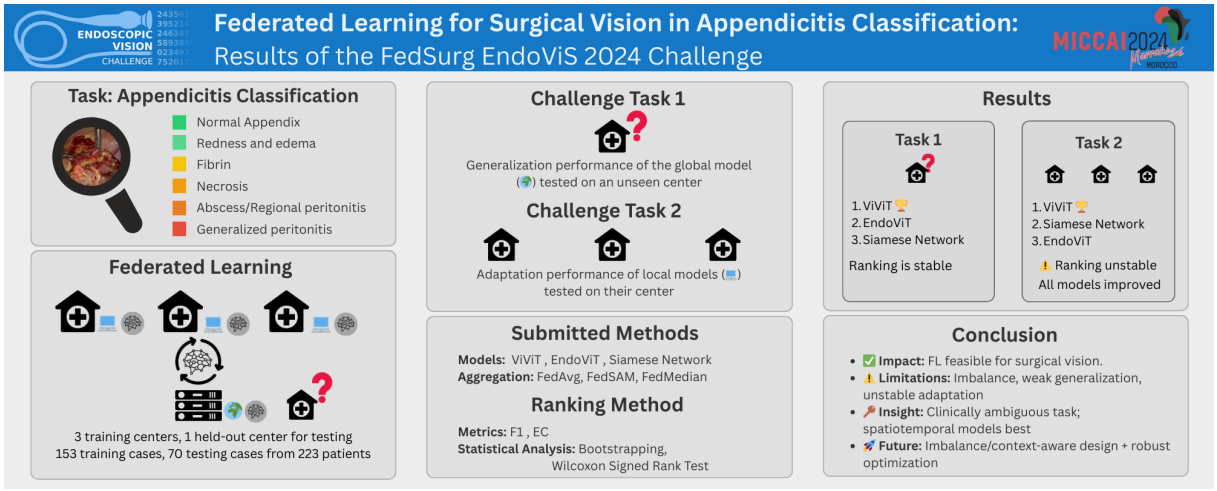
Graphical Abstract

Federated Learning for Surgical Vision in Appendicitis Classification: Results of the FedSurg EndoVis 2024 Challenge

Max Kirchner, Hanna Hoffmann, Alexander C. Jenke, Oliver L. Saldanha, Kevin Pfeiffer, Weam Kanjo, Julia Alekseenko, Claas de Boer, Santhi Raj Kolamuri, Lorenzo Mazza, Nicolas Padoy, Sophia Bano, Annika Reinke, Lena Maier-Hein, Danail Stoyanov, Jakob N. Kather, Fiona R. Kolbinger, Sebastian Bodenstedt, Stefanie Speidel



Highlights

Federated Learning for Surgical Vision in Appendicitis Classification: Results of the FedSurg EndoVis 2024 Challenge

Max Kirchner, Hanna Hoffmann, Alexander C. Jenke, Oliver L. Saldanha, Kevin Pfeiffer, Weam Kanjo, Julia Alekseenko, Claas de Boer, Santhi Raj Kolamuri, Lorenzo Mazza, Nicolas Padoy, Sophia Bano, Annika Reinke, Lena Maier-Hein, Danail Stoyanov, Jakob N. Kather, Fiona R. Kolbinger, Sebastian Bodenstedt, Stefanie Speidel

- First federated learning challenge in surgical AI
- First use of preliminary multi-center Appendix300 dataset
- Novel patient-level video classification beyond image-based tasks
- Findings on generalization vs. personalization in federated learning
- Demonstrated strengths and limitations of federated surgical AI

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*

Abstract

Purpose: The FedSurg challenge was designed to benchmark the state of the art in federated learning for surgical video classification. Its goal was to assess how well current methods generalize to unseen clinical centers and adapt through local fine-tuning while enabling collaborative model development without sharing patient data.

Methods: Participants developed strategies to classify inflammation stages in appendicitis using a preliminary version of the multi-center Appendix300 video dataset. The

*Corresponding author

Email address: max.kirchner@nct-dresden.de (Max Kirchner)

challenge evaluated two tasks: generalization to an unseen center and center-specific adaptation after fine-tuning. Submitted approaches included foundation models with linear probing, metric learning with triplet loss, and various FL aggregation schemes (FedAvg, FedMedian, FedSAM). Performance was assessed using F1-score and Expected Cost, with ranking robustness evaluated via bootstrapping and statistical testing.

Results: In the generalization task, performance across centers was limited. In the adaptation task, all teams improved after fine-tuning, though ranking stability was low. The ViViT-based submission achieved the strongest overall performance. The challenge highlighted limitations in generalization, sensitivity to class imbalance, and difficulties in hyperparameter tuning in decentralized training, while spatiotemporal modeling and context-aware preprocessing emerged as promising strategies.

Conclusion: The FedSurg Challenge establishes the first benchmark for evaluating FL strategies in surgical video classification. Findings highlight the trade-off between local personalization and global robustness, and underscore the importance of architecture choice, preprocessing, and loss design. This benchmarking offers a reference point for future development of imbalance-aware, adaptive, and robust FL methods in clinical surgical AI.

Keywords: Federated Learning, EndoViS Challenge, Appendectomy, Video Classification, Surgical Data Science

---

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 [1]. Recent work has demonstrated that AI can reliably analyze surgical video, comprehending anatomy, tool usage, and procedural events in real time [2]. 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 [3].

Surgical datasets are typically sourced from either a single institution or a small group of collaborators [3, 4]. While these datasets enable preliminary developments, they inherently suffer from limited diversity and scale if applied in real-world scenarios [5]. Single-center datasets often lack the heterogeneity required for generalizable AI models and involve prolonged data acquisition cycles [6]. 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) [7] or the General Data Protection Regulation (GDPR) [8] 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 [9]. 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) [9] or more robust alternatives such as FedMedian [10]. 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 [11]. It enables large-scale collaborative model development without the need to centralize sensitive, multi-institutional data, thereby preserving patient confidentiality and regulatory compliance [11]. However, FL also introduces new challenges. These include handling data and system heterogeneity across institutions [12, 13, 14], balancing personalization with generalization [15], managing communication overhead due to frequent model updates [9, 13], addressing fairness and potential model biases [13, 16], and coping with the increased complexity of evaluation and debugging in distributed settings [13]. In the context of SDS, recent studies have begun to explore FL for tasks such as surgical phase recognition, scene segmentation, and tool detection [17, 18, 19].

To address challenges in SDS, initiatives such as the Endoscopic Vision (EndoVis) Challenge [20] 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. It aimed to benchmark the state of the art in applying FL to surgical AI for privacy-preserving model development across institutions. 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 [21]. 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 benchmarking paper has been published for the complete dataset [22]. 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 [23].

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.
- A benchmark of FL strategies for the novel task of patient-level surgical video appendicitis grading classification, advancing the complexity of analysis beyond common static, image-based tasks.
- A rigorous, data-driven analysis of the critical trade-off between model generalization to unseen institutions and personalization to local client data.
- A clear demonstration of the current strengths and practical limitations of FL in surgical AI, highlighted by bootstrapping and Wilcoxon signed-rank test.

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 [21] (see Subsection 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 (Figure 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 (Figure 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

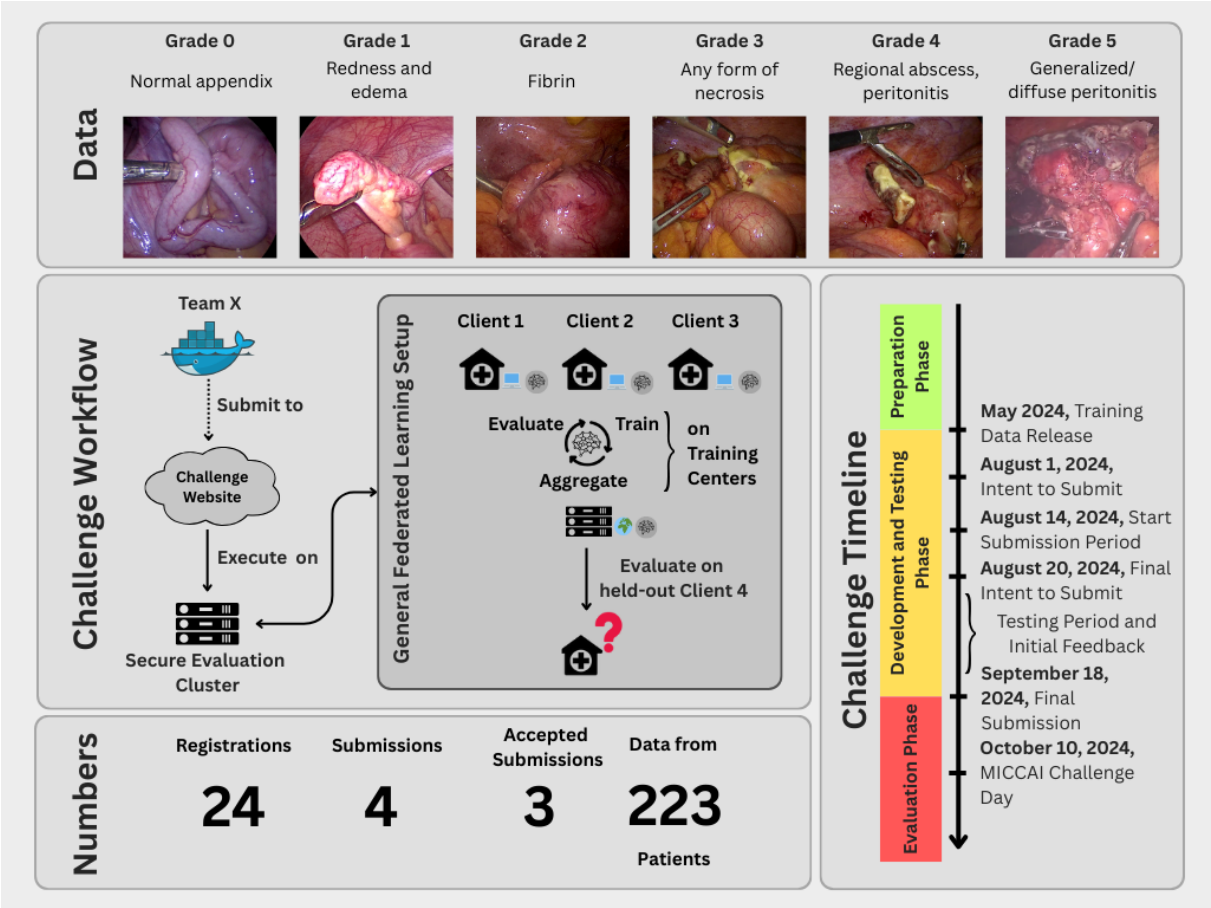


Figure 1: **FedSurg24 Challenge Highlights:** The top panel shows example images of intraoperative appendicitis grades, defined according to Gomes et al. [24], which were used for video annotation. The lower panel illustrates the FedSurg Challenge workflow: teams submitted Docker containers via Synapse, 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.

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 benchmarking study of different 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 [21].

This challenge's main contribution is its patient-level prediction task for appendicitis staging, a transferable innovation for AI in surgery. By establishing a framework for standardized, objective assessment of inflammation severity, this approach serves as a powerful proof-of-concept for addressing more critical clinical needs. These include advancing intraoperative assistance for surgical quality control, and providing a foundation for developing AI tools for accurate diagnosis, real-time surgical support, and effective prognosis for a variety of surgical conditions.

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 federated learning, 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, comprising 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 [21]. For this challenge, 200 frames were extracted per video using FFmpeg software [25], capturing a 100-second window sampled at two frames per second around an annotated timestamp [21].

Frames were selected at the timestamp identified by the operating surgeon when the appendix was fully visible prior to dissection [21]. 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 [21]) defines class descriptions and includes illustrative examples based on the definition by Gomes et al. [24]. Furthermore, fine-grained classes 3A/3B were merged into class 3, and classes 4A/4B into class 4 in this challenge dataset. An example overview of the data is visible in Figure 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. Therefore, we deem the risk of misclassification limited to borderline cases (i.e., cases between partial and total necrosis of the appendix). This potential source of error applies equally to all centers.

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:

- 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.

Table 1: Data Distribution Across Participating Centers.

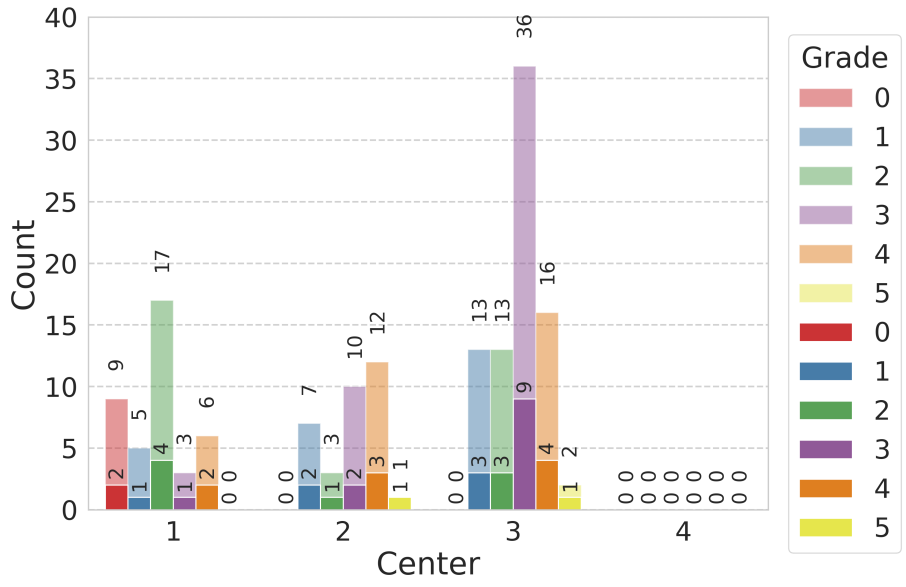
Center	**Training Videos**	**Testing Videos**	**Total**
1	40	10	50
2	33	9	42
3	80	22	102
4	0	29	29
Total	153	70	223

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 (Figure 1 and Figure 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.

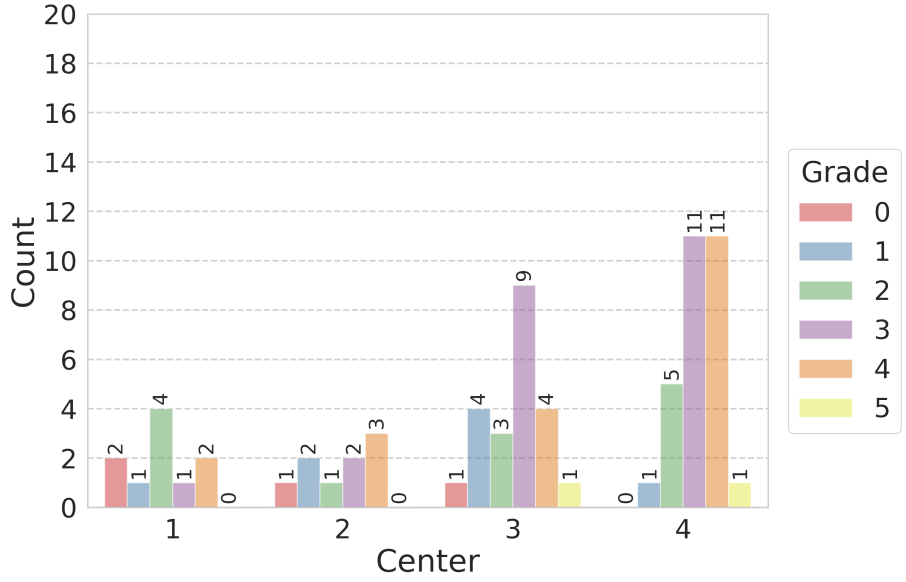
No predefined validation set was provided. Participants were expected to devise their own validation strategies. Final submissions were executed centrally by the organizers, who conducted full federated training and evaluation across all centers under secure and standardized conditions to ensure reproducibility 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 here[1]. In addition, two videos are released that were excluded from the final dataset because they were either shorter than 100 seconds or did not contain a clearly visible appendix in the center. The Appendix300 dataset is publicly available for non-commercial use under the Creative Commons Attribution (CC BY) license. Any use of this dataset requires citing both this paper and the final Appendix300 publication.

[1]Available upon acceptance or on request.



(a) Training Dataset Distribution.



(b) Testing Dataset Distribution.

Figure 2: **Label Distribution Across Data Subsets per Center.** Label distributions for (a) the training dataset and (b) the test dataset across the four centers. The plots highlight notable inter-center variability and class imbalance. 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 exact data distribution was unknown for the participants.

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, 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 [26, 27]. 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:

$$\begin{aligned}
\mathrm{TP}_c &= M_{c,c} \quad \text{(true positives)} \\
\mathrm{FP}_c &= \sum_{\substack{i=1 \\ i\neq c}}^{C} M_{i,c} \quad \text{(false positives)} \\
\mathrm{FN}_c &= \sum_{\substack{j=1 \\ j\neq c}}^{C} M_{c,j} \quad \text{(false negatives)}
\end{aligned}$$

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 [28, 29, 27]. 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:

$$\begin{aligned}
N &= \sum_{i=1}^{C} \sum_{j=1}^{C} M_{i,j} && \text{(total number of samples)} \\
w_{i,j} &= \frac{|i-j|}{C-1} && \text{(cost of predicting class } j \text{ when the true class is } i\text{)}
\end{aligned}$$

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 integrates multiple performance metrics with rigorous statistical analysis. 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 [30] was applied. Bootstrapping, as emphasized by Maier-Hein et al. [31], 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. To statistically compare team performances, we applied the Wilcoxon signed-rank test to the bootstrapped metric values obtained over all iterations, thereby quantifying whether observed performance differences were significant beyond random variation.

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 Mazana 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.

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.

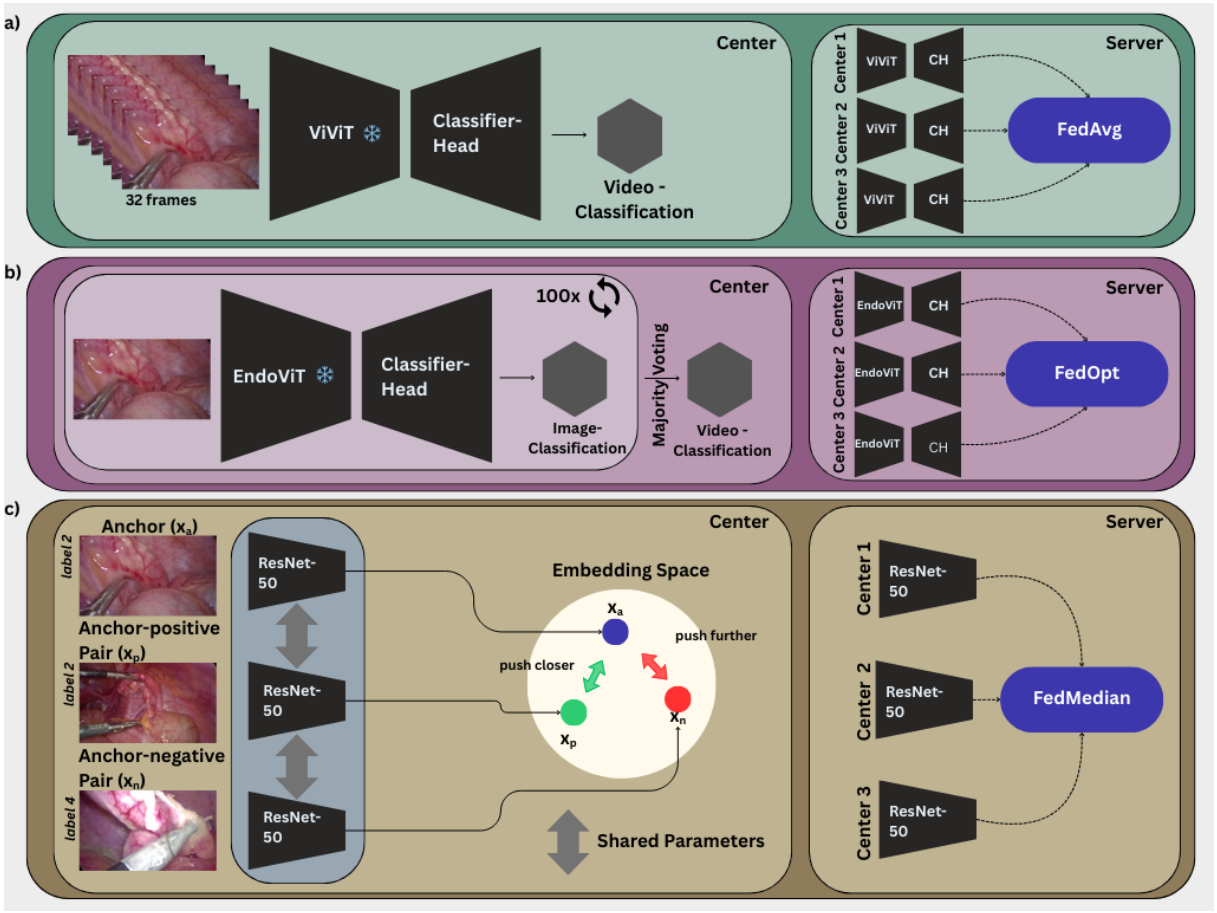


Figure 3: **Methods Overview:** The three submissions shown utilize different backbone architectures and federated strategies. A common approach is that in each server round, the best-performing model from a client's local training rounds is sent to the server for aggregation. (a) Team Santhi uses a 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	Pretrained Video Vision Transformer (ViViT) with a frozen backbone.	Pretrained EndoViT with a frozen encoder.	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	5	50	10
Local Rounds	20	2	5
Learning Rate	1×10^{-4}	1×10^{-3}	1×10^{-6}
Batchsize	4	128	1

3.1.1. Team Santhi

Santhi R. Kolamuri's approach leverages a pretrained Video Vision Transformer (ViViT) [32], which is well-suited for capturing spatio-temporal features in surgical video sequences (Figure 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 [33], 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 [34], 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 [35], a Vision Transformer pretrained on the Endo700k dataset consisting of diverse endoscopic images (Figure 3 (b)). Following a standard linear probing setup [33], 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 [36, 37], 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 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 [38, 39], 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 [34]. 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 [38]. Model aggregation on the server is done with FedOpt [40]. 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 [41], inspired by the original Siamese architecture [42]. 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 (Figure 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 [43], 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 [10], which is robust to outliers. Each center performs 10 local epochs followed by 5 epochs dedicated to triplet optimization. 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. 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, followed by the Wilcoxon signed-rank test for statistical significance. A comprehensive summary of results—including key metrics, confusion matrices, and rank stability—is provided in Tables 3 and 4 and Figures 4, 5, 6, and 7, with further details in Appendix Appendix G.

3.2.1. Task 1: Generalization to an Unseen Center

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%**

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 superior generalization under domain shift. This is evident in the confusion

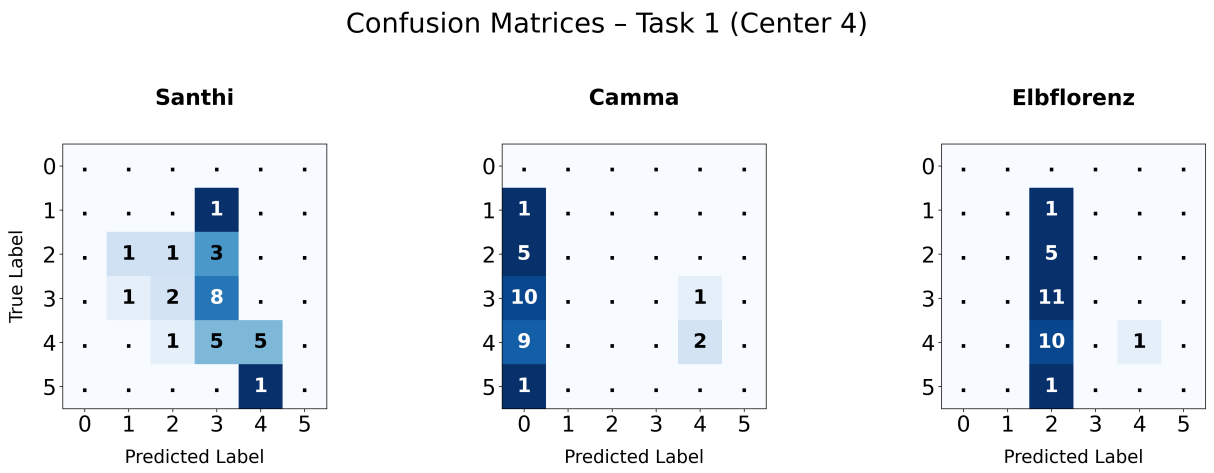


Figure 4: **Confusion Matrices – Task 1, Center 4.** Confusion matrices for the participating teams on Center 4 (Task 1). The values in the confusion matrices are not normalized. The color highlighting is normalized row-wise by true labels. The diagonal highlights class-wise recall, while off-diagonal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

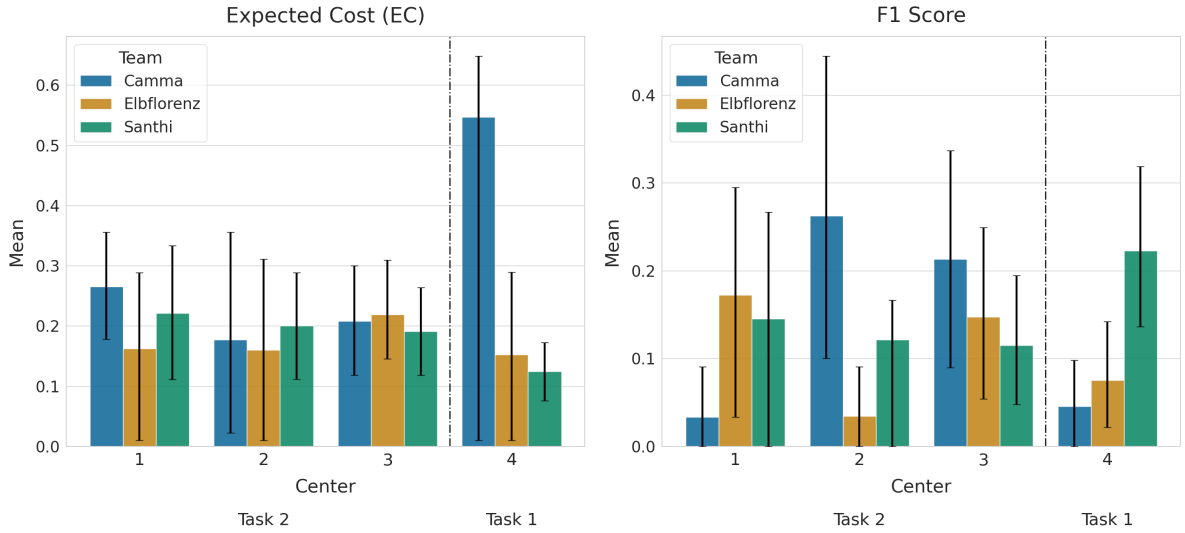
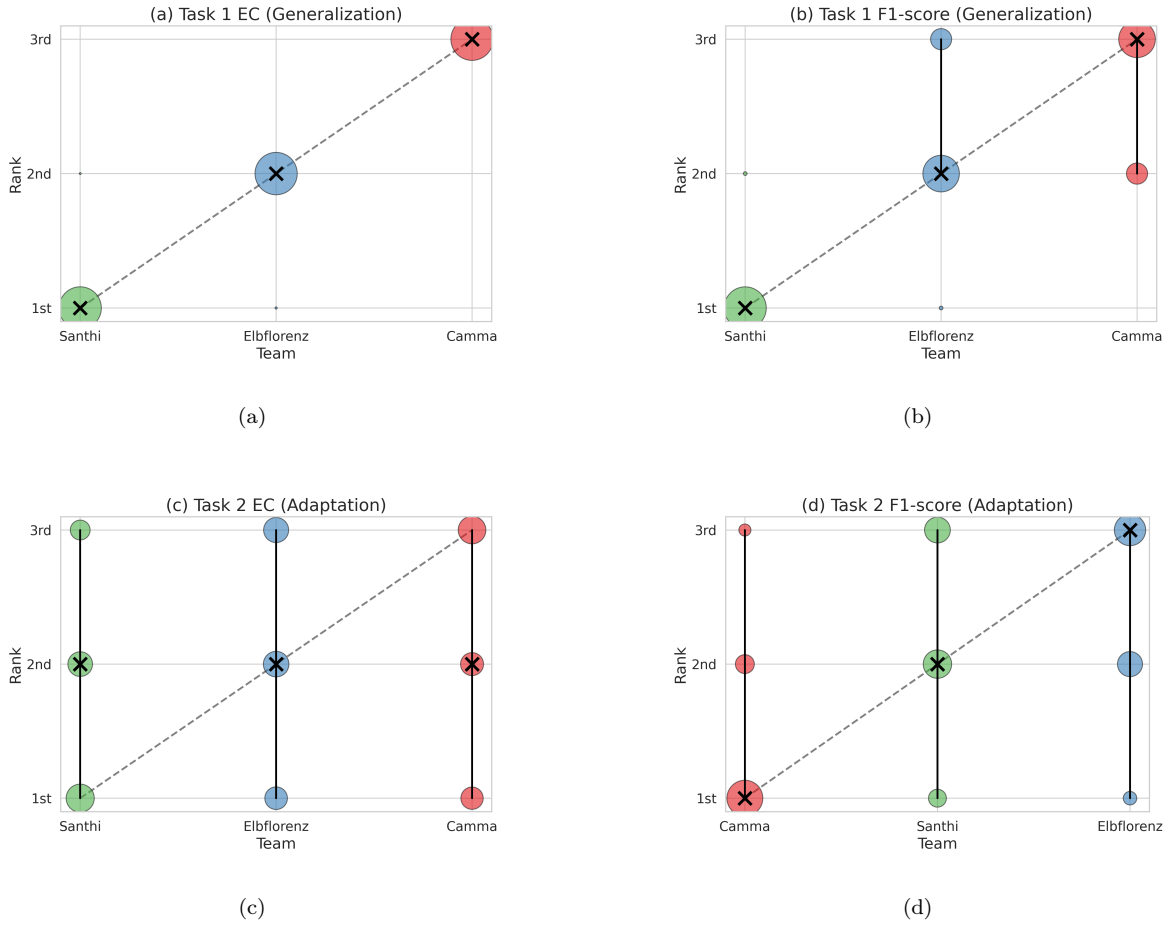


Figure 5: **Bootstrapped Performance Results.** Visualization of the performance results with standard deviation as error bars for all teams and tasks after bootstrapping with 10,000 repetitions. The plot illustrates the variability and stability of the outcomes across different centers.



(a) (b) (c) (d)

Figure 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.

matrix (Figure 4), where Santhi's predictions tend to align with 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 (Figure 5) confirm the statistical robustness of these findings. This is further supported by the ranking stability plots (Figures 6a and 6b), 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 trivial classifier is supported by bootstrapping analysis (Figure 5) and statistical significance testing (Table 3). Across 10,000 repetitions, Team Camma consistently ranked as the lowest-performing team in 100% of cases for the EC-Score and 75.09% for the F1-score (Figure 6).

Team Elbflorenz also demonstrated limited performance, predominantly predicting class 2 across inputs (Figure 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 (Figure 6, Figure 5).

Wilcoxon signed-rank tests confirmed statistically significant performance differences between all models at Center 4. This validates that the observed variations in EC and F1-score reflect meaningful distinctions in generalization ability under domain shift. The ranking stability plot (Figure 6) visually reinforces this trend, clearly showing Team Santhi in the lead, followed by Elbflorenz and Camma. Therefore, the leaderboard presented in Table 5 is robust.

3.2.2. Task 2: Center-Specific Adaptation

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		26.67%	3.70%
Elbflorenz	1	**17.78%**	20.20%
Santhi		22.22%	**20.83%**
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		21.79%	**18.91%**
Elbflorenz	Average	21.35%	13.14%
Santhi		**20.44%**	15.40%

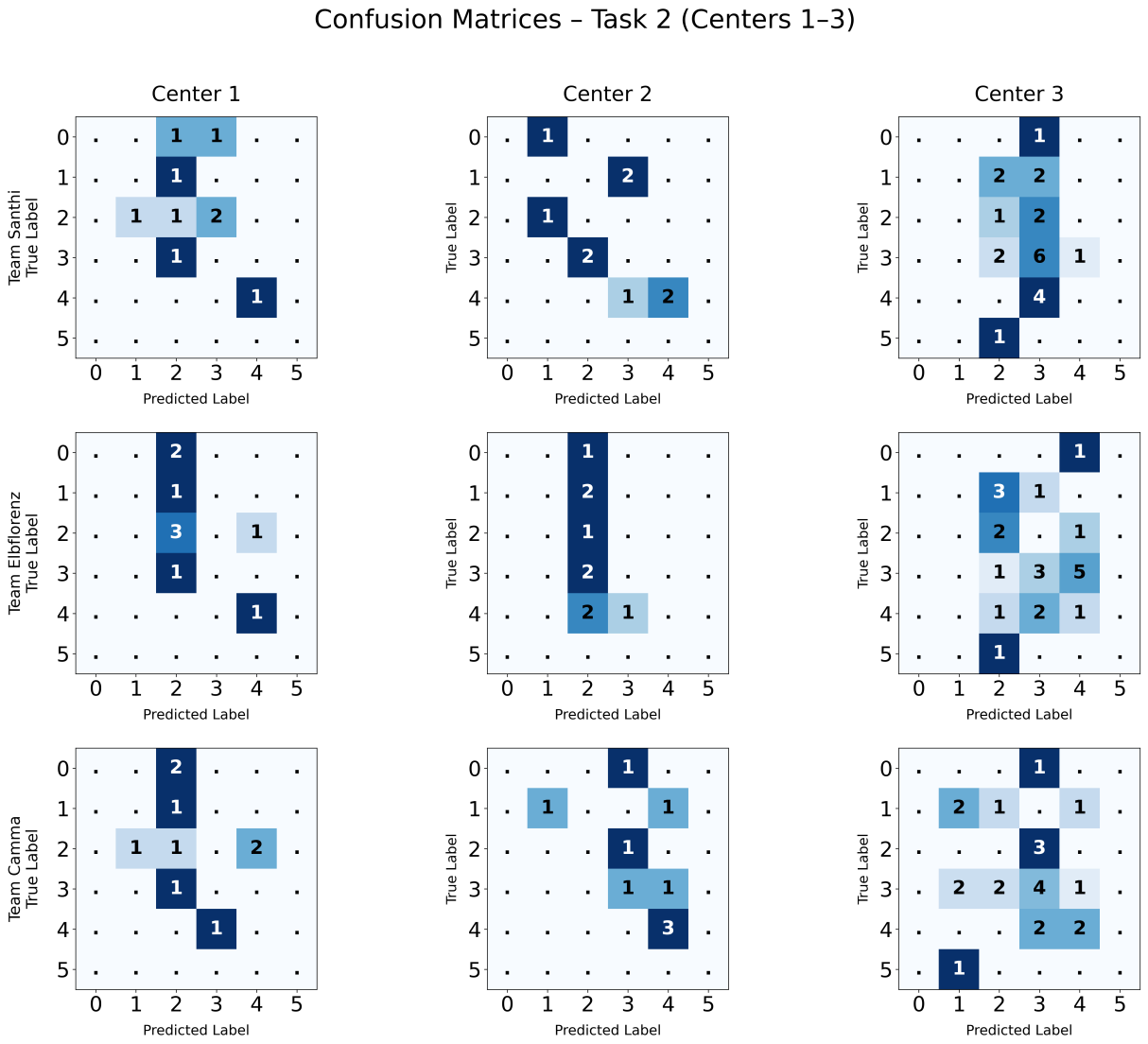


Figure 7: **Confusion Matrices – Task 2, Centers 1–3.** The values in the confusion matrices are not normalized. The color highlighting is normalized row-wise by true labels. The diagonal highlights class-wise recall, while off-diagonal values indicate common misclassification patterns.

Task 2 evaluated model performance in intra-center settings, where each local model was tested on its dedicated test set from its associated center. All teams demonstrated improvements over Task 1, although performance varied considerably across centers (Table 4, Figure 5).

Team Santhi demonstrated robust performance across centers, with F1-scores of 20.83% (Center 1), 13.33% (Center 2), and 12.04% (Center 3) (Table 4). Although not consistently the best at individual centers, their steady results, combined with relatively low EC values, indicate effective regularization and strong adaptability to the test data of the training centers. The confusion matrices in Figure 7 reveal predictions largely concentrated along the diagonal, suggesting good class differentiation. In Center 3, however, the local model—similar to Team Elbflorenz—showed limited discriminatory power, with a tendency to predict classes 2, 3, or 4 rather than providing balanced classification. Bootstrapping analysis (Figure 5) confirms consistent performance across centers, but the final ranking remained unstable, as indicated by the wide 95% confidence intervals in Figure 6. For the F1-score, Santhi ranked first in 17.86% of iterations, second in 45.02%, and third in 37.12%; for EC, the respective values were 44.12%, 33.94%, and 21.94%.

Team Elbflorenz displayed greater inconsistency across centers, especially in the F1-Score (Figure 5). The confusion matrices (Figure 7) reveal inconsistent local adaptation: the models for Centers 1 and 2 collapsed into trivial classifiers, biasing predictions towards a single class, while the model for Center 3 failed to discriminate effectively between classes 2, 3, and 4. Despite this, the model performed relatively well at Center 1 (F1: 20.20%), but performance dropped markedly at Center 2 (3.70%) and moderately at Center 3 (15.51%). Bootstrapping results suggest difficulties in achieving reliable within-center generalization, potentially attributable to data imbalance or overfitting. Similar to Team Santhi, Team Elbflorenz's ranking fluctuates between first, second, and third place across evaluation metrics (Figure 6). For the F1-Score, Elbflorenz ranked first in 10.23% of cases, second in 35.03%, and third in 54.74%. Regarding the EC metric, Elbflorenz achieved first place in 44.12% of cases, second place in 33.94%, and third place in 21.94%.

Team Camma exhibited pronounced center-dependent performance (Table 4, Figure 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 3.70%. Confusion matrices shown in Figure 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 71.29% of cases, 2nd place in 19.94%, and 3rd place in 8.14%. For the EC metric, the rankings were 1st place in 27.81%, 2nd place in 29.55%, and 3rd place in 42.64%.

Importantly, Wilcoxon signed-rank tests revealed statistically significant performance differences between the models for all centers and both metrics. Conducting the test after bootstrapping substantially supports the statistical power, ensuring that the detected differences were not due to chance. Despite these significant differences, the bootstrapped ranking stability analysis (Figures 6c and 6d) shows notable fluctuations in team positions within the bootstrapping. While the Task 2 F1-score rankings (Figure 6d) align with the overall standings in Table 5, the EC mean-based rankings (Figure 6c) 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. Such variability highlights that Task 2 outcomes are sensitive to small changes in the test set, and claims regarding

top-performing submissions should therefore be made with caution. In line with the interpretation of Maier-Hein et al. [31], the fact that no team retained its original EC rank in at least 50% of bootstrap replicates indicates that the Task 2 ranking is unstable. We therefore report the original ranking for completeness, while emphasizing that it should not be interpreted as conclusive evidence of submission superiority.

Table 5: Team rankings for the FedSurg Challenge are presented, where lower ranks indicate better performance.

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

4. Discussion

4.1. Task Wise performance

The FedSurg Challenge establishes the first benchmark for evaluating Federated Learning (FL) strategies in the surgical video classification of appendicitis grades, utilizing a heterogeneous, multi-institutional dataset. Three complete submissions were received and evaluated on a preliminary subset of the Appendix300 dataset.

Task 1, which focused on generalization to an unseen center, immediately revealed the limited generalization abilities of current FL models. This difficulty stems directly from significant data heterogeneity across the participating institutions. For instance, images originated from a wide array of surgical video systems, including Arthrex Synergy UHD4, Storz image1 s, B. Braun Aesculap EinsteinVision, and Richard Wolf Endocam 4k, leading to variations in lighting and resolution [21]. For example, Center 1 provided 4k images, while Center 4 provided cropped images in Full HD resolution. In addition to this feature skew, a label distribution skew existed between the different centers (Figure 2). In this challenging environment, models capable of temporal modeling, like the ViViT-based submission, achieved the strongest but still weak results. A slight tendency also emerged showing that approaches focused on the loss landscape, like FedSAM, performed slightly better than a FedMedian aggregation of a triplet margin loss and cosine similarity model. This aligns with the conclusions of Foret et al. [39] and Caldarola et al. [38] that the quest for flat minima, which is incorporated into FedSAM, can lead to better generalization ability. This difficulty was not unique to the challenge submissions; a benchmark study on the Appendix300 dataset [22] using a leave-one-out cross-validation design similarly confirmed the profound challenge of inter-center generalization.

Task 2, which involved center-specific adaptation after fine-tuning, led to only marginally better results across all teams. This marginal improvement, however, came at the cost of rank instability, highlighting that even minor gains in local adaptation can compromise the robustness of a model's comparative performance. The limited improvement can be attributed to several factors inherent to the federated setup. Due to the difficult knowledge of the complete data distribution and amounts, hyperparameter tuning was challenging, which could lead to local models overfitting to their local training data and, therefore, still failing to achieve high accuracy. Furthermore, the knowledge from other

centers within the global model can negatively influence client performance. At each training round, the client initializes its weights from the global model, which represents an aggregated compromise of the knowledge learned across all participating centers. The poor local performance could stem from conflicting updates that pull the model away from what is optimal for its own center's data.

Ultimately, these findings illustrate the central challenge of balancing generalization and adaptation, alongside the overarching difficulty of data heterogeneity in federated setups.

4.2. Stability assessment

The bootstrapped rank frequency analysis yielded distinct outcomes for Task 1 and Task 2.

For Task 1, the analysis indicated stable leader board rankings, with the predictions from submitted models being statistically distinct. A central finding was the critical role of temporal information; the highest-ranking submission utilized a video-based model, whereas models predicated on static frames exhibited the poorest performance. This aligns with the findings from the Appendix300 benchmarking paper [22], where the model trained only on middle frames performed poorly in comparison to the models trained on video models with a temporal horizon. Furthermore, using a higher frame rate up to 1 frame per second (fps) resulted into better results.

In stark contrast, the rankings for Task 2 demonstrated pronounced instability. No submission maintained its rank on the EC metric in more than 50% of the bootstrap replicates, indicating that the leader board outcomes were not robust. This instability suggests that leader boards for federated adaptation tasks should be interpreted with extreme caution, as top-ranking methods may not be reliably superior but merely fortunate in a specific data configuration.

4.3. Systemic Limitations and Data-Related Challenges in Federated Surgical AI

Several systemic issues emerged. FL does not resolve underlying dataset problems such as class imbalance and can amplify heterogeneity effects. Rare inflammation stages were underrepresented, leading to poor classification across all teams. Personalized losses and fine-tuning improved performance within centers (e.g., Team Camma) but compromised generalization to unseen centers, underscoring a fundamental trade-off in FL. This mirrors our findings from the primary tasks, where the unstable gains in center-specific adaptation (Task 2) failed to translate into the robust, generalizable models needed for unseen centers (Task 1). Similarly, design choices impacted outcomes: lightweight frozen foundation models provided stability across centers, whereas more adaptive approaches (e.g., Siamese networks with triplet loss) improved local performance but struggled with generalization under distribution shifts.

Furthermore, hyperparameter tuning proved difficult without access to the true data distribution, reflecting real-world FL constraints. The unknown class distribution contributed to lower-than-expected performance. Moreover, human experts showed only moderate agreement on inflammation grading ($\kappa_{weighted} = 0.62$ [21]), underscoring the inherent ambiguity of the classification task.

4.4. Future directions

Future work should address identified limitations. Imbalance-aware strategies, such as reweighting, resampling, or federated focal loss, are needed to improve performance

on clinically critical minority classes. Context-aware preprocessing and targeted frame sampling consistently outperformed uniform strategies and should be further explored. Methodological advances such as federated domain adaptation, personalized aggregation, and uncertainty-aware inference may improve robustness and safety. Finally, self-supervised pretraining on large-scale surgical video datasets could yield more transferable representations without compromising data privacy.

4.5. Challenge insights

Despite 24 registered teams, only three complete submissions were ultimately received. This discrepancy reflects both the niche character of FL within surgical AI and the complexity of the challenge setup. Unlike conventional challenges, FedSurg required fully encapsulated end-to-end submissions in Docker, with limited access to data and restricted debugging capabilities. These hurdles highlight practical barriers to broader participation and point to the technical demands of real-world FL applications.

5. Conclusion

The FedSurg Challenge provides the first benchmark for FL in surgical video classification, highlighting both opportunities and current limitations. The findings demonstrate that while spatiotemporal modeling and preprocessing choices can support generalization and adaptation, challenges persist in handling data imbalance, hyperparameter tuning, and stability of results. By exposing these trade-offs, FedSurg establishes an important reference point for the surgical AI community and guides future work toward robust, adaptive, and clinically relevant FL methods.

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 (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).

Conflict of Interest

JNK 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.

FRK declares advisory roles for Radical Healthcare, USA; and the Surgical Data Science Collective, USA, and has received research funding from Novartis.

SS received speaker's fees from Stryker Corporation.

AR, NP, SS, DS, and SBa serve as an Associate Editor for the Medical Image Analysis journal.

NP is co-founder and owns shares in Scialytics SAS.

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 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.

Appendix A. Dataset and Annotation Protocol

The challenge dataset is a preliminary subset of the publicly available Appendix300 dataset [21]. A CSV file detailing the samples used in the challenge is available here [2], along with two videos that were excluded from the final dataset but used in the challenge for reproducibility. The dataset is available for non-commercial use under a CC BY license and requires citation of the Appendix300 publication. The annotation protocol is available in [21]. 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[3].

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 S. Lestner 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

[2] Available upon acceptance or on request.

[3] Available at https://tinyurl.com/nhb5z6a3

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[4]. To support development, an example setup based on the FL Flower framework [34] was provided through a GitLab repository[5]. The evaluation framework was made transparent through another repository[6], which included the source code for metric computation as well as the ranking scripts (bootstrapping, statistical testing, and related plots).

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 Wilcoxon Signed-Rank Test Results

[4] Available at https://www.synapse.org/Synapse:syn53137385/wiki/625370
[5] Available at https://gitlab.com/nct_tso_public/challenges/miccai2024/FedSurg24
[6] Available at https://gitlab.com/nct_tso_public/challenges/miccai2024/snippet

Table G.6: Bootstrap rank frequency, win probability, and Wilcoxon signed-rank p-values for both tasks (F1-score).

Task	Center	Team	Rank Freq.			Win Prob.			Wilcox p		
			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	**0.7192**	0.1994	0.0814	NaN	0.8501	0.7876	NaN	0.0000	0.0000
2	Avg	Elb	0.1023	0.3503	**0.5474**	0.1499	NaN	0.4050	0.0000	NaN	5.04×10^{-106}
2	Avg	San	0.1786	**0.4502**	0.3712	0.2123	0.5950	NaN	0.0000	5.04×10^{-106}	NaN

Table G.7: Bootstrap rank frequency, win probability, and Wilcoxon signed-rank p-values for both tasks (EC).

Task	Center	Team	Rank Freq.			Win Prob.			Wilcox p		
			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	0.2781	0.2955	**0.4264**	NaN	0.4661	0.3791	NaN	3.13×10^{-13}	1.15×10^{-177}
2	Avg	Elb	0.2855	**0.3646**	0.3499	0.5307	NaN	0.3991	3.13×10^{-13}	NaN	1.17×10^{-128}
2	Avg	San	**0.4412**	0.3394	0.2194	0.6176	0.5983	NaN	1.15×10^{-177}	1.17×10^{-128}	NaN

Appendix H. Credit Authorship Contribution Statement

Max Kirchner: Conceptualization, Challenge Organization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review and Editing, Visualization, Project Administration **Hanna Hoffmann:** Conceptualization, Software, Writing - Review and Editing **Alexander C. Jenke:** Challenge Organization, Software, Writing - Review and Editing **Oliver L. Saldanha:** Conceptualization, Challenge Organization, Software, Investigation, Data Curation, Writing - Review and Editing **Kevin Pfeiffer:** Software, Data Curation, Writing - Review and Editing **Kanjo Weam:** Data Curation, Writing - Review and Editing **Julia Alekseenko:** Methodology, Software, Investigation, Visualization, Writing - Review and Editing **Claas de Boer:** Methodology, Software, Investigation, Writing - Review and Editing **Santhi Raj Kolamuri:** Methodology, Software, Investigation, Writing - Review and Editing **Lorenzo Mazza:** Methodology, Software, Investigation, Writing - Review and Editing **Nicolas Padoy:** Challenge Team Supervision, Writing - Review and Editing **Sophia Bano:** Challenge Organization, Supervision, Writing - Review and Editing **Annika Reinke:** Conceptualization, Challenge Organization, Supervision, Writing - Review and Editing **Lena Maier-Hein:** Challenge Organization, Supervision, Writing - Review and Editing **Danail Stoyanov:** Challenge Organization, Supervision, Writing - Review and Editing **Jakob N. Kather:** Conceptualization, Challenge Organization, Writing - Review and Editing, Supervision, Project administration, Funding Acquisition **Fiona R. Kolbinger:** Conceptualization, Challenge Organization, Investigation, Data Curation, Writing - Review and Editing, Supervision, Project Administration, Funding Acquisition **Sebastian Bodenstedt:** Conceptualization, Challenge Organization, Software, Investigation, Data Curation, Writing - Review and Editing, Supervision, Project Administration, Funding Acquisition, Challenge Team Supervision **Stefanie Speidel:** Conceptualization, Challenge Organization, Writing - Review and Editing, Supervision, Project Administration, Funding Acquisition, Challenge Team Supervision

References

[1] L. Maier-Hein, S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katić, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, P. Jannin, Surgical data science for next-generation interventions, Nature Biomedical Engineering 1 (Sep. 2017). doi:10.1038/s41551-017-0132-7.

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

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

[4] M. Carstens, S. Vasisht, Z. Zhang, I. Barbur, A. Reinke, L. Maier-Hein, D. A. Hashimoto, F. R. Kolbinger, Artificial intelligence for surgical scene understanding: A systematic review and reporting quality meta-analysis, ISSN: 3067-2007 Pages: 2025.07.12.25330122 (2025). doi:10.1101/2025.07.12.25330122.
URL https://www.medrxiv.org/content/10.1101/2025.07.12.25330122v1

[5] K. Kirtac, N. Aydin, J. Lavanchy, G. Beldi, M. Smit, M. Woods, F. Aspart, Surgical Phase Recognition: From Public Datasets to Real-World Data, Applied Sciences 12 (2022) 8746. doi:10.3390/app12178746.

[6] J. L. Lavanchy, S. Ramesh, D. Dall'Alba, C. Gonzalez, P. Fiorini, B. Muller-Stich, P. C. Nett, J. Marescaux, D. Mutter, N. Padoy, Challenges in Multi-centric Generalization: Phase and Step Recognition in Roux-en-Y Gastric Bypass Surgery, arXiv:2312.11250 [cs] (Dec. 2023). doi:10.48550/arXiv.2312.11250.
URL http://arxiv.org/abs/2312.11250

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

[8] General data protection regulation (GDPR) – legal text (2016).
URL https://gdpr-info.eu/

[9] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y. Arcas, Communication-Efficient Learning of Deep Networks from Decentralized Data, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282, iSSN: 2640-3498.
URL https://proceedings.mlr.press/v54/mcmahan17a.html

[10] D. Yin, Y. Chen, R. Kannan, P. Bartlett, Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates, in: Proceedings of the 35th International Conference on Machine Learning, PMLR, 2018, pp. 5650–5659, iSSN: 2640-3498.
URL https://proceedings.mlr.press/v80/yin18a.html

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

[12] T. Li, A. K. Sahu, A. Talwalkar, V. Smith, Federated Learning: Challenges, Methods, and Future Directions, IEEE Signal Processing Magazine 37 (3) (2020) 50–60. doi:10.1109/MSP.2020.2975749.
URL https://ieeexplore.ieee.org/document/9084352

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

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

[15] A. Z. Tan, H. Yu, L. Cui, Q. Yang, Towards personalized federated learning, IEEE Transactions on Neural Networks and Learning Systems 34 (12) (2023) 9587–9603. doi:10.1109/TNNLS.2022.3160699.

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

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

[18] M. Kirchner, A. C. Jenke, S. Bodenstedt, F. R. Kolbinger, O. L. Saldanha, J. N. Kather, M. Wagner, S. Speidel, Federated EndoViT: Pretraining Vision Transformers via Federated Learning on Endoscopic Image Collections, arXivArXiv:2504.16612 [cs] (May 2025). doi:10.48550/arXiv.2504.16612.
URL http://arxiv.org/abs/2504.16612

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

[20] S. Speidel, L. Maier-Hein, D. Stoyanov, S. Bodenstedt, A. Reinke, S. Bano, Endoscopic Vision Challenge – A MICCAI Challenge.
URL https://opencas.dkfz.de/endovis/

[21] F. R. Kolbinger, M. Kirchner, K. Pfeiffer, S. Bodenstedt, A. C. Jenke, J. Barthel, M. R. Carstens, K. Dehlke, S. Dietz, S. Emmanouilidis, G. Fitze, L. Leitermann, S. T. Mees, S. Pistorius, C. Prudlo, A. Seiberth, J. Schultz, K. Thiel, D. Ziehn, S. Speidel, J. Weitz, J. N. Kather, M. Distler, O. L. Saldanha, Appendix300: A multi-institutional laparoscopic appendectomy video dataset for computational modeling tasks, ISSN: 3067-2007 Pages: 2025.09.05.25335174. doi:10.1101/2025.09.05.25335174.
URL https://www.medrxiv.org/content/10.1101/2025.09.05.25335174v1

[22] O. L. Saldanha, K. Pfeiffer, S. Bodenstedt, M. Kirchner, A. C. Jenke, C. Barata, S. Barbosa, J. Barthel, M. Carstens, L. T. Castro, K. Dehlke, S. Dietz, S. Emmanouilidis, G. Fitze, M. Freitag, F. Holderried, W. Kanjo, L. Leitermann, S. T. Mees, A. S. Soares, M. Pascoal, S. Pistorius, C. Prudlo, J. Schultz, A. Seiberth, K. Thiel, X. Wu, D. Ziehn, S. Speidel, J. Weitz, M. Distler, J. N. Kather, F. R. Kolbinger, Decentralized, privacy-preserving surgical video analysis with swarm learning, ISSN: 3067-2007 Pages: 2025.10.02.25337106. doi:10.1101/2025.10.02.25337106.
URL https://www.medrxiv.org/content/10.1101/2025.10.02.25337106v1

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

[24] C. A. Gomes, T. A. Nunes, J. M. Fonseca Chebli, C. S. Junior, C. C. Gomes, Laparoscopy grading system of acute appendicitis: new insight for future trials, Sur-

gical Laparoscopy, Endoscopy & Percutaneous Techniques 22 (5) (2012) 463–466. doi:10.1097/SLE.0b013e318262edf1.

[25] S. Tomar, Converting video formats with ffmpeg, Linux Journal 2006 (146) (2006) 10.

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

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

[28] T. Hastie, R. Tibshirani, J. H. Friedman, J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, Vol. 2, Springer, 2009.

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

[30] B. Efron, Bootstrap methods: another look at the jackknife, in: Breakthroughs in statistics: Methodology and distribution, Springer, 1992, pp. 569–593.

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

[32] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, ViViT: A Video Vision Transformer, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, 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

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:2010.11929 [cs] (Jun. 2021). doi:10.48550/arXiv.2010.11929.
URL http://arxiv.org/abs/2010.11929

[34] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. d. Gusmão, N. D. Lane, Flower: A Friendly Federated Learning Research Framework, arXiv:2007.14390 [cs] (Mar. 2022). doi:10.48550/arXiv.2007.14390.
URL http://arxiv.org/abs/2007.14390

[35] D. Batić, F. Holm, E. Özsoy, T. Czempiel, N. Navab, EndoViT: pretraining vision transformers on a large collection of endoscopic images, International Journal of Computer Assisted Radiology and Surgery 19 (6) (2024) 1085–1091. doi:10.1007/s11548-024-03091-5.
URL https://doi.org/10.1007/s11548-024-03091-5

[36] S. Yang, F. Zhou, L. Mayer, F. Huang, Y. Chen, Y. Wang, S. He, Y. Nie, X. Wang, Ö. Sümer, Y. Jin, H. Sun, S. Xu, A. Q. Liu, Z. Li, J. Qin, J. Y. Teoh, L. Maier-Hein, H. Chen, Large-scale Self-supervised Video Foundation Model for Intelligent Surgery, arXiv:2506.02692 [cs] (Jun. 2025). doi:10.48550/arXiv.2506.02692.
URL http://arxiv.org/abs/2506.02692

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

[38] D. Caldarola, B. Caputo, M. Ciccone, Improving Generalization in Federated Learning by Seeking Flat Minima, in: S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham, 2022, pp. 654–672. doi:10.1007/978-3-031-20050-2_38.

[39] P. Foret, A. Kleiner, H. Mobahi, B. Neyshabur, Sharpness-Aware Minimization for Efficiently Improving Generalization, arXiv:2010.01412 [cs] (Apr. 2021). doi:10.48550/arXiv.2010.01412.
URL http://arxiv.org/abs/2010.01412

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

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, arXiv:1512.03385 [cs] (Dec. 2015). doi:10.48550/arXiv.1512.03385.
URL http://arxiv.org/abs/1512.03385

[42] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature Verification using a "Siamese" Time Delay Neural Network, in: Advances in Neural Information Processing Systems, Vol. 6, Morgan-Kaufmann, 1993.
URL https://proceedings.neurips.cc/paper/1993/hash/288cc0ff022877bd3df94bc9360b9c5d-Abstract.html

[43] P. Luo, R. Zhang, J. Ren, Z. Peng, J. Li, Switchable Normalization for Learning-to-Normalize Deep Representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2) (2021) 712–728. doi:10.1109/TPAMI.2019.2932062.
URL https://ieeexplore.ieee.org/abstract/document/8781758