability of rankings and the stability of observed performance differences. In this study, the test set was repeatedly resampled with replacement for 10,000 iterations, and team rankings were recalculated for each resample. For each team, we computed the proportion of iterations in which it retained its original rank as well as the proportions in which it achieved each of the other possible ranks. Resampling indices are drawn once per replicate and shared across all methods, and predictions are aligned by case identifier before resampling, so that all methods are compared on identical resampled sets. To characterize the consistency of observed differences under resampling, we applied the Wilcoxon signed-rank test to the bootstrapped metric values obtained over all iterations. Because these values are computed over resampling replicates rather than over the test cases themselves, their effective sample size is set by the number of iterations rather than by the data, and we therefore report them as stability diagnostics rather than as inferential tests.

To contextualize absolute performance, we additionally evaluate four reference classifiers derived from training labels only. TrivialGlobal predicts the pooled training mode on every case, and TrivialLocal predicts each center's own training mode. RandomUniform samples the six grades with equal probability, and RandomPrior samples according to training-set prevalence; both stochastic references are simulated over 10,000 draws. All four are contextual references and are excluded from every rank computation.

As an inferential complement to the bootstrap analysis, we additionally conduct paired permutation tests over the test cases themselves. The exchangeable unit is the individual case: for each of 10,000 permutations, every case is independently assigned to one of the two methods under comparison with probability 0.5. For Task 1 the metric is recomputed over the held-out center's test cases; for Task 2 it is recomputed per center from the permuted assignment and the three centers are then averaged with equal weight, matching the ranking estimand defined above. All tests are two-sided, comparing the absolute observed difference against the absolute differences of the permutation distribution. With 10,000 permutations the smallest attainable p-value is 10^{-4}; values reported as 0.0001 should therefore be read as $p \leq 10^{-4}$, and a value of 1.0000 indicates that the observed difference was equaled or exceeded in essentially every permutation.

3. Results

The FedSurg challenge received 24 registrations, with four final submissions. However, one submission was disqualified due to non-executable code after the final deadline. This left three complete submissions for evaluation. The participating teams were: Santhi R. Kolamuri, who submitted independently as Team Santhi; Lorenzo Mazza and Claas de Boer from the Translational Surgical Oncology group at the National Center for Tumor Diseases, submitting as Team Elbflorenz; and Julia Alekseenko and Nicolas Padoy from the CAMMA research group at IHU Strasbourg, submitting as Team Camma. Furthermore, we provide a centralized baseline, a federated Swarm Learning (SL) baseline, and two parameter-efficient fine-tuning baselines, together with constant and chance-level reference classifiers.

3.1. Participating teams and methods

The following section details the methodologies submitted by participants in the FedSurg challenge. Each subsection outlines the architectural choices, training strategies, and FL configurations employed. Where relevant, we contextualize each approach within existing literature on foundation models, metric learning, and federated optimization. A summary of each submission is presented in Table 2.

3.1.1. Team Santhi

Santhi R. Kolamuri's approach leverages a pretrained Video Vision Transformer (ViViT) (Arnab et al., 2021), which is well-suited for capturing spatio-temporal features in surgical video sequences (Fig. 3(a)). ViViT's transformer-based architecture has proven more effective than conventional CNNs in modeling temporal dependencies. By utilizing ViViT pretrained on Kinetics-400, only the final classification layer is fine-tuned, while all other weights remain frozen. This linear probing approach is common for foundation models (Dosovitskiy et al., 2021), as it allows efficient adaptation while retaining robust pretrained representations.

The frame loading strategy is customized to emphasize salient temporal regions. From 200 available video frames, 32 are sampled for training. Two-thirds are selected from a narrow window around frame 100 to capture high-information segments, while the remainder of the frames are drawn from the full sequence with 60% probability to the frames near the center. This hybrid sampling balances local relevance and temporal diversity.

Training is implemented via the Flower FL framework (Beutel et al., 2022), using FedAvg for aggregation. Centers train locally with a batch size of four for five epochs per round, across 20 communication rounds. Cross-entropy loss is used with mixed precision training to improve memory efficiency. The lightweight design, frozen backbone, focused sampling, and shallow head enable fast convergence with low resource demands. This submission builds on the success of ViViT in centralized video classification while adapting it to the constraints of FL.

3.1.2. Team Elbflorenz

Team Elbflorenz adopts a foundation model-based strategy using EndoViT (Batić et al., 2024), a Vision Transformer pretrained on the Endo700k dataset consisting of diverse endoscopic images (Fig. 3(b)). Following a standard linear probing setup (Dosovitskiy et al., 2021), the encoder is frozen and only a lightweight classification head is trained. This setup is motivated by recent successes of foundation models in medical imaging (Yang et al., 2025; Schmidgall et al., 2024), where pretrained encoders generalize well even with limited task-specific data.

The model uses 100 equidistant frames from each video and independently classifies each frame. A weighted cross-entropy loss is used to mitigate class imbalance, with weights inversely proportional to class frequency. Final video-level predictions are determined via majority voting. In the event of a tie, average confidence scores guide

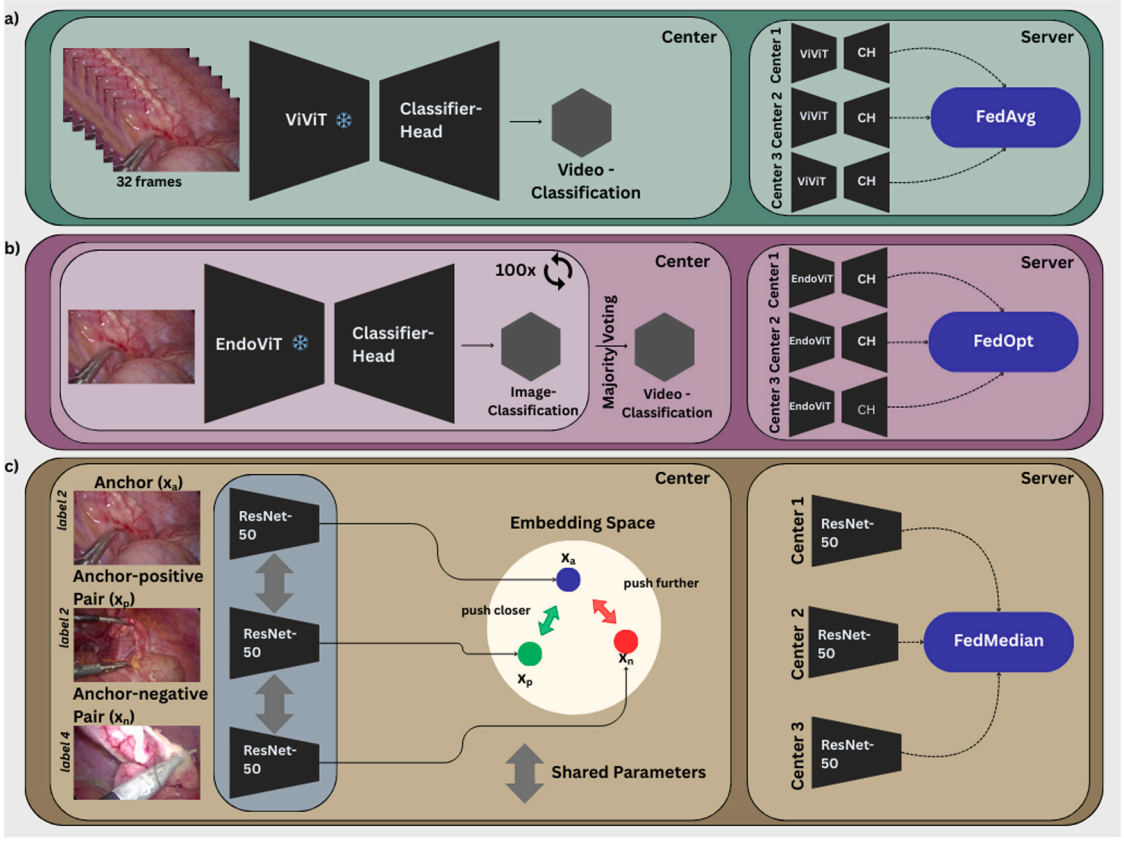


Fig. 3. Methods overview: The three submissions shown utilize different backbone architectures and federated strategies. A common approach is that in each server round, the best-performing model from a client's local training rounds is sent to the server for aggregation. (a) Team Santhi uses a frozen ViViT backbone with a fine-tuned classification head processing 32 frames per video, with updates aggregated via FedAvg. (b) Team Elbflorenz uses a frozen EndoViT backbone with a fine-tuned head, predicting single frames repeatedly and combining them via majority voting, with updates aggregated via FedSAM. (c) Team Camma uses ResNet50 models trained with a contrastive approach on positive and negative pairs, with updates aggregated via FedMedian. At inference, classification is performed by comparing the test embedding to a support set.

Table 2
Overview of submissions to the FedSurg Challenge, detailing each team's backbone model, prediction method, frame sampling, loss function, optimizer, learning rate, batch size, and federated learning configuration (FL strategy, FL-rounds, and local rounds).

Team	Santhi	Elbflorenz	Camma
Backbone model	Frozen pretrained Video Vision Transformer (ViViT) on Kinetics-400.	Frozen pretrained EndoViT on Endo700k.	Siamese network with a ResNet-50 backbone.
Prediction method	Video-level classification.	Majority voting on independent frame-level classifications.	Compares test embedding to class prototypes or a full support set.
Frame sampling	Samples 32 frames, with two-thirds from a narrow window around the center and the rest from the full sequence with a bias towards the center.	Uses 100 equidistant frames from each video.	Selects representative frames based on cosine similarity to a reference embedding, plus the central keyframe.
Loss function	Cross-entropy loss.	Weighted cross-entropy loss to address class imbalance.	Triplet margin loss with cosine similarity.
FL strategy	Federated Averaging (FedAvg).	Adaptive Federated Sharpness-Aware Minimization (FedSAM) on Client-side and FedOpt on Server Side.	Federated Median (FedMedian).
Optimizer	Adam	Adaptive Federated Sharpness Aware Minimization (FedSAM) based on SGD	Adam
FL-Rounds	20	50	10
Local rounds	5	2	5
Learning rate	1×10^{-4}	1×10^{-3}	1×10^{-6}
Batchsize	4	128	1

the decision. This per-frame classification strategy increases robustness by aggregating multiple frame-level predictions.

Federated optimization on the client side is handled using adaptive FedSAM (Caldarola et al., 2022; Foret et al., 2021), which extends Sharpness-Aware Minimization (SAM) to heterogeneous FL settings. This method encourages flatter minima and better generalization across non-IID (independent and identically distributed) client data. Similar to Team Santhi's approach, training was implemented via the Flower FL framework (Beutel et al., 2022). It employs 2 local epochs and 50 global rounds, with center-side class balancing and hyperparameter tuning. Compared to standard FedAvg, FedSAM has demonstrated improved convergence on non-IID data (Caldarola et al., 2022). Model aggregation on the server is done with FedOpt (Reddi et al., 2021). By combining EndoViT with adaptive optimization, Team Elbflorenz presents

a minimal yet effective pipeline for surgical appendicitis classification under federated constraints.

3.1.3. Team Camma

Team Camma employs a metric learning approach based on a Siamese network with a ResNet-50 backbone (He et al., 2015), inspired by the original Siamese architecture (Bromley et al., 1993). Unlike conventional classifiers that output class probabilities, this model maps input images to L2-normalized 256-dimensional embeddings. Training is performed using triplet margin loss with cosine similarity, encouraging embeddings from the same appendicitis stage to cluster while pushing apart those from different stages (Fig. 3(c)).

At inference, classification is performed by comparing the test embedding to a support set. This is done either via class prototypes (mean embeddings) or by averaging distances to all embeddings within each class. Camma observed that prototype-based inference worked best for Centers 2 and 3, while per-sample comparison was more effective for Center 1, allowing the approach to flexibly adapt to center-specific data distributions.

This metric learning paradigm is particularly well-suited for FL, as it focuses on learning a robust embedding space rather than simply aggregating classifier weights, which can be highly sensitive to the non-IID label distributions across centers.

To address domain heterogeneity, the model incorporates Switchable Normalization (Luo et al., 2021), which combines batch, instance, and layer normalization through softmax gating. This adaptive scheme enhances generalization across diverse centers, which is crucial in federated settings.

In addition, frame selection leverages embedding-based similarity: representative frames are chosen based on cosine similarity to a ResNet-50 reference embedding, alongside the keyframe (frame 100), to improve input consistency.

Training is conducted within a custom FL framework using the FedMedian aggregation strategy (Yin et al., 2018), which is robust to outliers. Each center performs 5 local epochs of triplet optimization per federation round, across 10 federation rounds (Table 2). The locally best-performing model (by F1-score) is selected for aggregation, ensuring only high-quality updates contribute to the global model. Overall, Team Camma's method demonstrates how metric learning, adaptive normalization, and robust aggregation can be combined for scalable, personalized FL in surgical video classification.

3.2. Baselines

In addition to the three participating teams, we provide four baselines for comparison: a centralized baseline, a federated Swarm Learning baseline, and two parameter-efficient fine-tuning baselines. The first two are based on the SurgTempoNet method (Saldanha et al., 2026).

SurgTempoNet employs a spatial–temporal architecture designed to predict patient-level clinical outcomes directly from laparoscopic videos. The model utilizes a ConvNeXt-Base (Liu et al., 2022) backbone, pre-trained on ImageNet-22k (Russakovsky et al., 2015), to extract high-level feature representations from individual frames. In its temporal configuration, these features are fed into a single-layer Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network with a hidden state size of 160. This allows the model to aggregate sequential information across the entire surgical procedure, mapping a sequence of frames to a single global label through a final linear classification layer.

To manage the high dimensionality of surgical footage, the pipeline uses a weakly supervised approach with sparse frame sampling. Videos are processed at a rate of one frame per second, and frames are resized to a resolution of 480 × 270 px. This sampling strategy reduces the computational burden of processing every raw video frame.

Both centralized and federated training is conducted using the AdamW optimizer with a learning rate of 1.0×10^{-3}. To address the inherent class imbalance in surgical datasets, the model utilizes Weighted Cross-Entropy Loss, where weights are inversely proportional to the frequency of each disease stage in the training set. Furthermore, a batch accumulation strategy is implemented, where gradients are computed over mini-batches of 64 clips, but weights are updated only every three steps, effectively stabilizing the training process for video sequences.

The defining characteristic of the federated baseline is its integration into a SL framework (Warnat-Herresthal et al., 2021; Hewlett Packard Enterprise, 2021). Unlike centralized training or standard FL, local models are trained on site-specific data, and only the model weights are synchronized via a peer-to-peer network managed by a SwarmCallback function. This synchronization occurs at a fixed frequency (e.g., every 100 batches) and requires a minimum number of active peers to proceed.

Furthermore, to probe whether parameter-efficient fine-tuning (PEFT) constitutes a viable adaptation strategy in this data regime, two additional preliminary models are included. Both apply Low Rank Adaptation (LoRA) (Hu et al., 2022) (rank 8, alpha 16) to the attention projections of a frozen backbone, combined with a shared three-layer MLP classification head and FedAvg aggregation over 25 rounds; all remaining hyperparameters (ASAM optimizer, cosine annealing, four local epochs per round) are held constant.

The *EndoViT+LoRA* baseline builds on the Team Elbflorenz architecture, adapting the frame-level EndoViT backbone (Batić et al., 2024) (ViT-Base, pretrained on endoscopic image data (Endo700k); ~940 K trainable parameters) with single-frame inputs (224 × 224). The *ViViT+LoRA* baseline extends this to temporal modeling by substituting a ViViT-B-16 × 2 backbone pretrained on Kinetics-400, processing 16-frame clips through native spatial–temporal attention (~969 K trainable parameters), following the architectural principle of Team Santhi. The two backbones therefore differ both in temporal modeling and in pretraining domain, endoscopic for EndoViT and general-purpose video for ViViT. Both baselines are post-hoc organizer analyses and are not part of the official challenge ranking. They are reported as point estimates throughout and are excluded from the bootstrap rank stability analysis, the permutation tests and the tabulated reference-classifier comparisons, all of which cover the three submissions together with the centralized and Swarm Learning baselines; where the two parameter-efficient baselines are compared against a reference classifier in the text, that comparison is made on point estimates alone.

3.3. Scores and rankings

The performance of the submitted federated models was evaluated in both tasks of the FedSurg Challenge: generalization to an unseen center (Task 1) and center-specific adaptation (Task 2). Performance was measured using Expected Cost (EC) and F1-score. Rank stability was assessed via 10,000-iteration bootstrapping, with statistical comparison by paired permutation tests over the test cases. Additionally, the baseline methods are reported separately in Section 3.4 as contextual references and are not part of the official challenge ranking. Overall, a comprehensive summary of results, including key metrics, confusion matrices, and rank stability, is provided in Tables 3 and 4 and Figs. 4, 5, 6, and 7, with further details in Appendix G.

3.3.1. Task 1: Generalization to an unseen center

Task 1 assessed the models' ability to generalize to Center 4, which was excluded during training. Overall, performance on this task was modest across all teams (Table 3).

Team Santhi achieved the highest F1-score (23.03%) and the lowest EC (12.41%), demonstrating the strongest generalization among the challenge submissions under domain shift. This is evident in the confusion matrix (Fig. 4), where Santhi's predictions tend to align with

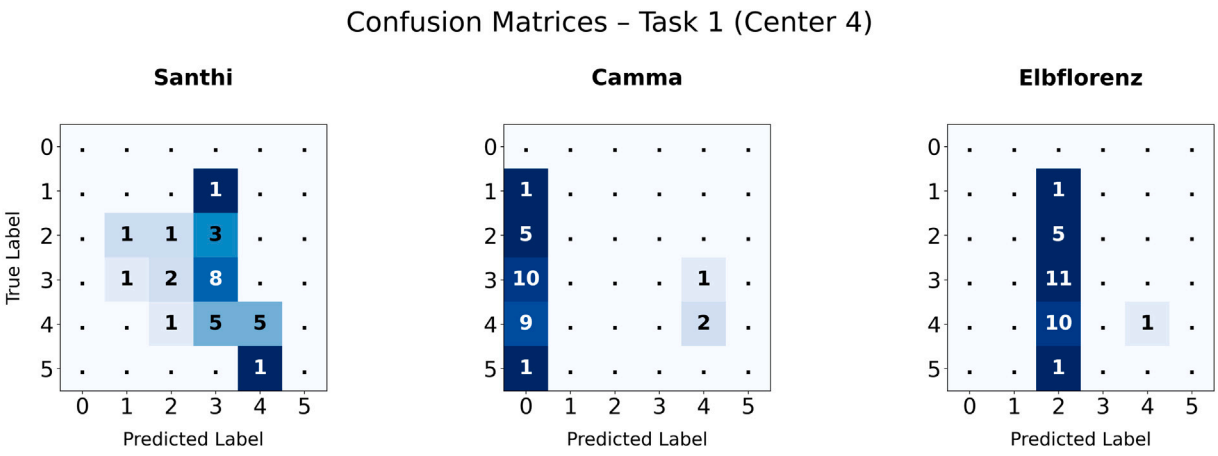


Fig. 4. Confusion matrices – Task 1, Center 4. Confusion matrices for the participating teams on Center 4 (Task 1). The values in the confusion matrices are not normalized. The color highlighting is normalized row-wise by true labels. The diagonal highlights class-wise recall, while off-diagonal values indicate common misclassification patterns.

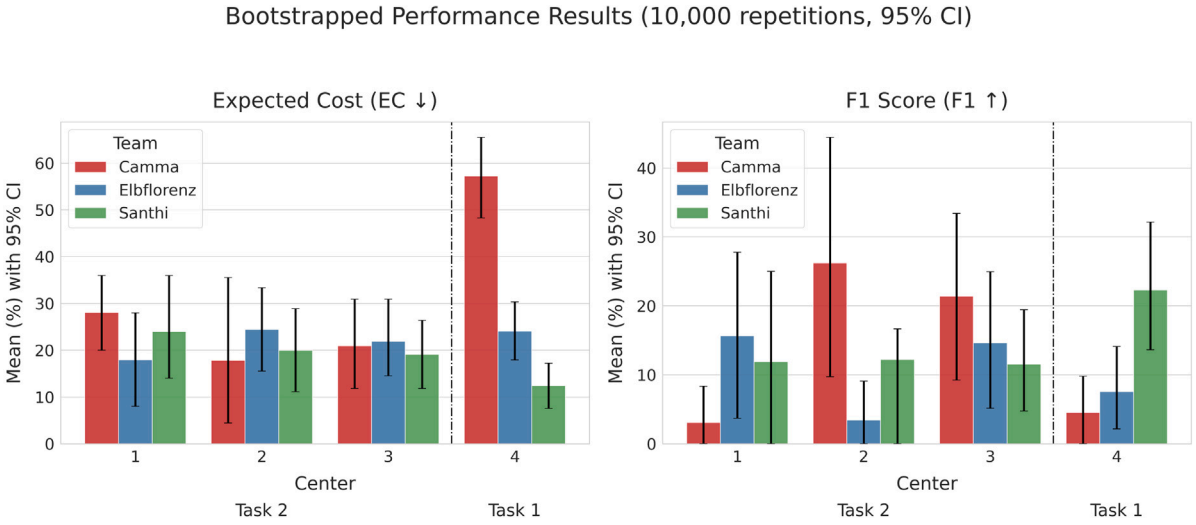


Fig. 5. Bootstrapped performance results. Expected Cost (EC ↓) and macro F1 (F1 ↑) for the three participating teams, shown as the bootstrap mean with 95% percentile confidence intervals as error bars, from 10,000 bootstrap repetitions. Centers 1–3 correspond to Task 2 (local adaptation) and Center 4 to Task 1 (generalization).

Table 3
Performance comparison of the three teams on Task 1 at the held-out center (Center 4), reporting Expected Cost (EC, lower is better) and F1-score (higher is better). Best results are highlighted in bold.

Team	EC ↓	F1 ↑
Camma	57.24%	4.76%
Elbflorenz	24.14%	7.83%
Santhi	**12.41%**	**23.03%**

the diagonal, in contrast to the predictions heavily biased towards a specific class. The model correctly identified class 3 in 8 out of 11 cases but struggled with adjacent stages (e.g., classes 2 and 4) and underrepresented ones (e.g., 0, 1, 5), likely due to class imbalance. Bootstrapped performance metrics (Fig. 5) indicate that this ordering is consistent under resampling. This is further supported by the ranking stability plots (Figs. 6(a) and 6(b)), where Team Santhi ranks first in 99.23% of bootstrapped samples for F1-score and 99.79% for EC.

Team Camma obtained the lowest F1-score (4.76%) and the highest EC (57.24%). The model predominantly predicted class 0, achieving only three correct predictions across the dataset. This behavior indicates a failure to generalize, likely caused by the model overfitting to the training distribution and being unable to adapt to the significant domain shift in the unseen test set. The poor performance of this collapsed classifier is supported by bootstrapping analysis (Fig. 5) and by the permutation tests, which separate Camma from every other method on Task 1 EC ($p \leq 10^{-4}$, the smallest value attainable with 10,000 permutations). Across 10,000 repetitions, Team Camma consistently ranked as the lowest-performing team in 100% of cases for the EC and 75.09% for the F1-score (Fig. 6).

Team Elbflorenz also demonstrated limited performance, predominantly predicting class 2 across inputs (Fig. 4). The model achieved an EC of 24.14% and an F1-score of 7.83%, ranking slightly above Team Camma. This is likely due to the classifier exhibiting a strong bias towards predicting classes adjacent to the most frequently observed labels in the training data. Confusion matrices and bootstrapped scores confirm weak generalization, placing Elbflorenz in the middle of the three submissions in terms of overall performance (Figs. 6, 5).

Paired permutation tests over the 29 Center-4 cases separate the two weakest submissions from the strongest: Camma and Elbflorenz each differ significantly from Santhi on both metrics ($p \leq 0.047$), and Camma differs from Elbflorenz on EC ($p = 0.0001$) though not on F1 ($p = 0.592$). The observed differences in generalization under domain shift are therefore resolvable at this sample size for the largest

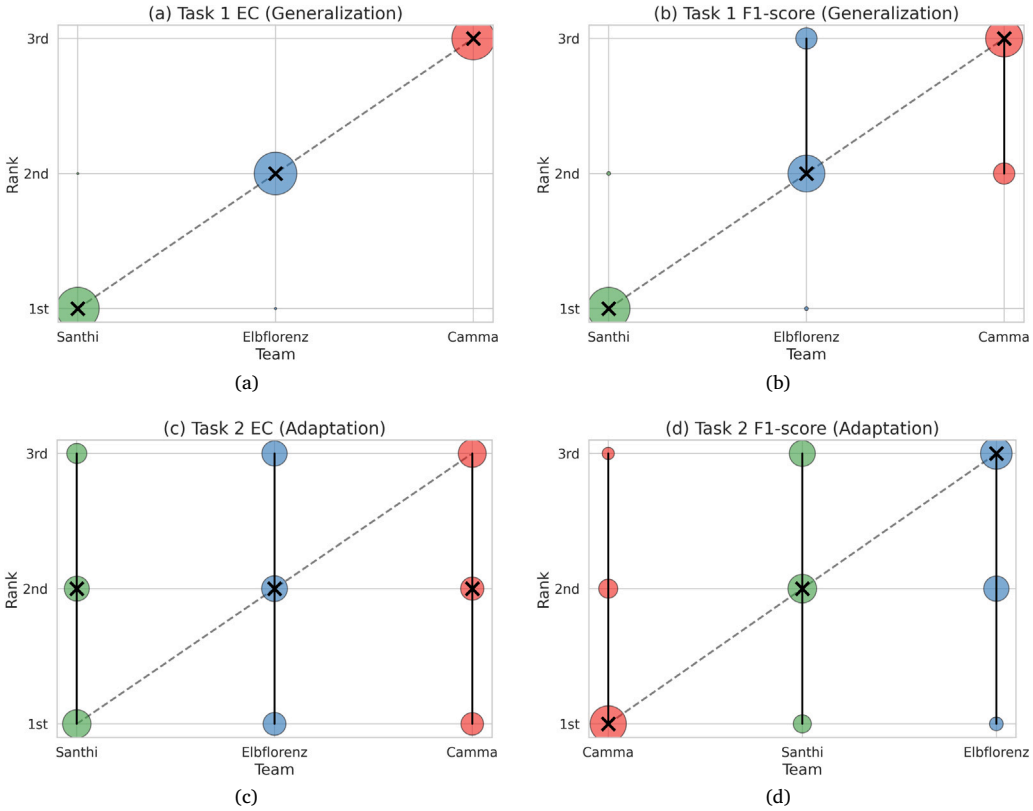


Fig. 6. Ranking stability. Bootstrapped ranking distributions for each metric and task, based on 10,000 bootstrap iterations. Circle size indicates the percentage of times a team's model achieved a specific rank across samples. Black crosses show median ranks, and black lines denote the 95% bootstrap confidence intervals. Subfigures (a) and (b) correspond to Task 1 (generalization ability) with metrics EC and F1-score, respectively, while (c) and (d) represent Task 2 (adaptation ability) using the same metrics.

Table 4
Performance comparison of the three teams on Task 2 across Centers 1, 2, and 3. The table reports Expected Cost (EC, lower is better) and F1-score (higher is better). Best results per metric and center are highlighted in bold.

Team	Center	EC ↓	F1 ↑
Camma		28.00%	3.33%
Elbflorenz	1	**18.00%**	**17.42%**
Santhi		24.00%	14.81%
Camma		**17.78%**	**30.28%**
Elbflorenz	2	24.44%	3.70%
Santhi		20.00%	13.33%
Camma		20.91%	**22.76%**
Elbflorenz	3	21.82%	15.51%
Santhi		**19.09%**	12.04%
Camma		22.23%	**18.79%**
Elbflorenz	Average	21.42%	12.21%
Santhi		**21.03%**	13.40%

gaps, while the ordering between the two weakest submissions rests on the EC metric alone. The ranking stability plot (Fig. 6) visually reinforces this trend, clearly showing Team Santhi in the lead, followed by Elbflorenz and Camma. The Task 1 component of the leaderboard presented in Table 5 is therefore well supported.

3.3.2. Task 2: Center-specific adaptation

Task 2 evaluated model performance in intra-center settings, where each local model was tested on its dedicated test set from its associated center. Absolute scores are therefore not directly comparable between the two tasks.

Team Santhi returned the narrowest spread of F1-scores across the three centers, at 14.81% (Center 1), 13.33% (Center 2) and 12.04% (Center 3), a range of under three percentage points, and the lowest Task 2 average EC of the three submissions (21.03%, Table 4). It was not the best submission at any individual center on F1, and it attained the lowest EC at Center 3 only. The confusion matrices in Fig. 7 show predictions distributed over grades 2 to 4 rather than concentrated on a single grade, with 2 of 10, 2 of 9 and 7 of 22 cases on the diagonal; separation between grades 2, 3 and 4 is limited at all three centers. The bootstrap intervals in Fig. 5 are wide and overlap between centers, and the rank distribution in Fig. 6 is correspondingly unstable. For the F1-score, Santhi ranked first in 12.39% of cases, second in 48.62%, and third in 38.99%. Regarding the EC metric, the values were 40.36%, 36.36% and 23.28%.

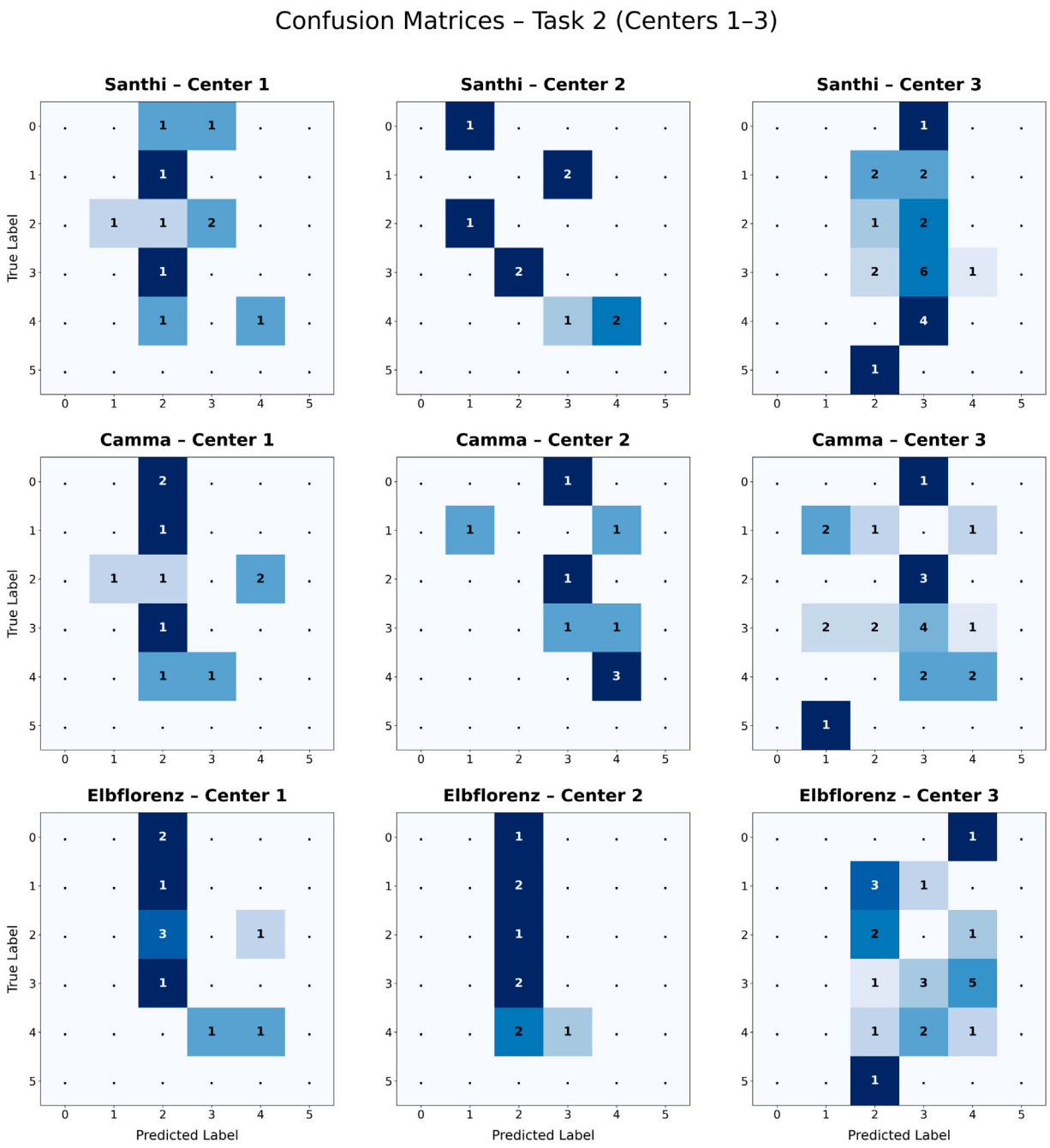


Fig. 7. Confusion matrices – Task 2, Centers 1–3. The values in the confusion matrices are not normalized. The color highlighting is normalized row-wise by true labels. The diagonal highlights class-wise recall, while off-diagonal values indicate common misclassification patterns.

Team Elbflorenz displayed greater inconsistency across centers, especially in the F1-score (Fig. 5). The confusion matrices (Fig. 7) reveal inconsistent local adaptation: the models for Centers 1 and 2 concentrate their predictions on a single grade, with 7 of 10 and 8 of 9 cases assigned to grade 2, while the model for Center 3 does not separate grades 2, 3 and 4. Notably, the Center 1 model records the highest F1-score of the three submissions at that center (17.42%) despite this concentration, whereas the same behavior at Center 2 yields the lowest value of the section (3.70%). Center 3 is intermediate at 15.51%. The bootstrap intervals in Fig. 5 indicate that these per-center values are not stable under resampling. For the F1-score, Elbflorenz ranked first in 9.09% of cases, second in 37.10%, and third in 53.81%. Regarding the EC metric, Elbflorenz achieved first place in 36.00% of cases, second place in 32.30%, and third place in 31.70%.

Team Camma shows the widest center-to-center variation of the three submissions (Table 4, Fig. 5), excelling at Center 2 with an F1-score of 30.28% and at Center 3 with 22.76%, but underperforming significantly at Center 1, where the F1-score was 3.33%. Confusion matrices shown in Fig. 7 indicate intermediate performance relative to Teams Santhi and Elbflorenz, as the models demonstrate a modest ability to differentiate between classes, evidenced by a visible diagonal tendency. Similar to the other submissions, bootstrapping analyses confirm variability in performance across centers. The ranking for Team Camma also fluctuates across centers, with F1-score rankings of 1st place in 78.52% of cases, 2nd place in 14.29%, and 3rd place in 7.19%. For the EC metric, the rankings were first place in 23.68%, second place in 31.42%, and 3rd place in 44.90%.

Table 5
Team rankings for the FedSurg Challenge are presented, where lower ranks indicate better performance. The Task 2 EC ranking is not reproducible under resampling and is reported for completeness only; no team retains its position in at least 50% of bootstrap replicates, and no Task 2 comparison reaches significance under the permutation test (Section 3). The Final Rank column inherits this instability through the Task 2 average.

Team	Task 1 Ranks			Task 2 Ranks			Final Rank
	EC	F1-score	Avg	EC	F1-score	Avg	
Camma	3	3	3	3	1	2	2
Elbflorenz	2	2	2	2	3	3	2
Santhi	1	1	1	1	2	1	1

The bootstrapped ranking stability analysis (Figs. 6(c) and 6(d)) shows notable fluctuations in team positions within the bootstrapping. While the Task 2 F1-score rankings (Fig. 6(d)) align with the overall standings in Table 5, the EC mean-based rankings (Fig. 6(c)) reveal no clear winner with respect to the median rank. Nevertheless, the frequency of rank positions matches the trend of the general ranking in Table 5. The Task 2 EC ranking is close to uniform, with first-rank probabilities of 0.40, 0.36 and 0.24 for Santhi, Elbflorenz and Camma and mean ranks of 1.83, 1.96 and 2.21. Such variability highlights that Task 2 outcomes are sensitive to small changes in the test set, and claims regarding top-performing submissions on this task are not supported by the evidence available here. Following the criterion of Maier-Hein et al. (2018), no team retained its original EC rank in at least 50% of bootstrap replicates, and no Task 2 comparison reaches significance under the permutation test. The Task 2 EC ordering therefore carries essentially no information about submission superiority. We report it for completeness as a record of the challenge outcome, not as evidence of relative method quality.

The Wilcoxon signed-rank statistics reported in Appendix G are computed over the 10,000 bootstrap replicates rather than over the underlying test cases. Their effective sample size is therefore the number of resampling iterations, which is a chosen parameter, and not the 29 test cases of the held-out center or the 41 cases underlying the Task 2 average. These values quantify how consistently one method exceeds another under resampling of a fixed test set. They do not constitute inference about a population of centers or patients, and they are reported as stability diagnostics rather than as hypothesis tests, labeled as bootstrap discordance in Appendix G.

To provide an inferential counterpart, we additionally conducted paired permutation tests over the actual test cases, as described in Section 2.4.1. Five of the six pairwise comparisons on Task 1 reach $p < 0.05$, the exception being Camma versus Elbflorenz on F1 ($p = 0.59$); the significant comparisons involve EC differences of 12 to 45 percentage points, or F1 differences of 15 percentage points or more. No Task 2 comparison reaches significance on either metric. The contrast with the bootstrap diagnostics is instructive: comparisons that the resampling diagnostics report as separated at arbitrarily small values are not resolved at the sample size actually available. At 70 test cases, only very large effects are detectable, and the Task 2 ranking in particular should be interpreted accordingly.

3.4. Further analysis

In this section the challenge participant results are compared with the centralized and federated baselines. Furthermore, we analyze the communication efficiency of the different approaches. The methods used are described in Section 2.4.1.

3.4.1. Baseline comparison

To provide contextual reference points, we introduced four baselines. A centralized (Cen) baseline, representing training on pooled data from all centers for Task 1 and on each center's data separately for Task 2, and a SL (SWARM) baseline, applying the same architecture in a decentralized peer-to-peer setup, together provide a matched comparison of pooled and decentralized training under otherwise similar conditions. Two additional PEFT baselines, EndoViT+LoRA and ViViT+LoRA, probe whether parameter-efficient adaptation can serve as a practical alternative to full local fine-tuning. Table 6 summarizes absolute raw test performance across both tasks, while Table 7 reports center-wise Task 2 results for all baseline models.

Table 6
Absolute raw test performance comparison including baselines. EC: lower is better; F1: higher is better. Baseline rows in the lower half of the table are contextual references and are not part of the official challenge ranking. Bold numbers mark the best point estimations only.

Method	Task 1 EC	Task 1 F1	Task 2 Avg EC	Task 2 Avg F1
Camma	57.24%	4.76%	22.23%	18.79%
Elbflorenz	24.14%	7.83%	21.42%	12.21%
Santhi	12.41%	23.03%	21.03%	13.40%
Cen	**10.34%**	**26.31%**	19.82%	**23.51%**
SWARM	13.10%	19.00%	20.41%	17.77%
EndoViT+LoRA	17.24%	9.24%	20.13%	8.70%
ViViT+LoRA	22.76%	15.93%	**17.60%**	21.47%

Table 7
Center-wise Task 2 baseline results. EC: lower is better; F1: higher is better.

Baseline	Center 1		Center 2		Center 3	
	EC	F1	EC	F1	EC	F1
Cen	24.00%	11.43%	20.00%	**26.43%**	**15.45%**	**32.67%**
SWARM	28.00%	10.32%	17.78%	20.95%	**15.45%**	22.03%
EndoViT+LoRA	**17.78%**	10.32%	24.44%	6.06%	18.18%	9.68%
ViViT+LoRA	22.22%	**12.22%**	**13.33%**	22.62%	17.27%	29.57%

Comparative performance. The centralized baseline achieved the highest point estimates on both metrics in Task 1 and on F1 in Task 2; on the Task 2 average Expected Cost, ViViT+LoRA (17.60%) is lower than the centralized baseline (19.82%). In Task 1 (Generalization), the centralized baseline reached an F1-score of 26.31% and an EC of 10.34%, followed by Santhi and the SWARM baseline (compare Fig. 9).

In Task 2 (Adaptation), the performance gap narrowed, reflecting the challenge of center-specific fine-tuning. While the centralized model maintained the lead (Avg. F1: 23.51%), the SWARM baseline placed second on the Task 2 average, though its per-center results vary considerably. Camma and Elbflorenz both improved substantially in Task 2 relative to their Task 1 performance, yet neither reached the performance level of the centralized or SL baselines.

The two PEFT baselines differ substantially in Task 1 F1. EndoViT+LoRA achieved 9.24% F1 on Task 1 and 8.70% average F1 on Task 2. Relative to Elbflorenz, the frame-level submission it is derived from, it is the better of the two on three of the four task and metric combinations (Task 1 EC 17.24% against 24.14%, Task 1 F1 9.24% against 7.83%, Task 2 average EC 20.13% against 21.42%) and worse only on the Task 2 average F1 (8.70% against 12.21%). Inspection of per-class predictions confirms near-complete collapse to the majority class, and its comparatively low Task 1 Expected Cost should be read in that light: a constant predictor of the pooled training mode attains 13.79% on the same test set, so a value of 17.24% reflects concentration on a central grade rather than discrimination. ViViT+LoRA, by contrast, achieved 15.93% F1 on Task 1 and 21.47% average F1 on Task 2, ranking second among all methods in Task 2. Notably, at Center 3, ViViT+LoRA reached 29.57% F1, the highest Task 2 F1 among all non-centralized methods and approaching the centralized baseline (32.67%). Performance at the smaller centers remained limited (Center 1: 12.22%, Center 2: 22.62%). The ordering of the two parameter-efficient baselines is not consistent across metrics: on Task 1 Expected Cost, EndoViT+LoRA (17.24%) is lower than ViViT+LoRA (22.76%), reversing the F1 ordering, although both remain above

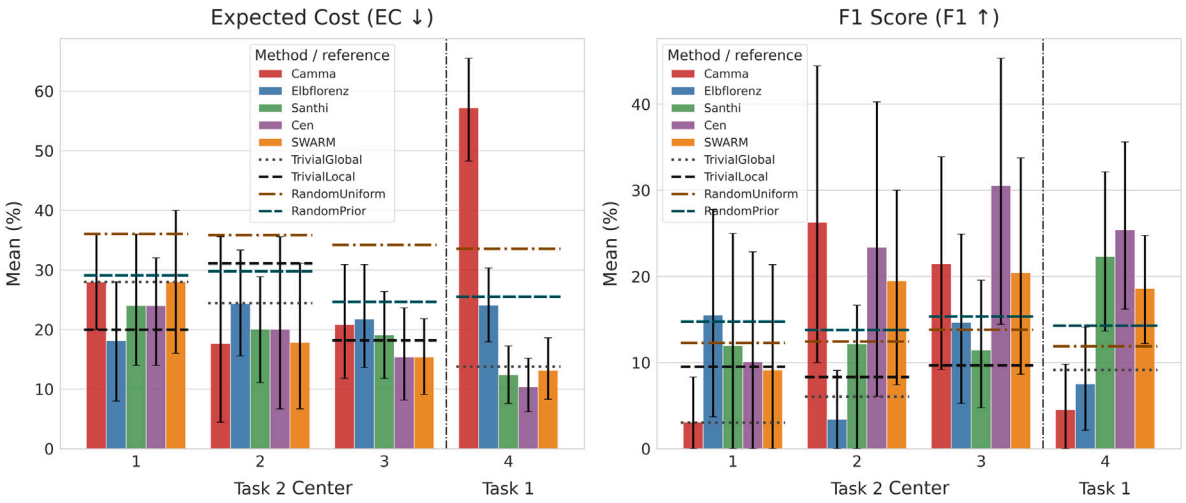


Fig. 8. Bootstrapped performance results with baselines. Expected Cost (EC ↓) and macro F1 (F1 ↑) for the three participating teams and baselines, shown as the bootstrap mean with 95% percentile confidence intervals as error bars, from 10,000 bootstrap repetitions. Centers 1–3 correspond to Task 2 (local adaptation) and Center 4 to Task 1 (generalization). Horizontal lines mark four non-ranked reference classifiers derived from training labels only: TrivialGlobal, a constant predictor of the pooled training mode (grade 3); TrivialLocal, a constant predictor of each center's own training mode (grades 2, 4 and 3 for Centers 1 to 3); RandomUniform, sampling all six grades with equal probability; and RandomPrior, sampling according to training set prevalence. The two stochastic references are simulated over 10,000 draws and shown as their mean. The two constant references coincide at Center 3, where the local and pooled training modes are identical, and TrivialLocal is undefined for Task 1, whose held-out center contributes no training labels. Reference values are contextual only and are excluded from all rank computations.

the 13.79% of a constant predictor. The two backbones differ both in temporal modeling and in pretraining corpus (Section 3.2); the attribution of this difference is discussed in Section 4.1.

Reference classifiers. Absolute values on this task are difficult to interpret without a reference point, and macro F1 in particular behaves unintuitively under the observed label distribution. We therefore evaluated four reference classifiers, in two families, on every test set. Two constant classifiers predict a fixed grade: TrivialGlobal predicts the pooled training mode (grade 3), and TrivialLocal predicts each center's own training mode (grades 2, 4 and 3 for Centers 1 to 3 respectively). Two stochastic classifiers were simulated over 10,000 draws: RandomUniform samples all six grades with equal probability, and RandomPrior samples according to training set prevalence. All references are derived from training labels only and are excluded from every rank column. Table H.25 reports their values.

Four observations follow.

First, the attainable macro F1 is bounded below 100% wherever a grade is absent from the test set, since absent classes contribute zero against a fixed denominator of six. The ceiling is 83.33% at Centers 1, 2 and 4, and 88.89% for the Task 2 average. Reported F1 values should be read against this bound rather than against unity. At the other end of the scale, the most generous chance model is a stochastic classifier matching the *test* prevalence exactly, for which each class contributes $F1_c = p_c$ in expectation and the macro average is therefore $1/C$, that is 16.67% for six grades. That model presupposes knowledge of the test distribution and is not attainable in practice; the two stochastic references reported here use training prevalence or a uniform draw and consequently sit lower, at 14.30% and 11.91% on the held-out center.

Second, on Task 1 the constant classifier reaches 13.79% EC, which two of the three submissions fail to improve upon (Elbflorenz 24.14%, Camma 57.24%); both parameter-efficient baselines also fail to improve on it (EndoViT+LoRA 17.24%, ViViT+LoRA 22.76%). Camma's cost exceeds even uniform random guessing (33.55%), reflecting a predicted distribution systematically displaced from the ordinal position of the true labels.

Third, at Center 1 the constant classifier predicting that center's training mode reaches 20.00% EC, which five of the seven evaluated methods fail to improve upon, including the centralized baseline (24.00%) and the Swarm baseline (28.00%). Only Elbflorenz (18.00%) and EndoViT+LoRA (17.78%) are more accurate under the ordinal cost, and both by at most 2.2 percentage points. On macro F1 the same center places both baselines below the prior matched chance level. Center 1 is therefore a setting in which neither centralized nor decentralized training establishes a meaningful advantage over trivial prediction.

Fourth, at Center 3 no submission improves on the constant classifier on EC (Santhi 19.09%, Camma 20.91%, Elbflorenz 21.82%, against 18.18%), while the centralized, Swarm and ViViT+LoRA baselines do (15.45%, 15.45% and 17.27%); EndoViT+LoRA matches the reference exactly at 18.18%. Center 3 contributed 80 of 153 training videos, so the submissions fail to exceed a constant predictor at precisely the node where local data is most abundant.

Across all test sets and both metrics, eleven method, test set and metric combinations fall below at least one chance classifier, giving twenty individual comparisons, since a result may fall below both stochastic references. On the Task 2 average F1, Elbflorenz (12.21%) and Santhi (13.40%) both sit below the prior matched chance level (14.65%). These comparisons are reported in full in Table H.26.

The constant references require one caveat. TrivialLocal at Center 2 predicts grade 4, which lies near the edge of the ordinal scale, so the linear distance penalty inflates its cost to 31.11% EC against 24.44% for the mid scale TrivialGlobal on identical data. Comparisons at Center 2 are therefore made against TrivialGlobal, and the fact that all methods improve on TrivialLocal there should not be read as evidence of adaptation quality.

Rank stability and statistical significance. To address the stability of these findings beyond point estimates, we conducted a stability analysis using 10,000-iteration bootstrapping. This process quantifies the probability of each team achieving a specific rank under different data distributions, as visualized in Fig. 9. As shown in Fig. 9(a) and (b), the rankings for Task 1 were comparatively stable. For the F1-score, the centralized baseline retained the 1st rank in 65.7% of iterations, while SWARM maintained the 3rd rank in 69.8% of cases. In contrast, Task 2 exhibited higher volatility (Fig. 9(c) and (d)), particularly in the EC metric. No method retained its nominal Task 2 EC rank in more than

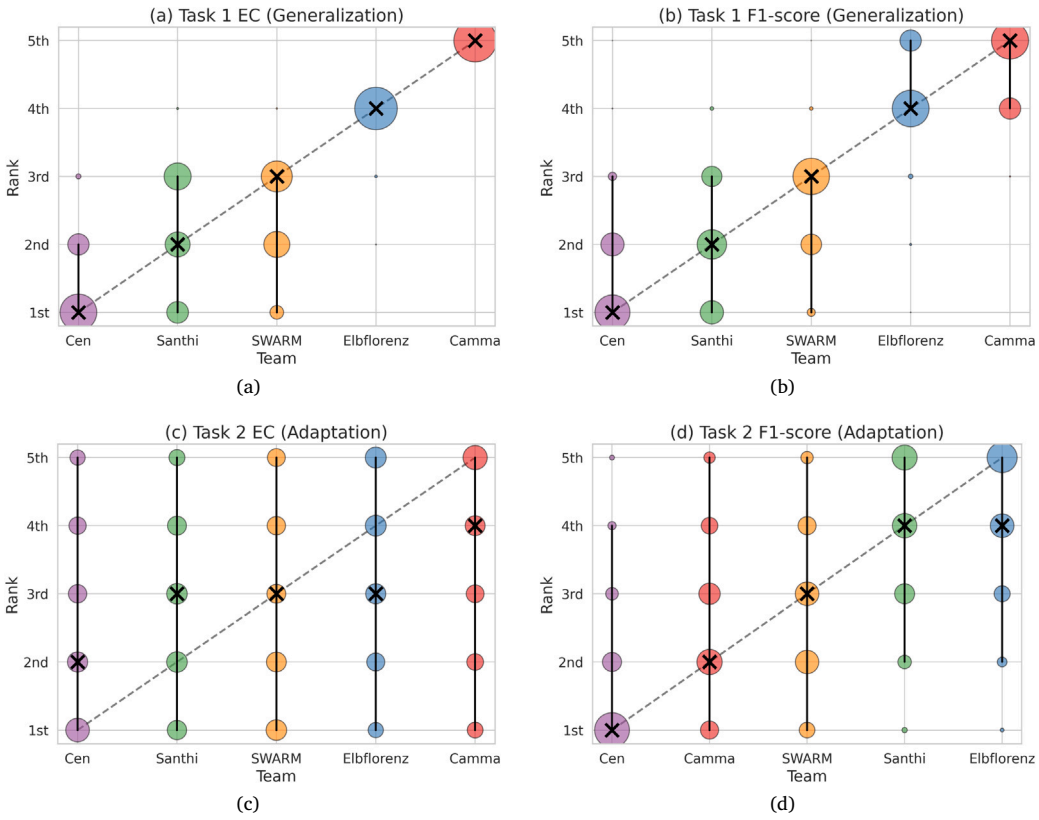


Fig. 9. Ranking stability. Bootstrapped ranking distributions for each metric and task, based on 10,000 bootstrap iterations. Circle size indicates the percentage of times a team's model achieved a specific rank across samples. Black crosses show median ranks, and black lines denote the 95% bootstrap confidence intervals. Subfigures (a) and (b) correspond to Task 1 (generalization ability) with metrics EC and F1-score, respectively, while (c) and (d) represent Task 2 (adaptation ability) using the same metrics.

37.9% of replicates, including the centralized baseline, and Camma attained the first rank in 8.6% of replicates despite ranking last on point estimates. Restricted to the three submissions alone, the Task 2 EC ranking is close to uniform, with 1st-rank probabilities of 0.40, 0.36 and 0.24 for Santhi, Elbflorenz and Camma and mean ranks of 1.83, 1.96 and 2.21. Following the criterion of Maier-Hein et al. (2018), the Task 2 EC ordering carries essentially no information about submission superiority.

As set out in Section 3.3.2, the Wilcoxon signed-rank statistics in Appendix G are computed over bootstrap replicates rather than over test cases and are reported as stability diagnostics rather than as hypothesis tests. They apply here in the same way, with the baselines included.

To provide an inferential counterpart, we additionally conducted paired permutation tests over the actual test cases (Section 2.4.1). Twelve of forty pairwise comparisons reach $p < 0.05$, all of them on Task 1 and all involving Camma or Elbflorenz, whose Task 1 EC differences from the remaining methods range from 11 to 47 percentage points. No Task 2 comparison reaches significance on either metric (minimum $p = 0.507$ for EC and 0.073 for F1). In particular, the Santhi submission and the SWARM baseline achieve Task 2 average Expected Cost values of 21.03% and 20.41%, a difference of 0.62 percentage points, with heavily overlapping bootstrap intervals (16.00%–26.22% and 14.46%–26.77%), a bootstrap win probability of 0.407, and a permutation p-value of 0.844; their adaptation performance under the ordinal cost is statistically indistinguishable at this sample size, notwithstanding their divergence in F1-score. We note that this conclusion rests on the width of the intervals relative to the effect rather than on the point estimates being close. At 70 test cases, only very large effects are resolvable, and the ranking should be interpreted accordingly.

The variability of these outcomes is further illustrated in Fig. 8, where error bars represent the 95% percentile confidence interval across bootstrap replicates. The narrower distributions for Task 1 compared to the wider variance in Task 2 underscore the inherent difficulty of the adaptation task, where local center characteristics significantly influence model stability.

3.4.2. *Communication efficiency analysis*

Communication cost is a critical metric in decentralized machine learning because the frequent transmission of high-dimensional model parameters, weights, or gradients over limited or unstable network bandwidth can become a significant bottleneck (McMahan et al., 2017). In many medical AI applications, the time required for network synchronization often dwarfs the time spent on local GPU computation. Consequently, optimizing these costs is essential for making collaborative learning scalable and practical for real-world clinical scenarios.

In a traditional FL architecture, the total communication costs can be calculated using the following formula:

$$C_{\text{total}} = 2 \times K \times T \times |W| \tag{5}$$

where:

- C_{total} is the total communication cost (data transferred).
- The factor of 2 accounts for the bi-directional transfer (upload and download) of model parameters between the client and the central server in each round.
- K is the number of participating clients (or nodes) per round.

- T is the total number of communication rounds.
- $|W|$ represents the size of the model parameters or weights.

However, for SL, the communication cost is calculated differently due to its peer-to-peer (P2P) nature. Since a central server does not exist, nodes synchronize weights directly through a decentralized mesh. In a leader-based synchronization (as utilized in the SurgTempoNet framework), one node is elected to aggregate weights from its peers. Consequently, the communication cost formula is adjusted to reflect the $K-1$ active connections:

$$C_{\text{total}} = 2 \times (K-1) \times S \times |W| \tag{6}$$

where S represents the total number of synchronization events across the entire training duration. In the SurgTempoNet implementation, synchronization is governed by a frequency parameter rather than fixed rounds. The total number of synchronization events S for a node is defined as:

$$S = \sum_{n=i}^{K} \left\lfloor \frac{E \times B_i}{f_{\text{sync}}} \right\rfloor + 1 \tag{7}$$

where:

- E is the number of local epochs performed.
- B_i is the number of batches contained in each epoch for center i.
- f_{sync} is the synchronization frequency (e.g., every 100 batches).
- The $+1$ accounts for the initial synchronization at the start of training.

By utilizing $2(K-1)$, the model accounts for the fact that the aggregating node does not incur external network traffic for its own local weights. This provides a more rigorous representation of the physical network traffic generated during the decentralized training of surgical video classifiers across heterogeneous hospital sites.

Table 8 reveals again significant variance in both architectural complexity and the resulting communication overhead during the FL process.

Model architecture and paramerization. The models utilized by Santhi, Elbflorenz, and our SL Baseline share a similar scale, ranging around 88M parameters. In contrast, team Camma implemented a significantly more compact architecture. With only 24M parameters, Camma's model is approximately 73% smaller than the other methods. All models utilize 32-bit floating-point precision, resulting in a model size of approximately 340 MB for the larger architectures and around 92 MB for Camma's compact model.

Synchronization events and total communication costs. As discussed, the total communication cost is a function of both the model size and the frequency of synchronization events. The SL Baseline, with its high synchronization frequency (100 events), incurs the highest communication cost, totaling approximately 133.14 GB across all clients and rounds, despite saving communication compared to the other methods by using a P2P approach. Furthermore, the impact of synchronization frequency is most evident when comparing Santhi and Elbflorenz.

Team Elbflorenz, with 50 synchronization events, results in a total cost of around 97.47 GB. Team Santhi, with only 20 synchronization events, significantly reduces the communication overhead to approximately 39.63 GB.

Finally, team Camma, with the smallest model and lowest synchronization frequency (10 events), achieves the lowest communication cost of about 5.38 GB.

In summary, team Camma emerged as the most communication-efficient team by a wide margin. By combining a reduced parameter count with a highly conservative synchronization strategy (10 sync events), they achieved a total communication cost of only 5.38 GB. Compared to the SL Baseline, this represents a 95.9% reduction in network traffic, highlighting the efficacy of aggressive model compression and sparse synchronization in federated environments.

In the current implementation, both PEFT baselines transmitted full model snapshots during aggregation, incurring communication costs comparable to full-model FL methods. Since only the LoRA adapter weights differ between federation rounds (~940–969 K parameters, ~3.8 MB per model), a dedicated LoRA-only synchronization strategy would reduce per-round transmission to less than 2% of the costs reported in Table 8. This efficiency advantage was not exploited here, but represents a compelling practical benefit of PEFT-based FL for future implementations.

4. Discussion

The FedSurg challenge should be understood as a proof-of-concept benchmark: the appendectomy task was selected for the combination of a standardized workflow, an established grading system, and genuine multi-center availability, in a procedure not previously represented in surgical vision datasets (Kolbinger et al., 2026). The following discussion focuses on methodological insights transferable to future FL applications on surgical video, structured around two central questions: first, what architectural and algorithmic choices best support global generalization to unseen centers; and second, whether and how local adaptation can be reliably achieved without sacrificing the globally learned representation.

4.1. The role of temporal modeling

A key finding of this challenge is that temporal modeling was the design choice most clearly associated with performance in this heterogeneous setting. Frame-level approaches, whether operating through majority voting (Team Elbflorenz) or embedding-based prototype comparison (Team Camma), are inherently limited by the high intra-video visual variance present in laparoscopic recordings. Differences in laparoscopic hardware, lighting conditions, and patient anatomy render individual frames unreliable predictors of patient-level inflammation grade. The metric learning approach of Team Camma illustrates this most clearly: while it achieved competitive performance at centers whose visual characteristics were represented in training, it failed to map embeddings of the unseen center to established training prototypes, revealing the brittleness of a feature space built on individual frames.

Video-level temporal models, by contrast, capture the progression of visual features across the surgical sequence, providing contextual information no single frame can supply. The ViViT-based submission of Team Santhi, processing 32 frames jointly, consistently achieved the strongest generalization to the unseen center in Task 1, outperforming both frame-level approaches despite using fewer communication resources. This finding is consistent with prior work on decentralized surgical video analysis using the full Appendix300 cohort, of which the challenge data is a subset (Saldanha et al., 2026), in which temporal architectures similarly outperformed frame-level methods across patient-level prediction tasks. The two parameter-efficient baselines come closest to a controlled comparison in this work. EndoViT+LoRA and ViViT+LoRA hold the aggregation strategy, optimizer, learning-rate schedule, round count and classification head fixed, and differ in the frozen backbone. That backbone differs in two respects, frame-level versus temporal processing and endoscopic versus general-domain pretraining, so the resulting gap on Task 1 (F1 9.24% versus 15.93%) does not isolate temporal modeling as the sole cause. It is consistent with the ordering observed among the submissions, and the fact that the domain-specific backbone is the weaker of the two argues against pretraining domain as the explanation. The gap is also metric-dependent. On Task 1 Expected Cost the ordering reverses, with the frame-level EndoViT+LoRA at 17.24% against 22.76% for ViViT+LoRA, because collapse onto a central grade is inexpensive under a linear ordinal penalty even when it carries no discriminative information; neither improves on the 13.79% of a constant predictor there. In the controlled

Table 8
Summary of architectural parameters and cumulative synchronization data costs per team.

Team	SL Baseline	Santhi	Elbflorenz	Camma
Parameters	89,351,470	88,651,014	87,213,038	24,086,014
Data type	float32 (4 bytes/param)	float32 (4 bytes/param)	float32 (4 bytes/param)	float32 (4 bytes/param)
Model size	357,405,880 B (340.85 MB)	354,604,056 B (338.18 MB)	348,852,152 B (332.69 MB)	96,344,056 B (91.88 MB)
Clients	3	3	3	3
Sync events	100	20	50	10
Total costs	142,962,352,000 B (133.14 GB)	42,552,486,720 B (39.63 GB)	104,655,645,600 B (97.47 GB)	5,780,643,360 B (5.38 GB)

comparison the temporal advantage therefore holds on three of the four task and metric combinations rather than on all four, and the claim that temporal modeling is the dominant architectural factor rests primarily on the submissions, where the ordering is consistent across both metrics on Task 1.

With respect to the generalization vs. adaptation trade-off, temporal modeling also offered a partial advantage in Task 2: Team Santhi's frozen ViViT backbone preserved global spatiotemporal representations during local fine-tuning, limiting the degree to which center-specific updates could corrupt the globally learned feature space. This suggests that architecture choice is not only a generalization strategy but also a safeguard against the instability of local adaptation. This reading requires one qualification: Santhi's Task 2 average F1 of 13.40% falls below the prevalence-matched chance level of 14.65% (Section 3.4), so the advantage is visible in the consistency of the Expected Cost across centers rather than in absolute discriminative performance, and no Task 2 comparison is statistically significant in any case. This point is revisited in Section 4.4.

4.2. Dataset scale as the binding constraint

The generally low absolute scores across all submissions raise an important question: how much of the low performance is attributable to the task and data regime rather than to federation? The centralized baseline provides a partial answer. Even when pooling all multi-institutional data onto a single server, thereby removing all federated constraints, the centralized model achieved only 26.31% F1-score on the unseen center. The underlying task of patient-level appendicitis grading from surgical video is therefore intrinsically difficult, regardless of the learning paradigm employed, due to the high visual variance across centers and the inherent ambiguity of fine-grained clinical grading.

The reference classifiers reported in Section 3.4 sharpen this picture. On the unseen center both baselines exceed all reference classifiers on point estimates, but the margins differ in kind. Under macro F1 the centralized model sits 2.14 standard deviations above the prevalence-matched chance level, a value reached by only 3.19% of chance draws, whereas the decentralized baseline sits 0.80 standard deviations above it and is matched or exceeded by 18.25% of draws; on this metric it is therefore above chance descriptively but not distinguishable from it. Under the EC both baselines fall far below the chance references, and the binding comparison is instead the constant classifier, which they improve on by 3.45 and 0.69 percentage points respectively. The two metrics behave differently because macro F1 is a weak instrument here: absent classes score zero against a fixed denominator of six, and the metric ignores how far a prediction lies from the true grade. The EC, which penalizes ordinal distance, retains signal that macro F1 discards, and should be treated as the primary indicator of model quality throughout.

This does not establish that decentralization causes the gap. The two baselines are not statistically separable from one another at this sample size, and we therefore make no claim of a quantified decentralization penalty. What the comparison does show is that the two paradigms sit on opposite sides of the chance threshold on one metric, and that the margin over trivial prediction is narrow for both. This distinction should be drawn carefully: that one baseline separates from chance while the other does not is a statement about two separate contrasts, not about the contrast between them. The direct comparison of the two is not significant on either metric ($p = 0.243$ for F1 and $p = 0.357$ for EC, Table H.24), and a difference in separability from a common reference is not evidence of a difference between the methods themselves.

The adaptation setting is less favorable still. At Center 1 only two of the seven evaluated methods improve on a constant classifier, and the centralized baseline is not among them; at Center 3 no participant submission improves on it, while three of the four organizer baselines do. Aggregate Task 2 scores therefore conceal substantial variation between centers, and center-level results should be read alongside the averages rather than replaced by them.

The most plausible explanation is data scale. With 153 training videos across three training centers, federated training fragments an already small pool, leaving individual nodes with too few samples to learn robust local representations. A subsequent study applying a related decentralized pipeline to a larger multinational cohort reports performance comparable to centralized training (Saldanha et al., 2026); that work differs in cohort size, center count and task definition, so it is not a controlled comparison, but the contrast points to the data regime rather than the decentralization mechanism as the binding constraint. Scaling the dataset, in both total videos and contributing centers, is a necessary condition for a more definitive evaluation of FL on surgical video.

4.3. Data heterogeneity and the limits of local adaptation

The challenge results highlight that non-IID data remains the central unsolved problem in federated surgical video analysis. Distribution shift across centers originates from two compounding sources: hardware-induced feature skew, including differences in laparoscopic systems, resolution, lighting, and cropping, and patient population heterogeneity, including differences in age distribution, disease severity prevalence, and surgical practice between academic and community centers. Current aggregation strategies, including FedAvg and FedSAM, partially mitigate but do not resolve these shifts.

These effects are further amplified by inter-center data quantity skew. Center 3 contributed 80 of the 153 training videos, causing global models to disproportionately reflect Center 3's data distribution. The bootstrapped rank stability analysis illustrates the consequence: Task 1 rankings were comparatively stable, and the permutation tests separate the frame-level submissions from the temporally-informed ones, though not the leading methods from one another; Task 2 rankings showed pronounced instability, with no submission retaining its rank on the EC metric in more than 50% of bootstrap replicates.

Degenerate prediction behavior is present across submissions, but its distribution differs from what a purely fine tuning based account would predict. The most pronounced collapse occurs on Task 1, where the global model is evaluated on an unseen center without any local adaptation: Elbflorenz assigned 28 of 29 cases to grade 2 and Camma assigned 26 of 29 to grade 0. Within Task 2, collapse is concentrated at the smaller nodes, most clearly Elbflorenz at Center 2 with 8 of 9 predictions on a single grade, while Center 3, the largest node, is the least degenerate setting for all three submissions. Two distinct mechanisms are therefore at work. Failure to transfer under domain

shift produces collapse in the global model itself, independent of adaptation, and unconstrained fine tuning on small, skewed local datasets produces collapse where local sample counts are lowest. Both inflate nominal F1 without reflecting genuine discrimination. As discussed in Section 4.2, the EC metric is more robust to this artifact and should be considered the primary indicator of adaptation quality in imbalanced ordinal settings.

4.4. Toward personalized federated learning

The results of both tasks point toward the same conclusion: a single globally aggregated model is a compromise that satisfies no center optimally. However, naive local fine-tuning, as evaluated in Task 2, does not resolve this, as unconstrained adaptation on skewed local data leads to catastrophic forgetting rather than genuine personalization, particularly at the smaller nodes. As shown in Section 4.3, this is one of two routes to collapse: the global model itself may already fail to transfer to an unseen center, in which case no amount of constrained adaptation will recover discrimination. The recommendations below therefore address both, a globally federated component that resists center-specific drift, and a locally adapted component that cannot destabilize it. The path forward lies in structured personalized FL: approaches that separate a shared global representation from center-specific adaptation components, and train both with objectives that explicitly account for local data imbalance.

For Task 1 generalization, the FedSurg results support the use of a frozen pretrained temporal backbone as the globally federated component. By fixing backbone weights across federation rounds, center-specific visual variance cannot propagate into the shared representation, preserving its generalizability to unseen centers. Only lightweight task-specific heads need to be aggregated globally, substantially reducing both communication cost and the risk of conflicting updates between heterogeneous nodes.

For Task 2 adaptation, parameter-efficient fine-tuning (PEFT) methods offer a principled alternative to full local fine-tuning. Methods such as LoRA or adapter layers constrain the parameter space available for local updates, limiting the degree to which a small, imbalanced local dataset can destabilize the global representation, while still enabling meaningful center-specific adjustment. This approach is particularly relevant given the data regime of this challenge, where local node datasets ranged from 33 to 80 videos.

Looking further ahead, surgical video foundation models represent a promising direction for initializing the shared backbone (Yang et al., 2025; Schmidgall et al., 2024). However, currently available surgical foundation models are trained on limited procedure-specific data, and general-purpose video foundation models produce feature spaces too broad for the fine-grained visual distinctions required in intraoperative grading tasks. Recent work combining FL with endoscopic foundation model pretraining (Kirchner et al., 2026) suggests a more principled path: federating the pretraining process itself to build privacy-preserving, domain-specific representations from the outset. The integration of such approaches with PEFT-based local adaptation remains a hypothesis for future evaluation rather than a validated recommendation.

The exploratory PEFT baselines included in this work provide a first and partly contrary empirical test for these recommendations. The EndoViT+LoRA variant, which applies LoRA to a frozen frame-level backbone, collapsed to near majority-class prediction (Task 1 F1: 9.24%), consistent with the interpretation that parameter-efficient adaptation cannot compensate for the absence of temporal modeling in this data regime, though the two backbones also differ in pretraining domain (Section 4.1). Under the ordinal cost this variant is nevertheless the cheaper of the two on Task 1 (17.24% against 22.76%), which is what collapse onto a central grade produces, and neither improves on a constant predictor there. The ViViT+LoRA variant, applying LoRA to a frozen temporal backbone, achieved 15.93% F1 on Task 1 and 29.57% F1 at Center 3 in Task 2, approaching the centralized baseline (32.67%) at the largest local node. Performance at smaller centers remained limited (Center 1: 12.22%, Center 2: 22.62%), consistent with the quantity-skew hypothesis: PEFT-based local adaptation is most effective where sufficient local samples are available to support stable fine-tuning.

4.5. Implications for task design

Beyond algorithmic choices, the FedSurg results suggest that the design of the clinical prediction task itself should inform the choice of FL strategy. This dimension has received limited attention in the literature. The consequences of prediction errors differ fundamentally between tasks, and these differences have direct implications for how generalization and adaptation should be balanced.

Consider two contrasting task types. In high-sensitivity settings, for example the detection of malignant lesions in colonoscopy where false negatives carry direct patient harm, global generalization to unseen centers is critical. FL strategies that prioritize robustness across non-IID distributions, such as flat-minima optimization approaches like FedSAM, are preferable over those that optimize local performance at the cost of cross-center generalizability. The fine-grained 6-class inflammation grading evaluated in this challenge approaches the opposite end of the spectrum: overcalling severity could influence clinical decision-making, placing greater weight on specificity. Here, the instability of local adaptation observed in Task 2 is particularly concerning. Classifier collapse, the reversion to a predictor that assigns nearly all cases to a single grade, is not only a performance failure but a specific risk in high specificity grading contexts. As shown in Section 4.3, collapse in this challenge arises from two sources: failure of the global model to transfer to an unseen center, and unconstrained fine tuning on small, imbalanced local datasets. A collapsed model may retain a superficially acceptable nominal score while providing no clinically actionable discrimination, and the direction of collapse determines the clinical consequence: collapse toward a low grade systematically undercalls severity, whereas collapse toward a high grade overcalls it. Neither failure mode is detectable from an aggregate F1-score alone.

These observations lead to a practical hypothesis for future work: FL strategy selection should be guided not only by dataset characteristics, but by an explicit analysis of the clinical error cost structure of the target task. High-sensitivity tasks call for globally robust aggregation; high-specificity tasks require adaptation strategies that prevent classifier collapse, such as the PEFT-based approaches discussed in Section 4.4. Systematic evaluation of this hypothesis across diverse surgical AI tasks remains an important open problem.

4.6. Limitations and future directions

The limitations of this challenge should be read against the state of the field. A recent systematic review of 188 surgical scene understanding studies reports a median of 65 videos and 40 patients per study, with 70.7% of studies relying on a single institution and 59.0% on laparoscopic cholecystectomy (Carstens et al., 2025). The field is also dominated by frame-level instrument and anatomy detection tasks, which are substantially easier than patient-level ordinal grading and on which higher scores are routinely reported. We consider the reporting of weak absolute performance on a clinically meaningful task, measured against chance and constant references, more useful to realistic expectations of clinical AI performance than stronger numbers on tasks of limited clinical relevance would be.

That said, the specific constraints of this edition are real and bear directly on what can be concluded from it. The evaluation rests on 153 training and 70 test videos across four centers, which constrains what it can resolve. Paired permutation tests over the test cases reach significance only for large differences, and no adaptation comparison reaches significance on either metric. Bootstrap confidence intervals on

individual scores span roughly ten percentage points, and the Task 2 rank ordering is not reproducible under resampling. Reported point estimates should therefore be read as indicative rather than as reliable estimates of relative method quality, and the leaderboard is reported for completeness rather than as evidence of submission superiority. No clinically acceptable performance threshold has been established for intraoperative appendicitis grading, so the reported scores cannot be assessed against a target; the six-class ordinal label itself carries substantial ambiguity, with a weighted inter-annotator kappa of 0.614 reported for the complete dataset (Kolbinger et al., 2026), which bounds attainable performance independently of the learning paradigm.

Despite 24 registered teams, only three complete submissions were received, reflecting both the niche character of FL within surgical AI and the high technical barriers to participation. Several aspects of the challenge design likely contributed. The requirement to submit a containerized end-to-end federated pipeline raised the implementation burden substantially relative to a typical segmentation or detection challenge; expertise in federated optimization and in surgical video analysis rarely coincides within a single group; and the public and private data partition, introduced to discourage centralized development, prevented conventional development workflows. Future editions should consider enlarging the development subset or releasing a separate development split, since the guarantee against centralized training derives from the organizers executing the final federated run rather than from withholding data. Providing feedback runs on the complete training data ahead of the final deadline would allow participants to observe federated convergence and tune accordingly.

The three submissions differed in backbone architecture, frame sampling, loss function and aggregation strategy simultaneously. Observed performance differences therefore cannot be attributed to any single design choice, and in particular the contribution of the FL strategy cannot be separated from that of the architecture. The parameter-efficient baselines discussed in Section 4.1 come closest to a controlled comparison here, though their backbones differ in pretraining corpus as well as in temporal modeling, and the organizer-provided baselines are post-hoc analyses rather than competitive entries.

The dataset, as a preliminary subset of Appendix300 (Kolbinger et al., 2026), was insufficient in scale for robust federated training, and the simulated FL setup, where all data resided centrally during evaluation, does not capture the full complexity of real-world distributed deployments, including network latency, node availability, and hardware heterogeneity. A further practical lesson concerns the evaluation infrastructure itself. Frame extraction does not yield an identical frame count for every recording, since the available footage around the annotated timestamp is occasionally shorter than the nominal window. An evaluation harness that filters inputs on an expected frame count will silently discard such cases, which can produce divergent test sets between submission and baseline pipelines without any error being raised. Challenge harnesses should therefore assert explicitly on test set size and composition at evaluation time rather than skipping inputs that fail a format check. This occurred in the present challenge. One of the ten Center 1 test cases was initially skipped in the evaluation of the participant submissions, while the organizer baselines, which sample frames at a fixed rate from the full recording rather than from a fixed index range, were always evaluated on all ten. Predictions for that case were subsequently obtained from all three teams, and every result reported here is computed on the complete test set of 70 cases. The rank ordering is unaffected.

Future work should prioritize scaling the dataset across a broader network of institutions and transitioning from simulated to genuine multi-institutional FL. On the algorithmic side, imbalance-aware objectives, such as federated focal loss or client-weighted aggregation, and personalized aggregation methods that account for quantity skew are necessary next steps. Self-supervised pretraining on large-scale surgical video collections, ideally in a federated manner (Kirchner et al., 2026), offers a promising route toward more transferable representations without compromising data privacy.

5. Conclusion

The FedSurg Challenge establishes the first benchmark for FL in surgical video classification, evaluated here as a proof-of-concept using a multi-center dataset of laparoscopic appendectomies. The results reveal that temporal modeling is the architectural factor most consistently associated with cross-center generalization, though the effect is not uniform across metrics in the controlled baseline comparison, while absolute performance remains far below clinical usability under both centralized and decentralized training, and several configurations fail to improve on constant or chance-level prediction. We do not establish a quantified cost of decentralization: at this sample size the centralized and federated baselines are not statistically separable from one another, and only large differences between methods are resolvable at all. Local fine-tuning without imbalance-aware objectives leads to classifier collapse rather than genuine personalization, as does failure of the global model to transfer to an unseen center, underscoring the need for structured personalized FL with parameter-efficient adaptation. By characterizing these failure modes, and the resolution limits of the evaluation itself, this work provides a methodological reference point for the design of robust, privacy-preserving AI systems for surgical video analysis and motivates future benchmarking on tasks with broader clinical scope.

CRediT authorship contribution statement

Max Kirchner: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Hanna Hoffmann:** Writing – review & editing, Software, Conceptualization. **Alexander C. Jenke:** Writing – review & editing, Software, Conceptualization. **Oliver L. Saldanha:** Writing – review & editing, Software, Investigation, Data curation, Conceptualization. **Kevin Pfeiffer:** Writing – review & editing, Software, Data curation. **Weam Kanjo:** Writing – review & editing, Data curation. **Julia Alekseenko:** Writing – review & editing, Visualization, Software, Methodology, Investigation. **Claas de Boer:** Writing – review & editing, Software, Methodology, Investigation. **Santhi Raj Kolamuri:** Writing – review & editing, Software, Methodology, Investigation. **Lorenzo Mazza:** Writing – review & editing, Software, Methodology, Investigation. **Nicolas Padoy:** Writing – review & editing. **Sophia Bano:** Writing – review & editing, Supervision. **Annika Reinke:** Writing – review & editing, Supervision, Conceptualization. **Lena Maier-Hein:** Writing – review & editing, Supervision. **Danail Stoyanov:** Writing – review & editing, Supervision. **Jakob N. Kather:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Fiona R. Kolbinger:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization. **Sebastian Bodenstedt:** Writing – review & editing, Supervision, Software, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization. **Stefanie Speidel:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT, Gemini 2.5 Pro, Claude Code, and DeepL Write in order to improve the readability and language of the manuscript. After using these tools/services, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

Research Support

This work was co-funded by the European Union through NEARDATA under grant agreement ID 101092644, the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Dresden University of Technology, and the Federal Ministry of Education and Research of Germany in the program of "Souverän. Digital. Vernetzt.", a joint project 6G-life with the project identification number 16KISK001K. Furthermore, **Fiona R. Kolbinger** receives support from the German Cancer Research Center (CoBot 2.0), the Joachim Herz Foundation (Add-On Fellowship for Interdisciplinary Life Science), the Central Indiana Corporate Partnership AnalytiXIN Initiative, the Evan and Sue Ann Werling Pancreatic Cancer Research Fund, and the Indiana Clinical and Translational Sciences Institute (EPAR4157) funded, in part, by Grant Number UM1TR004402 from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Team CAMMA (**Nicolas Padoy, Julia Alekseenko**) was supported by French State Funds managed by the Agence Nationale de la Recherche (ANR) under Grants ANR-22-FAI1-0001 (Project DAIOR) and ANR-10-IAHU-02 (IHU Strasbourg).

Relationships

Jakob N. Kather declares consulting services for Bioptimus, France; Panakeia, UK; AstraZeneca, UK; and MultiplexDx, Slovakia. Furthermore, he holds shares in StratifAI, Germany, Synagen, Germany, Ignition Lab, Germany; has received an institutional research grant by GSK; and has received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. **Fiona R. Kolbinger** declares advisory roles for Radical Healthcare, USA; and the Surgical Data Science Collective, USA, and has received research funding from Novartis. **Stefanie Speidel** received speaker's fees from Stryker Corporation. **Nicolas Padoy** is co-founder and owns shares in Scialytics SAS.

Patents and Intellectual Property

There are no patents to disclose.

Other Activities

Annika Reinke, Nicolas Padoy, Stefanie Speidel, Danail Stoyanov, and Sophia Bano serve as an Associate Editor for the Medical Image Analysis journal.

Acknowledgments

This work was co-funded by the European Union through NEARDATA under grant agreement ID 101092644, the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Dresden University of Technology, and the Federal Ministry of Education and Research of Germany in the program of "Souverän. Digital. Vernetzt.", a joint project 6G-life with the project identification number 16KISK001K.

Furthermore, FRK receives support from the German Cancer Research Center (CoBot 2.0), the Joachim Herz Foundation, Germany (Add-On Fellowship for Interdisciplinary Life Science), the Central Indiana Corporate Partnership AnalytiXIN Initiative, the Evan and Sue Ann Werling Pancreatic Cancer Research Fund, and the Indiana Clinical and Translational Sciences Institute (EPAR4157) funded, in part, by Grant Number UM1TR004402 from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Team Camma (NP, JA) was supported by French State Funds managed by the Agence Nationale de la Recherche (ANR) under Grants ANR-22-FAI1-0001 (Project DAIOR) and ANR-10-IAHU-02 (IHU Strasbourg).

Appendix A. Dataset and annotation protocol

The challenge dataset is a preliminary subset of the publicly available Appendix300 dataset (Kolbinger et al., 2026). A CSV file detailing the samples used in the challenge, along with two recordings that were excluded from the final dataset but used in the challenge, is deposited for reproducibility.[3] Both deposits are available under a CC BY license and require citation of the Appendix300 publication.[4] The annotation protocol is available in Kolbinger et al. (2026). Usage of the data for individual publications was prohibited before the release of this study and the Appendix300 dataset.

Appendix B. Challenge rules

During registration, participants signed the EndoVis rules document.[5]

Appendix C. Challenge organization

This work was jointly organized by four parties. Fiona R. Kolbinger, from the Department of Visceral, Thoracic and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus of the TUD Dresden University of Technology and the Weldon School of Biomedical Engineering, Purdue University coordinated the challenge data collection from four hospitals in Germany. Oliver L. Saldanha and Jakob N. Kather, also from the Else Kröner Fresenius Center for Digital Health, supported the data aggregation from the technical point of view and obtained the source data from these centers. Max Kirchner, Alexander C. Jenke, Sebastian Bodenstedt, and Stefanie Speidel, from the Department of Translational Surgical Oncology (TSO) of the National Center for Tumor Diseases (NCT) Dresden, supported data aggregation and carried out the challenge organization and the technical implementation, including preprocessing, data provision, participant registration and administration, submission handling, evaluation, and results presentation. The NearData Horizon Europe program provided €500 in prize money, equally distributed across both tasks. If a single team achieved the top score in both tasks, they were eligible to receive both awards.

Appendix D. Submission instructions

The submission process was described in detail on the official challenge website.[6] To support development, an example setup based on the FL Flower framework (Beutel et al., 2022) was provided through a GitLab repository.[7] The evaluation framework was made transparent through another repository,[8] which included the source code for metric computation as well as the ranking scripts (bootstrapping, statistical testing, and related plots).

[3] Available at https://doi.org/10.25532/OPARA-1534.
[4] Available at https://doi.org/10.25532/OPARA-1173.
[5] Available at https://tinyurl.com/nhb5z6a3.
[6] Available at https://www.synapse.org/Synapse:syn53137385/wiki/625370.
[7] Available at https://gitlab.com/nct_tso_public/challenges/miccai2024/FedSurg24.
[8] Available at https://gitlab.com/nct_tso_public/challenges/miccai2024/snippet.

Table G.9
Bootstrap rank frequency, win probability, and bootstrap discordance (descriptive) for both tasks (F1-score ↑). Values in the discordance group are computed over 10,000 bootstrap replicates and describe resampling stability. They are not population-level p-values; see Section 2.4.1 and the permutation tests in Table H.24.

Task	Center	Team	Rank Freq.			Win Prob.			Bootstrap discordance (descriptive)		
			1	2	3	Cam	Elb	San	Cam	Elb	San
1	4	Cam	0.0000	0.2491	**0.7509**	NaN	0.2481	0.0000	NaN	0.0000	0.0000
1	4	Elb	0.0077	**0.7442**	0.2481	0.7509	NaN	0.0077	0.0000	NaN	0.0000
1	4	San	**0.9923**	0.0077	0.0000	1.0000	0.9923	NaN	0.0000	0.0000	NaN
2	Avg	Cam	**0.7852**	0.1429	0.0719	NaN	0.8706	0.8427	NaN	0.0000	0.0000
2	Avg	Elb	0.0909	0.3710	**0.5381**	0.1294	NaN	0.4233	0.0000	NaN	1.31×10^{-70}
2	Avg	San	0.1239	**0.4862**	0.3899	0.1573	0.5766	NaN	0.0000	1.31×10^{-70}	NaN

Participants submitted their solutions as Docker containers via the challenge website. Docker Compose scripts with additional code were also accepted, provided they encapsulated the complete FL algorithm for both training and inference. Submissions had to run fully automatically without user interaction. Participants received email notifications about submission status and were allowed unlimited resubmissions until the final deadline. While the organizers do not distribute Docker images, teams were encouraged to release their code publicly.

Appendix E. Challenge design document

See Supplementary file S1.

Appendix F. Ethics approval

This study was prospectively reviewed and approved by the Institutional Review Board of the TUD Dresden University of Technology, Germany (approval number: BO-EK-332072022, approval date: August 4, 2022). The corresponding study was prospectively registered at the German Registry of Clinical Trials (DRKS, URL, registration ID DRKS00030874).

Appendix G. Additional bootstrapping and resampling stability results

The third column group of each table reports a bootstrap discordance value, computed by applying the Wilcoxon signed-rank test to the two methods' metric values across the 10,000 bootstrap replicates. Because the test is applied to resampling replicates rather than to the test cases, its sample size is set by the number of replicates and not by the data, and the resulting values become arbitrarily small for any consistent difference as the number of replicates increases. A small value indicates that one method exceeds the other in nearly every replicate, and a value reported as 0.000 indicates that this holds so consistently that the statistic underflows double precision ($<10^{-308}$); intermediate values such as 8.38×10^{-90} indicate a slightly less consistent ordering, but the difference between such magnitudes carries no interpretable meaning. These values therefore describe how consistently one method exceeds another under resampling of a fixed test set, and are reported alongside the rank frequencies and win probabilities as stability diagnostics rather than as hypothesis tests. Inference over the actual test cases is provided by the paired permutation tests in Table H.24.

Appendix H. Further analysis results with baselines

H.1. Detailed bootstrap analysis tables

Task 1: Generalization (F1-score). Tables H.12–H.14 report rank frequency, pairwise win probabilities, and bootstrap discordance for Task 1 F1-score. Rankings are comparatively stable: the centralized baseline holds rank 1 in 65.7% of bootstrap iterations and Santhi in 32.1%, so those two are not resolved by the resampling. The paired permutation test over the 29 Center-4 cases agrees ($p = 0.671$ for Santhi vs. Cen). It does separate the two lowest-placed submissions from the leaders: Camma and Elbflorenz each differ significantly from Santhi and from Cen ($p \leq 0.047$), and Camma additionally from Swarm ($p = 0.005$), though Elbflorenz does not ($p = 0.122$). Camma and Elbflorenz are indistinguishable from each other ($p = 0.592$).

Task 1: Generalization (Expected cost). Tables H.15–H.17 report the equivalent analysis for the EC metric, which gives the sharpest separation anywhere in the challenge. Camma retains the last rank in 100% of iterations and Elbflorenz the fourth in 98.6%, and the permutation test confirms this: Camma differs from all four other methods ($p = 10^{-4}$) and Elbflorenz from Santhi, Cen and SWARM ($p = 0.013$, 0.004 and 0.039). The leading three are not separated from one another — Santhi, Cen and SWARM exchange the top three positions across replicates, and no permutation comparison among them approaches significance (Santhi vs. Cen $p = 0.638$, Santhi vs. SWARM $p = 1.000$, Cen vs. SWARM $p = 0.357$).

Task 2: Adaptation (F1-score). Tables H.18–H.20 show markedly higher rank volatility compared to Task 1. No method retains its nominal rank in more than 72.3% of iterations (Cen), and among the three participant submissions none exceeds 52.0% (Elbflorenz); Camma, the highest-placed submission on this metric, holds its nominal rank 2 in only 36.7% of iterations and takes rank 1 in 17.8%. The pairwise win probabilities between Camma, Santhi, and SWARM indicate substantial overlap, reflecting the instability of local adaptation across centers. No pairwise permutation comparison on Task 2 F1 reaches significance; the smallest is $p = 0.073$ for Santhi vs. Cen.

Task 2: Adaptation (Expected cost). Tables H.21–H.23 report Task 2 EC bootstrap results, where rank instability is most pronounced. No method retains its nominal rank in even 40% of iterations: the best-retained is Cen's at 37.9%, followed by Camma's at 36.5%, with the remaining three below 26%. Two further observations underline the instability: Camma takes rank 1 in 8.6% of iterations despite placing last on the point estimate, and Elbflorenz's modal rank (5) differs from its nominal rank (4). Win probabilities sit close to 0.5 throughout, the most extreme being 0.756 for Cen over Camma, and no permutation comparison approaches significance (smallest $p = 0.507$, Camma vs. Cen). In particular Santhi and SWARM differ by 0.62 percentage points with a permutation p of 0.844 and heavily overlapping bootstrap intervals. EC-based adaptation rankings are not reproducible under resampling and carry essentially no information about submission superiority.

The ranking summary is provided in Table H.11.

H.2. Paired permutation tests

Table H.24 reports the inferential counterpart to the bootstrap diagnostics above. Twelve of the forty pairwise comparisons reach $p < 0.05$ and all twelve are on Task 1; no Task 2 comparison is significant on either metric. The comparisons among Santhi, Cen and SWARM are not significant on any task or metric, so the top of the ranking is not resolved by this evidence.

Table G.10
Bootstrap rank frequency, win probability, and bootstrap discordance (descriptive) for both tasks (EC ↓). Values in the discordance group are computed over 10,000 bootstrap replicates and describe resampling stability. They are not population-level p-values; see Section 2.4.1 and the permutation tests in Table H.24. At Task 2 both Santhi and Elbflorenz take rank 1 most frequently, in 40.4% and 36.0% of replicates respectively, so the modal ranking does not reproduce the nominal ordering.

Task	Center	Team	Rank Freq.			Win Prob.			Bootstrap discordance (descriptive)		
			1	2	3	Cam	Elb	San	Cam	Elb	San
1	4	Cam	0.0000	0.0000	**1.0000**	NaN	0.0000	0.0000	NaN	0.0000	0.0000
1	4	Elb	0.0027	**0.9973**	0.0000	1.0000	NaN	0.0021	0.0000	NaN	0.0000
1	4	San	**0.9979**	0.0021	0.0000	1.0000	0.9973	NaN	0.0000	0.0000	NaN
2	Avg	Cam	0.2368	0.3142	**0.4490**	NaN	0.4227	0.3640	NaN	8.38×10^{-90}	4.68×10^{-226}
2	Avg	Elb	**0.3600**	0.3230	0.3170	0.5772	NaN	0.4652	8.38×10^{-90}	NaN	4.80×10^{-20}
2	Avg	San	**0.4036**	0.3636	0.2328	0.6350	0.5343	NaN	4.68×10^{-226}	4.80×10^{-20}	NaN

Table H.11
Comprehensive ranking summary for Task 1, Task 2, and overall performance including contextual baselines. Lower rank indicates better performance. The two parameter-efficient baselines are not ranked here; as stated in Section 3.2 they are reported as point estimates only.

Team	Task 1			Task 2			Overall	Final
	EC	F1	Avg	EC	F1	Avg	Avg	Rank
Camma	5	5	5	5	2	3	4.00	3
Elbflorenz	4	4	4	4	5	4	4.00	3
Santhi	2	2	2	3	4	3	2.50	2
Cen	1	1	1	1	1	1	1.00	1
SWARM	3	3	3	2	3	2	2.50	2

Table H.12
Task 1 F1-score ↑: Rank frequency distribution.

Team	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Camma	0.0000	0.0000	0.0001	0.2490	0.7509
Elbflorenz	0.0001	0.0023	0.0126	0.7369	0.2481
Santhi	0.3206	0.4332	0.2397	0.0065	0.0000
Cen	0.6566	0.2953	0.0480	0.0001	0.0000
SWARM	0.0242	0.2689	0.6984	0.0085	0.0000

Table H.13
Task 1 F1-score ↑: Pairwise win probabilities $P(\text{team}_i > \text{team}_j)$.

Team	Camma	Elbflorenz	Santhi	Cen	SWARM
Camma	–	0.2481	0.0000	0.0000	0.0001
Elbflorenz	0.7509	–	0.0077	0.0003	0.0095
Santhi	1.0000	0.9923	–	0.3266	0.7477
Cen	1.0000	0.9997	0.6728	–	0.9339
SWARM	0.9999	0.9905	0.2516	0.0647	–

Table H.14
Task 1 F1-score ↑: Bootstrap discordance (descriptive). Values are computed over 10,000 bootstrap replicates and describe resampling stability. They are not population-level p-values; see Section 2.4.1 and the permutation tests in Table H.24. Every entry underflows double precision ($<10^{-308}$), including pairs the permutation test cannot separate.

Team	Camma	Elbflorenz	Santhi	Cen	SWARM
Camma	–	0.000	0.000	0.000	0.000
Elbflorenz	0.000	–	0.000	0.000	0.000
Santhi	0.000	0.000	–	0.000	0.000
Cen	0.000	0.000	0.000	–	0.000
SWARM	0.000	0.000	0.000	0.000	–

H.3. Reference classifiers

Table H.25 gives the reference levels against which the reported scores should be read. All four references are constructed from training labels only and none uses test prevalence. TrivialGlobal predicts the pooled training mode (grade 3) everywhere; TrivialLocal predicts each center's own training mode (grade 2 at Center 1, grade 4 at Center 2, grade 3 at Center 3). TrivialLocal is undefined for Task 1, whose test center is held out and contributes no training labels of its own. RandomUniform draws each of the six grades with probability 1/6;

Table H.15
Task 1 EC ↓: Rank frequency distribution. Santhi's modal rank (3) differs from its nominal rank (2).

Team	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Camma	0.0000	0.0000	0.0000	0.0000	1.0000
Elbflorenz	0.0002	0.0011	0.0124	0.9863	0.0000
Santhi	0.2619	0.3560	0.3810	0.0011	0.0000
Cen	0.7452	0.2309	0.0238	0.0001	0.0000
SWARM	0.0890	0.3882	0.5132	0.0096	0.0000

Table H.16
Task 1 EC ↓: Pairwise win probabilities.

Team	Camma	Elbflorenz	Santhi	Cen	SWARM
Camma	–	0.0000	0.0000	0.0000	0.0000
Elbflorenz	1.0000	–	0.0021	0.0002	0.0097
Santhi	1.0000	0.9973	–	0.2094	0.5482
Cen	1.0000	0.9996	0.7287	–	0.8855
SWARM	1.0000	0.9879	0.3905	0.0692	–

Table H.17
Task 1 EC ↓: Bootstrap discordance (descriptive). Values are computed over 10,000 bootstrap replicates and describe resampling stability. They are not population-level p-values; see Section 2.4.1 and the permutation tests in Table H.24. Santhi and SWARM receive 1.40×10^{-95} here, while the permutation test over the 29 test cases returns $p = 1.000$ for that pair.

Team	Camma	Elbflorenz	Santhi	Cen	SWARM
Camma	–	0.000	0.000	0.000	0.000
Elbflorenz	0.000	–	0.000	0.000	0.000
Santhi	0.000	0.000	–	0.000	1.40×10^{-95}
Cen	0.000	0.000	0.000	–	0.000
SWARM	0.000	0.000	1.40×10^{-95}	0.000	–

Table H.18
Task 2 F1-score ↑: Rank frequency distribution.

Team	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Camma	0.1776	0.3672	0.2812	0.1083	0.0657
Elbflorenz	0.0063	0.0395	0.0991	0.3351	0.5200
Santhi	0.0048	0.0593	0.1353	0.4257	0.3749
Cen	0.7232	0.1820	0.0792	0.0125	0.0031
SWARM	0.0881	0.3521	0.4051	0.1185	0.0362

Table H.19
Task 2 F1-score ↑: Pairwise win probabilities.

Team	Camma	Elbflorenz	Santhi	Cen	SWARM
Camma	–	0.8706	0.8427	0.2094	0.5600
Elbflorenz	0.1294	–	0.4233	0.0183	0.1059
Santhi	0.1573	0.5766	–	0.0152	0.1441
Cen	0.7906	0.9817	0.9848	–	0.8526
SWARM	0.4400	0.8941	0.8558	0.1474	–

Table H.20
Task 2 F1-score ↑: Bootstrap discordance (descriptive). Values are computed over 10,000 bootstrap replicates and describe resampling stability. They are not population-level p-values; see Section 2.4.1 and the permutation tests in Table H.24. No Task 2 F1 comparison reaches $p < 0.05$ under the permutation test.

Team	Camma	Elbflorenz	Santhi	Cen	SWARM
Camma	–	0.000	0.000	0.000	3.71×10^{-42}
Elbflorenz	0.000	–	1.31×10^{-70}	0.000	0.000
Santhi	0.000	1.31×10^{-70}	–	0.000	0.000
Cen	0.000	0.000	0.000	–	0.000
SWARM	3.71×10^{-42}	0.000	0.000	0.000	–

Table H.21
Task 2 EC ↓: Rank frequency distribution. Elbflorenz's modal rank (5) differs from its nominal rank (4).

Team	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Camma	0.0861	0.1395	0.1716	0.2378	0.3650
Elbflorenz	0.1535	0.1737	0.1965	0.2276	0.2487
Santhi	0.1416	0.2055	0.2512	0.2467	0.1550
Cen	0.3794	0.2267	0.1714	0.1283	0.0942
SWARM	0.2414	0.2540	0.2090	0.1602	0.1354

Table H.22
Task 2 EC ↓: Pairwise win probabilities.

Team	Camma	Elbflorenz	Santhi	Cen	SWARM
Camma	–	0.4227	0.3640	0.2440	0.3114
Elbflorenz	0.5772	–	0.4652	0.3275	0.3843
Santhi	0.6350	0.5343	–	0.3513	0.4074
Cen	0.7557	0.6721	0.6473	–	0.5911
SWARM	0.6882	0.6152	0.5915	0.4084	–

Table H.23
Task 2 EC ↓: Bootstrap discordance (descriptive). Values are computed over 10,000 bootstrap replicates and describe resampling stability. They are not population-level p-values; see Section 2.4.1 and the permutation tests in Table H.24. Every entry here is numerically extreme, yet no Task 2 EC comparison reaches $p < 0.05$ under the permutation test (smallest $p = 0.507$).

Team	Camma	Elbflorenz	Santhi	Cen	SWARM
Camma	–	8.38×10^{-90}	4.68×10^{-226}	0.000	0.000
Elbflorenz	8.38×10^{-90}	–	4.80×10^{-20}	0.000	5.23×10^{-164}
Santhi	4.68×10^{-226}	4.80×10^{-20}	–	2.56×10^{-272}	9.30×10^{-95}
Cen	0.000	0.000	2.56×10^{-272}	–	1.26×10^{-89}
SWARM	0.000	5.23×10^{-164}	9.30×10^{-95}	1.26×10^{-89}	–

Table H.24
Paired permutation tests over the actual test cases. The exchangeable unit is the individual case: for each permutation the two methods' predictions are exchanged independently per case, with 10,000 permutations and a fixed seed. For Task 2 the exchange is performed within each center, the metric is recomputed per center, and the per-center values are then averaged with equal weight across centers. Sample sizes are $n = 29$ cases for Task 1 (Center 4) and $n = 41$ cases for Task 2 (10, 9 and 22 cases at Centers 1, 2 and 3). Observed differences are method A minus method B in percentage points; for F1 ↑ a positive value favors A, for EC ↓ a negative value favors A.

Task	Metric	Comparison (A vs. B)	Obs. diff. (pp)	Permutation p
1	F1 ↑	Camma vs. Elbflorenz[a]	−3.07	0.5919
1	F1 ↑	Camma vs. Santhi[a]	−18.27	**0.0036**
1	F1 ↑	Camma vs. Cen	−21.54	**0.0005**
1	F1 ↑	Camma vs. SWARM	−14.24	**0.0052**
1	F1 ↑	Elbflorenz vs. Santhi[a]	−15.20	**0.0465**
1	F1 ↑	Elbflorenz vs. Cen	−18.48	**0.0084**
1	F1 ↑	Elbflorenz vs. SWARM	−11.17	0.1221
1	F1 ↑	Santhi vs. Cen	−3.27	0.6709
1	F1 ↑	Santhi vs. SWARM	+4.03	0.5876
1	F1 ↑	Cen vs. SWARM	+7.31	0.2429
1	EC ↓	Camma vs. Elbflorenz[a]	+33.10	**0.0001**
1	EC ↓	Camma vs. Santhi[a]	+44.83	**0.0001**
1	EC ↓	Camma vs. Cen	+46.90	**0.0001**
1	EC ↓	Camma vs. SWARM	+44.14	**0.0001**
1	EC ↓	Elbflorenz vs. Santhi[a]	+11.72	**0.0131**
1	EC ↓	Elbflorenz vs. Cen	+13.79	**0.0035**
1	EC ↓	Elbflorenz vs. SWARM	+11.03	**0.0390**

(*continued on next page*)

Table H.24 (*continued*).

Task	Metric	Comparison (*A* vs. *B*)	Obs. diff. (pp)	Permutation *p*
1	EC ↓	Santhi vs. Cen	+2.07	0.6380
1	EC ↓	Santhi vs. SWARM	−0.69	1.0000
1	EC ↓	Cen vs. SWARM	−2.76	0.3566
2	F1 ↑	Camma vs. Elbflorenz[a]	+6.58	0.4208
2	F1 ↑	Camma vs. Santhi[a]	+5.39	0.4200
2	F1 ↑	Camma vs. Cen	−4.72	0.4256
2	F1 ↑	Camma vs. SWARM	+1.02	0.8444
2	F1 ↑	Elbflorenz vs. Santhi[a]	−1.18	0.8218
2	F1 ↑	Elbflorenz vs. Cen	−11.30	0.1161
2	F1 ↑	Elbflorenz vs. SWARM	−5.56	0.4024
2	F1 ↑	Santhi vs. Cen	−10.11	0.0733
2	F1 ↑	Santhi vs. SWARM	−4.37	0.4060
2	F1 ↑	Cen vs. SWARM	+5.74	0.3124
2	EC ↓	Camma vs. Elbflorenz[a]	+0.81	0.8355
2	EC ↓	Camma vs. Santhi[a]	+1.20	0.7431
2	EC ↓	Camma vs. Cen	+2.41	0.5068
2	EC ↓	Camma vs. SWARM	+1.82	0.6513
2	EC ↓	Elbflorenz vs. Santhi[a]	+0.39	0.9069
2	EC ↓	Elbflorenz vs. Cen	+1.60	0.6732
2	EC ↓	Elbflorenz vs. SWARM	+1.01	0.7845
2	EC ↓	Santhi vs. Cen	+1.21	0.7313
2	EC ↓	Santhi vs. SWARM	+0.62	0.8444
2	EC ↓	Cen vs. SWARM	−0.59	0.8701

[a] Rows constitute the participants-only set of Tables G.9 and G.10; a pair's permutation result does not depend on which other methods are included. Significant comparisons ($p < 0.05$) in bold.

Table H.25
Reference classifier levels, in percent. F1 ↑, EC ↓. Constant references are point estimates; stochastic references are simulation means with 95% percentile intervals in brackets. Task 2 average is the equal-weight mean over Centers 1–3. At Center 2 the edge position of grade 4 inflates the linear cost of TrivialLocal, so comparisons there should be made against TrivialGlobal.

Constant references (point estimates)

Test set	TrivialGlobal			TrivialLocal		
	Grade	F1 ↑	EC ↓	Grade	F1 ↑	EC ↓
Task 1 Center 4	3	9.17	13.79	n/a	n/a	n/a
Task 2 Center 1	3	3.03	28.00	2	9.52	20.00
Task 2 Center 2	3	6.06	24.44	4	8.33	31.11
Task 2 Center 3	3	9.68	18.18	3[a]	9.68	18.18
Task 2 average	–	6.26	23.54	–	9.18	23.10

Stochastic references (simulation mean [95% percentile interval])

Test set	RandomUniform		RandomPrior	
	F1 ↑	EC ↓	F1 ↑	EC ↓
Task 1 Center 4	11.91 [2.4–23.9]	33.55 [25.5–42.1]	14.30 [5.5–26.1]	25.52 [18.6–33.1]
Task 2 Center 1	12.29 [0.0–34.7]	36.04 [20.0–52.0]	14.78 [0.0–36.9]	29.11 [16.0–42.0]
Task 2 Center 2	12.47 [0.0–36.1]	35.84 [20.0–53.3]	13.80 [0.0–34.4]	29.80 [15.6–44.4]
Task 2 Center 3	13.83 [2.4–30.4]	34.20 [24.5–44.5]	15.37 [5.0–30.1]	24.65 [17.3–31.8]
Task 2 average	12.86 [4.0–24.3]	35.36 [27.2–43.7]	14.65 [6.3–25.2]	27.85 [21.1–34.5]

[a] At Center 3 the two constant references coincide, because that center's training mode and the pooled training mode are both grade 3.

Table H.26
The twenty comparisons in which a method falls below a chance classifier on a given test set and metric. These cover eleven distinct method, test set and metric combinations, since a result may fall below both stochastic references, and the Task 2 average rows are not independent of the Centers 1 to 3 rows. The two parameter-efficient baselines are not included, as stated in Section 3.2. Values in percent; the chance level is the simulation mean from Table H.25. For F1 ↑ a lower value is worse, for EC ↓ a higher value is worse.

Test set	Method	Falls below	Method value	Chance level
Macro F1 ↑				
Task 2 Center 1	Camma	RandomUniform	3.33	12.29
Task 2 Center 1	Cen	RandomUniform	11.43	12.29
Task 2 Center 1	SWARM	RandomUniform	10.32	12.29
Task 2 Center 1	Camma	RandomPrior	3.33	14.78
Task 2 Center 1	Cen	RandomPrior	11.43	14.78
Task 2 Center 1	SWARM	RandomPrior	10.32	14.78

(*continued on next page*)

RandomPrior samples the training prior (pooled for Task 1, the center's own for Task 2). The two constant references are evaluated with 10,000 paired bootstrap replicates; the two stochastic references are simulated over 10,000 draws with a fresh prediction vector per draw, so they are reported with percentile intervals and are never passed through the ranking or win-probability machinery.

Table H.26 (*continued*).

Test set	Method	Falls below	Method value	Chance level
Macro F1 ↑				
Task 2 Center 2	Elbflorenz	RandomUniform	3.70	12.47
Task 2 Center 2	Elbflorenz	RandomPrior	3.70	13.80
Task 2 Center 2	Santhi	RandomPrior	13.33	13.80
Task 2 Center 3	Santhi	RandomUniform	12.04	13.83
Task 2 Center 3	Santhi	RandomPrior	12.04	15.37
Task 2 average	Elbflorenz	RandomUniform	12.21	12.86
Task 2 average	Elbflorenz	RandomPrior	12.21	14.65
Task 2 average	Santhi	RandomPrior	13.40	14.65
Task 1 Center 4	Camma	RandomUniform	4.76	11.91
Task 1 Center 4	Elbflorenz	RandomUniform	7.83	11.91
Task 1 Center 4	Camma	RandomPrior	4.76	14.30
Task 1 Center 4	Elbflorenz	RandomPrior	7.83	14.30
Expected cost ↓				
Task 1 Center 4	Camma	RandomUniform	57.24	33.55
Task 1 Center 4	Camma	RandomPrior	57.24	25.52

These references are floors rather than competitive levels, and several reported results fall below them. At Center 1 the constant TrivialLocal classifier attains an expected cost of 20.00%, better than five of the seven evaluated methods; only Elbflorenz (18.00%) and the EndoViT+LoRA baseline (17.78%) improve on it. At Center 3 no participant submission reaches the 18.18% expected cost of the constant classifier. Table H.26 lists the twenty comparisons in which a method falls below a chance classifier, covering eleven distinct method, test set and metric combinations.

Appendix I. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2026.104290.

Data availability

The Appendix300 dataset is available at https://doi.org/10.25532/OPARA-1173 (CC BY). Evaluation code is on GitLab. The challenge subset specification is available at https://doi.org/10.25532/OPARA-1534 (CC BY).

References

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C., 2021. ViViT: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6836–6846, URL https://openaccess.thecvf.com/content/ICCV2021/html/Arnab_ViViT_A_Video_Vision_Transformer_ICCV_2021_paper.html?ref=https://githubhelp.com.

Batić, D., Holm, F., Özsoy, E., Czempiel, T., Navab, N., 2024. EndoViT: pretraining vision transformers on a large collection of endoscopic images. Int. J. Comput. Assist. Radiol. Surg. 19 (6), 1085–1091. http://dx.doi.org/10.1007/s11548-024-03091-5.

Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K.H., Parcollet, T., Gusmão, P.P.B.d., Lane, N.D., 2022. Flower: A friendly federated learning research framework. http://dx.doi.org/10.48550/arXiv.2007.14390, arXiv:2007.14390 [cs]. URL http://arxiv.org/abs/2007.14390.

Brandenburg, J.M., Jenke, A.C., Stern, A., Daum, M.T.J., Schulze, A., Younis, R., Petrynowski, P., Davitashvili, T., Vanat, V., Bhasker, N., Schneider, S., Mündermann, L., Reinke, A., Kolbinger, F.R., Jörns, V., Fritz-Kebede, F., Dugas, M., Maier-Hein, L., Klotz, R., Distler, M., Weitz, J., Müller-Stich, B.P., Speidel, S., Bodenstedt, S., Wagner, M., 2023. Active learning for extracting surgomic features in robot-assisted minimally invasive esophagectomy: a prospective annotation study. Surg. Endosc. 37 (11), 8577–8593. http://dx.doi.org/10.1007/s00464-023-10447-6.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1993. Signature verification using a "Siamese" time delay neural network. In: Advances in Neural Information Processing Systems. Vol. 6, Morgan-Kaufmann, URL https://proceedings.neurips.cc/paper/1993/hash/288cc0ff022877bd3df94bc9360b9c5d-Abstract.html.

Caldarola, D., Caputo, B., Ciccone, M., 2022. Improving generalization in federated learning by seeking flat minima. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022. Springer Nature Switzerland, Cham, pp. 654–672. http://dx.doi.org/10.1007/978-3-031-20050-2_38.

Carstens, M., Vasisht, S., Zhang, Z., Barbur, I., Reinke, A., Maier-Hein, L., Hashimoto, D.A., Kolbinger, F.R., 2025. Artificial intelligence for surgical scene understanding: a systematic review and reporting quality meta-analysis. Npj Digit. Med. 9 (1), 59. http://dx.doi.org/10.1038/s41746-025-02227-4.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302. http://dx.doi.org/10.2307/1932409, _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.2307/1932409. URL https://onlinelibrary.wiley.com/doi/abs/10.2307/1932409.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. http://dx.doi.org/10.48550/arXiv.2010.11929, arXiv:2010.11929 [cs]. URL http://arxiv.org/abs/2010.11929.

Efron, B., 1992. Bootstrap methods: another look at the jackknife. In: Breakthroughs in Statistics: Methodology and Distribution. Springer, pp. 569–593.

Elkan, C., 2001. The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence. IJCAI, pp. 973–978.

Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B., 2021. Sharpness-aware minimization for efficiently improving generalization. http://dx.doi.org/10.48550/arXiv.2010.01412, arXiv:2010.01412 [cs]. URL http://arxiv.org/abs/2010.01412.

2016. General data protection regulation (GDPR) – Legal text. URL https://gdpr-info.eu/.

Gomes, C.A., Nunes, T.A., Fonseca Chebli, J.M., Junior, C.S., Gomes, C.C., 2012. Laparoscopy grading system of acute appendicitis: new insight for future trials. Surg. Laparosc. Endosc. Percutan. Tech. 22 (5), 463–466. http://dx.doi.org/10.1097/SLE.0b013e318262edf1.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2, Springer.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. http://dx.doi.org/10.48550/arXiv.1512.03385, arXiv:1512.03385 [cs]. URL http://arxiv.org/abs/1512.03385.

Hewlett Packard Enterprise, 2021. Swarm learning. https://github.com/HewlettPackard/swarm-learning.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations.

Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R.G.L., Eichner, H., Rouayheb, S.E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P.B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S.U., Sun, Z., Suresh, A.T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F.X., Yu, H., Zhao, S., 2021. Advances and open problems in federated learning. http://dx.doi.org/10.48550/arXiv.1912.04977, arXiv:1912.04977 [cs]. URL http://arxiv.org/abs/1912.04977.

Kassem, H., Alapatt, D., Mascagni, P., AI4SafeChole, C., Karargyris, A., Padoy, N., 2022. Federated cycling (FedCy): Semi-supervised federated learning of surgical phases. IEEE Trans. Med. Imaging 1. http://dx.doi.org/10.1109/TMI.2022.3222126, Conference Name: IEEE Transactions on Medical Imaging.

Kirchner, M., Jenke, A.C., Bodenstedt, S., Kolbinger, F.R., Saldanha, O.L., Kather, J.N., Wagner, M., Speidel, S., 2026. Federated EndoViT: Pretraining vision transformers via federated learning on endoscopic image collections. In: Huo, Y., Gao, M., Kuo, C.-F., Jin, Y., Deng, R. (Eds.), Proceedings of the 9th International Conference on Medical Imaging with Deep Learning. In: Proceedings of Machine Learning

Research, vol. 315, PMLR, pp. 1903–1934, URL https://proceedings.mlr.press/v315/kirchner26a.html.

Kirtac, K., Aydin, N., Lavanchy, J., Beldi, G., Smit, M., Woods, M., Aspart, F., 2022. Surgical phase recognition: From public datasets to real-world data. Appl. Sci. 12, 8746. http://dx.doi.org/10.3390/app12178746.

Kolbinger, F.R., Kirchner, M., Pfeiffer, K., Bodenstedt, S., Jenke, A.C., Barthel, J., Carstens, M., Dehlke, K., Dietz, S., Emmanouilidis, S., Fitze, G., Freitag, M., Holderried, F., Jacobi, T., Kanjo, W., Leitermann, L., Mees, S.T., Pistorius, S., Prudlo, C., Seiberth, A., Schultz, J., Thiel, K., Ziehn, D., Speidel, S., Weitz, J., Kather, J.N., Distler, M., Saldanha, O.L., 2026. Appendix300: Surgical video and patient metadata of 330 laparoscopic appendectomy cases from five institutions. Sci. Data 13 (1), 886. http://dx.doi.org/10.1038/s41597-026-07593-6.

Lavanchy, J.L., Ramesh, S., Dall'Alba, D., Gonzalez, C., Fiorini, P., Muller-Stich, B., Nett, P.C., Marescaux, J., Mutter, D., Padoy, N., 2023. Challenges in multi-centric generalization: Phase and step recognition in Roux-en-Y gastric bypass surgery. http://dx.doi.org/10.48550/arXiv.2312.11250, arXiv:2312.11250 [cs]. URL http://arxiv.org/abs/2312.11250.

Li, Y., Kundu, S.S., Boels, M., Mahmoodi, T., Ourselin, S., Vercauteren, T., Dasgupta, P., Shapey, J., Granados, A., 2025. UltraFlwr – An efficient federated medical and surgical object detection framework. http://dx.doi.org/10.48550/arXiv.2503.15161, arXiv:2503.15161 [cs]. URL http://arxiv.org/abs/2503.15161.

Li, T., Sahu, A.K., Talwalkar, A., Smith, V., 2020a. Federated learning: Challenges, methods, and future directions. IEEE Signal Process. Mag. 37 (3), 50–60. http://dx.doi.org/10.1109/MSP.2020.2975749, URL https://ieeexplore.ieee.org/document/9084352.

Li, T., Sanjabi, M., Beirami, A., Smith, V., 2020b. Fair resource allocation in federated learning. http://dx.doi.org/10.48550/arXiv.1905.10497, arXiv:1905.10497 [cs]. URL http://arxiv.org/abs/1905.10497.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Luo, P., Zhang, R., Ren, J., Peng, Z., Li, J., 2021. Switchable normalization for learning-to-normalize deep representation. IEEE Trans. Pattern Anal. Mach. Intell. 43 (2), 712–728. http://dx.doi.org/10.1109/TPAMI.2019.2932062, URL https://ieeexplore.ieee.org/abstract/document/8781758.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., Feldmann, C., Frangi, A.F., Full, P.M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B.A., März, K., Maier, O., Maier-Hein, K., Menze, B.H., Müller, H., Neher, P.F., Niessen, W., Rajpoot, N., Sharp, G.C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A.A., van der Sommen, F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., Kopp-Schneider, A., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. 9 (1), 5217. http://dx.doi.org/10.1038/s41467-018-07619-7, Publisher: Nature Publishing Group. URL https://www.nature.com/articles/s41467-018-07619-7.

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., Nakawala, H., Park, A., Pugh, C., Stoyanov, D., Vedula, S.S., Cleary, K., Fichtinger, G., Forestier, G., Gibaud, B., Grantcharov, T., Hashizume, M., Heckmann-Nötzel, D., Kenngott, H.G., Kikinis, R., Mündermann, L., Navab, N., Onogur, S., Roß, T., Sznitman, R., Taylor, R.H., Tizabi, M.D., Wagner, M., Hager, G.D., Neumuth, T., Padoy, N., Collins, J., Gockel, I., Goedeke, J., Hashimoto, D.A., Joyeux, L., Lam, K., Leff, D.R., Madani, A., Marcus, H.J., Meireles, O., Seitel, A., Teber, D., Ückert, F., Müller-Stich, B.P., Jannin, P., Speidel, S., 2022. Surgical data science – from concepts toward clinical translation. Med. Image Anal. 76, 102306. http://dx.doi.org/10.1016/j.media.2021.102306, URL https://www.sciencedirect.com/science/article/pii/S1361841521003510.

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., Wiesenfarth, M., Kavur, A.E., Sudre, C.H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M.B., Cardoso, M.J., Cheplygina, V., Cimini, B.A., Collins, G.S., Farahani, K., Ferrer, L., Galdran, A., Van Ginneken, B., Haase, R., Hashimoto, D.A., Hoffman, M.M., Huisman, M., Jannin, P., Kahn, C.E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B.A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A.L., Mattson, P., Meijering, E., Menze, B., Moons, K.G.M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C.I., Shetty, S., Van Smeden, M., Summers, R.M., Taha, A.A., Tiulpin, A., Tsaftaris, S.A., Van Calster, B., Varoquaux, G., Jäger, P.F., 2024. Metrics reloaded: recommendations for image analysis validation. Nature Methods 21 (2), 195–212. http://dx.doi.org/10.1038/s41592-023-02151-z, URL https://www.nature.com/articles/s41592-023-02151-z.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., 2020. BIAS: Transparent reporting of biomedical image analysis challenges. Med. Image Anal. 66, 101796. http://dx.doi.org/10.1016/j.media.2020.101796, URL https://www.sciencedirect.com/science/article/pii/S1361841520301602.

Maier-Hein, L., Vedula, S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., Hashizume, M., Katić, D., Kenngott, H., Kranzfelder, M., Malpani, A., März, K., Neumuth, T., Padoy, N., Pugh, C., Jannin, P., 2017. Surgical data science for next-generation interventions. Nat. Biomed. Eng. 1, http://dx.doi.org/10.1038/s41551-017-0132-7.

McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y., 2017. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1273–1282, ISSN: 2640-3498. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Rauniyar, A., Hagos, D.H., Jha, D., Håkegård, J.E., Bagci, U., Rawat, D.B., Vlassov, V., 2023. Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. http://dx.doi.org/10.48550/arXiv.2208.03392, arXiv:2208.03392 [cs]. URL http://arxiv.org/abs/2208.03392.

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., McMahan, H.B., 2021. Adaptive federated optimization. http://dx.doi.org/10.48550/arXiv.2003.00295, version: 5. arXiv:2003.00295 [cs]. URL http://arxiv.org/abs/2003.00295.

Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R.M., Trask, A., Xu, D., Baust, M., Cardoso, M.J., 2020. The future of digital health with federated learning. Npj Digit. Med. 3 (1), 119. http://dx.doi.org/10.1038/s41746-020-00323-1, Publisher: Nature Publishing Group. URL https://www.nature.com/articles/s41746-020-00323-1.

Rights (OCR), O.f.C., 2021. Health information privacy. Last Modified: 2025-06-27T11:38:47-0400. URL https://www.hhs.gov/hipaa/index.html.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Saldanha, O.L., Pfeiffer, K., Bodenstedt, S., Kirchner, M., Jenke, A.C., Barata, C., Barbosa, S., Barthel, J., Carstens, M., Castro, L.T., Dehlke, K., Dietz, S., Emmanouilidis, S., Fitze, G., Holderried, F., Kanjo, W., Leitermann, L., Mees, S.T., Soares, A.S., Pascoal, M., Pistorius, S., Prudlo, C., Schultz, J., Seiberth, A., Thiel, K., Wu, X., Ziehn, D., Speidel, S., Weitz, J., Distler, M., Kather, J.N., Kolbinger, F.R., 2026. Privacy-preserving surgical video analysis with swarm learning — Results from a multinational appendectomy cohort. NEJM AI 3 (7), AIoa2501116. http://dx.doi.org/10.1056/AIoa2501116, URL https://ai.nejm.org/doi/full/10.1056/AIoa2501116.

Schmidgall, S., Kim, J.W., Jopling, J., Krieger, A., 2024. General surgery vision transformer: A video pre-trained foundation model for general surgery. http://dx.doi.org/10.48550/arXiv.2403.05949, arXiv:2403.05949 [cs]. URL http://arxiv.org/abs/2403.05949.

Speidel, S., Maier-Hein, L., Stoyanov, D., Bodenstedt, S., Reinke, A., Bano, S., 2024. Endoscopic vision challenge – A MICCAI challenge. URL https://opencas.dkfz.de/endovis/.

Tan, A.Z., Yu, H., Cui, L., Yang, Q., 2023. Towards personalized federated learning. IEEE Trans. Neural Networks Learn. Syst. 34 (12), 9587–9603. http://dx.doi.org/10.1109/TNNLS.2022.3160699.

Tomar, S., 2006. Converting video formats with FFmpeg. Linux J. 2006 (146), 10.

Warnat-Herresthal, S., Schultze, H., Shastry, K.L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N.A., et al., 2021. Swarm learning for decentralized and confidential clinical machine learning. Nature 594, 265–270.

Yang, S., Zhou, F., Mayer, L., Huang, F., Chen, Y., Wang, Y., He, S., Nie, Y., Wang, X., Sümer, Ö., Jin, Y., Sun, H., Xu, S., Liu, A.Q., Li, Z., Qin, J., Teoh, J.Y., Maier-Hein, L., Chen, H., 2025. Large-scale self-supervised video foundation model for intelligent surgery. http://dx.doi.org/10.48550/arXiv.2506.02692, arXiv:2506.02692 [cs]. URL http://arxiv.org/abs/2506.02692.

Yin, D., Chen, Y., Kannan, R., Bartlett, P., 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In: Proceedings of the 35th International Conference on Machine Learning. PMLR, pp. 5650–5659, ISSN: 2640-3498. URL https://proceedings.mlr.press/v80/yin18a.html.