Endoscopic Vision Challenge 2025: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

Use the title to convey the essential information on the challenge mission.

Endoscopic Vision Challenge 2025

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

EndoVis25

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

With the advent of artificial intelligence as key technology in modern medicine, surgical data science (SDS) promises to improve the quality and value of the particular domain of interventional healthcare through capturing, organization, analysis, and modeling of data, thus creating benefit for both patients and medical staff. Holistic SDS concepts span the topics of context-aware perception in and beyond the operating room, data interpretation and real-time assistance or decision support. At the same time, minimally invasive surgery using cameras to observe the internal anatomy has become the state-of-the-art approach to many surgical procedures. Contributing to the key aspect of perception, endoscopic vision thus constitutes a central component of SDS and computer-assisted interventions. From this arises the necessity for high-quality common datasets that allow the scientific community to perform comparative benchmarking and validation of endoscopic vision algorithms. EndoVis (http://endovis.org/) organizes highprofile international challenges for the comparative validation of endoscopic vision algorithms that focus on different problems each year at MICCAI, comprising various computer vision tasks (classification, segmentation, detection, localization, etc) and subdisciplines ranging from laparoscopy to coloscopy and surgical training. It acts umbrella for several sub-challenges in this field, for the 10th anniversary this year we propose 4 different sub-challenges within EndoVis as well as keynotes from world leading experts in this field.

Challenge keywords

List the primary keywords that characterize the challenge.

Surgical Vision, Endoscopy, Classification, Detection, Segmentation

Year

2025

Novelty of the challenge

Briefly describe the novelty of the challenge.

EndoVis consists of different sub-challenges, while three of them are an extension of sub-challenges that took place last year (SurgVu, STIR, OSS) with novel tasks and data, one novel sub-challenges is proposed as well (RARE). SurgVu was already accepted as a lighthouse challenge and will be integrated into EndoVis.

Task description and application scenarios

Briefly describe the application scenarios for the tasks in the challenge.

N/A

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

NA

Duration

How long does the challenge take?

Full day

In case you selected half or full day, please explain why you need a long slot for your challenge.

EndoVis acts as an umbrella and consists of several sub-challenges. Each sub-challenge has specific challenge tasks itself and needs a least one hour to present the challenge results. In addition we have two high-profile keynote speakers (Shekoofeh Azizi, Google DeepMind and Pete Mountney, Founder of Odin Vision), each of them with a 1-hour slot.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

60-80 (based on the number of previous EndoVis challenges)

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

The joint publication will be coordinated by the particular sub-challenge organizers.

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

depends on the specific sub-challenges, e.g. DREAM/synapse platform for example
normal conference infrastructure on the challenge day (beamer, loud speaker, ...)

TASK 1: SurgVu: Surgical Visual Understanding

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Growing at an exponential rate with well over a million cases performed each year, robotic assisted surgery (RAS) promises to transform surgical intervention. Both the nature of the data these cases produce, and the sheer amount of it, present altogether new possibilities for study. Pursuits such as the quantification of surgical performance, efficiency and tool choreography, OR resource planning, AI-guided surgical planning, and surgical data science in general can all exploit this new source of clinical data. Not surprisingly, machine learning techniques that can extract meaning from these vast amounts of data seem poised to play an integral role. With this goal in mind we invite the surgical data science community to take part in two challenges, utilizing the largest publicly available dataset released to date (840+ hours of data). The first task of this sub challenge requires participants to build a model that localizes tools and their corresponding key-points in videos, using only tool presence data (surgical tool detection). The second task of this sub challenge invites participants to segment videos into different surgical steps being performed (surgical step recognition). Winning solutions will help advance surgical assistive technologies significantly.

Keywords

List the primary keywords that characterize the task.

computer vision, machine learining, surgical data science, surgical tool detection, surgical activity recognition

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

Aneeq Zia, Max Berniker, Conor Perreault, Rogerio Nespolo, Anthony Jarc, Intuitive Surgical

b) Provide information on the primary contact person.

Aneeq Zia
Aneeq.Zia@intusurg.com

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Repeated event with annual fixed conference submission deadline

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI 2025/EndoVis

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

TBA

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Publicly available data is allowed

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate and be visible on leaderboard, but will not be eligible for awards

d) Define the award policy. In particular, provide details with respect to challenge prizes.

for each task: 3 monetary prizes for 1st, 2nd , and 3rd place – exact amount per task to be decided

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Complete leaderboard will be visible publicly on the challenge website. The top 3 teams for each tasks will also be announced within the prizes sub-page of the challenge website

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

The organizers will publish a challenge paper within 6 months of the challenge day. All participating teams with valid submission will be eligible for being part of this combined paper. Following this publication, the participating teams will be allowed to publish their own results from the challenge citing the challenge paper. Each team will be asked to provide 2 team member names for authorship in challenge publication (but it may depend on the number of total participating teams)

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

The teams will need to submit an algorithm container (type 2) on grand-challenge. Detailed instructions will be provided to the teams on how to create such containers and submit them along with example submission containers.

Check out a submission instructions page from one of our previous challenge at https://surgtoolloc23.grand-challenge.org/submission-instructions/

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

There will be preliminary testing phases for each task where the teams will be allowed multiple (up to 10) algorithm submissions for testing. The dataset within these preliminary testing phases will only be 10-15% of the actual testing data set. The final rankings of the teams will only be based on their algorithm performances on the final testing phase submission.

Please check out an example preliminary and final testing phase leaderboards from one of our previous challenges below:

Prelim testing - https://surgtoolloc23.grand-challenge.org/evaluation/challenge/leaderboard/

Final testing - https://surgtoolloc23.grand-challenge.org/evaluation/final-testing-phase/leaderboard/

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)

- the release date(s) of the results

Registration open and release of training data: 1st April, 2025

Start of preliminary testing phase: 1st June, 2025

Start of final test phase: 5th August, 2025

New registrations deadline: 26th August, 2025

End of preliminary testing phase: 2nd September, 2025

End of final test phase: 10th September, 2025

Methodology reports (with all submission requirements) due: 20th September, 2025

Winners announced: On challenge day at MICCAI 2025

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

An existing Western IRB will be used

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The code to generate the evaluation metrics within the evaluation container (in grand-challenge) will be made public through a github repo.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

The participating teams will be required to make their algorithm container generating code public through a github repo.

An example of such a repo by a team in our previous challenge can be checked at https://github.com/vu-maple-lab/surgtool_challenge

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Only the organizing team (all within Intuitive) will have access to test cases and labels, hence there will be no conflict of interest with any other institution. All awards will also be sponsored by Intuitive, while any team from within Intuitive wanting to participate will not be eligible for any award.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research

Task category(ies)

State the task category(ies)

Examples:

- Classification

- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Detection,Classification

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Surgical tasks performed on porcine model by trainees during robotic surgical training. Tasks include suturing of different styles (1-hand, 2-hand, running), and dissection performed on various anatomy (uterine horn, rectal vein/artery, etc.). Tools include (but are not limited to) graspers, needle drivers, scissors, staplers, clip appliers, and energy instruments

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Surgical tasks performed on porcine model by trainees of various skill levels during robotic surgical training

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Single channel of endoscopic video

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Surgical tool detection: N/A
Surgical step recognition: The training videos will come with ground truth tool presence labels along with surgical step labels while the
testing data will have ground truth tool bounding box labels along with surgical step labels.

b) ... to the patient in general (e.g. sex, medical history).

N/A

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The data will be acquired from basic tasks being performed on a porcine model using a da Vinci Xi or Si system

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Surgical tool detection: Prediction of surgical tool bounding boxes.
Surgical step recognition: Prediction of surgical step being performed.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Tool detection: For bounding box detection, the assessment will be done using mean average precision over multiple intersection-over-union (IOU) values - this metric is standard for COCO dataset. Step recognition: For surgical step classification task, average f1-score across all steps will be used for evaluation.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

The Intuitive Data Recorder (IDR) will be used to capture video at 720p and 30fps from one channel of the endoscope on da Vinci Xi or Si system

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

The videos will be captured at 720p and 30fps from one channel of the endoscope on da Vinci Xi or Si system

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Data will be collected at 2 Intuitive Surgical training labs

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Experience of study participants will mostly be beginners (early in their learning curve) with a few experts (practicing surgeons) if possible.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case in this challenge will comprise of a video of a surgical training being performed on a porcine model. These videos will be long and variable in length where multiple surgical training steps will be present in each video along with different surgical tools.
The exact tools and steps present in the challenge dataset are given below (this information is also available in greater detail in our dataset publication and will be available to the teams):

Tools: Needle driver, monopolar curved scissor, force bipolar, clip applier, cadiere forceps, bipolar forceps, vessel sealer, permanant cautery hook/spatula, prograsp forceps, stapler, grasping retractor, tip-up fenestered grasper.

Steps: Suturing, Uterine horn, Suspensory ligaments, rectal artery/vein, skills application, range of motion, retraction and collision avoidance, other

b) State the total number of training, validation and test cases.

We will have 200+ cases for training and 50+ for testing. We will ensure variability in the dataset through the variety of tasks completed on the porcine model on different anatomy. Each case has an average length of around 4 hrs.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The numbers indicated were kept keeping in mind data collection technicalities and to provide enough data to the participants for developing meaningful models

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

We will try to ensure that the dataset has a balanced range of different tools and surgical steps within the training and testing set. We expect our dataset to have around 10+ unique tool labels with unequal distribution across classes (as some tools occur much more often than other e.g needle driver) while the surgical step distribution

will be much more balanced across training and testing sets.

e) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

The test data will be completely unseen (and unpublished). The test data will be ~20% of the training data size

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

We will use a crowd to annotate tool bounding boxes for our testing set. The annotations will not be redundant as bounding box annotations are not that subjective. The surgical step annotations will be done by a team of domain knowledge experts for training and testing sets.

At least 5 annotators for bounding boxes and 2-3 annotators for step annotations.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

For test set annotation, the crowd-sourced annotators were already trained and experienced in spatial annotation for surgical tools. Each frame will be annotated then reviewed by the annotation team to ensure quality. Bounding box labels will be placed around the surgical tools along with an object ID for object tracking. Additional tool classification label, such as left or right side will also be annotated. For surgical steps, the annotators will be provided by clear step starting and ending times for annotations.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Annotators will have significant experience in labelling bounding boxes for surgical tools and surgical steps.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

N/A

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Raw video frames will not be altered

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Image annotation will only be needed for the test set. Main sources of error would include the bounding box not being "tight" around the tool. It's hard to estimate the error quantitatively but we dont expect it to be more than 5% For surgical steps, there can be some sources of error as this type of annotation is temporal in nature. However, as the robotic surgical training steps are very well defined (as opposed to steps in clinical procedures), we do not expect there to be any significant error in annotations (<5%)

b) In an analogous manner, describe and quantify other relevant sources of error.

The tool presence labels will be generated using the events stream from the da Vinci system. There is a possibility of a dropped event that can cause error in the training tool presence labels. However, we do not expect this to happen frequently. The step labels would not have any other relevant source of error.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Tool detection: Mean average precision (mAP) for different intersection over union (IoU) values 0.50:0.05:0.95 will be used to assess performance of tool bounding box prediction algorithms.

Step recognition: average f-1 score across all steps will be used to assess performance of surgical step prediction algorithms

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

Tool detection: This is the standard metric used for bounding box prediction algorithms (and is also the COCO primary challenge metric). By varying the thresholds, this metric provides a more thorough evaluation of tool localization/keypoint detection accuracy. The surgical step category is a standard classification problem where average f1-score can accurately measure the performance of the models. This metric has proven to be a good measure of model performances in our previous challenges as well.

Step recognition: The surgical step category is a standard classification problem where average f1-score can accurately measure the performance of the models

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

The performance rank will be based on the rank of the evaluation metric (e.g mAP IoU 0.5:0.05:0.95 for bounding box detection, e.g average f-1 score for surgical step
recognition) - the higher the value of this metric, the higher the ranking of that team will be.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Missing results will be penalized and no score will be given for those cases

c) Justify why the described ranking scheme(s) was/were used.

Tool detection: Using the standard metrics being used within the object detection research seems like the right way to rank teams. The spatial detection metric tests the algorithms for detection of objects of different sizes (for tool detection category) which will be useful in differentiating high and low performing teams. Similarly, mean f1-score will reward and penalize the predictions for correct/incorrect predictions accordingly.

Step recognition: Using the standard metrics being used within the activity recognition research seems like the right way to rank teams. For step recognition, we would want to consider precision and recall together for prediction quality, hence mean f1-score will reward and penalize the predictions for correct/incorrect predictions accordingly.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Standard statistical methods to test for significance in results like t-test, ANOVA etc., will be used. Bootstrapping will also be used for uncertainty analysis.

b) Justify why the described statistical method(s) was/were used.

The stated statistical methods are fairly standard and used extensively in literature to test for statistical significance and uncertainty of prediction models.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Tool detection: Additional analysis will be performed on the results to check for ranking variability. On top of the mAP across
multiple IoU values for the first category, we will test ranking of the teams when using individual IoU values (e.g 0.5, 0.6, etc) instead of averaging over all. In addition to this, we will evaluate per class metrics as well.

Step recognition: Additional analysis will be performed on the results to check for ranking variability e.g per class metrics to see if rankings change if any step is ignored in evaluation

TASK 2: RARE: Recognition of Anomalies in low-pREvalance cancer

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Developing computer-aided detection (CADe) systems for cancer detection in low-prevalence scenarios presents a significant challenge. In clinical practice, early-stage cancers are often rare events, vastly outnumbered by normal or non-pathological findings. This inherent class imbalance makes it difficult to train models that are both sensitive to rare anomalies and robust to the overwhelming majority of normal cases. The subtle nature of early cancerous changes further adds to this challenge, as these
anomalies are easily overlooked or misclassified. As a result, models trained on artificially balanced datasets often fail to generalize to real-world clinical conditions, where accurate detection of rare cases is critical for timely intervention and improved patient outcomes.
The challenge centers on developing a classification system that can accurately identify early-stage cancers while maintaining a balance between sensitivity and specificity. Improper evaluation of CADe systems during development can have significant consequences during clinical deployment. For instance, systems that are overly sensitive to rare anomalies may generate an excessive number of false positives. Conversely, systems that are not sensitive enough risk missing early cancer cases, delaying
crucial interventions and adversely affecting patient outcomes. Striking the right balance between sensitivity and specificity is critical, particularly in low-prevalence settings, where false positives are more likely to dominate unless rigorously controlled during evaluation.
Detecting early-stage cancer in Barrett's Esophagus (BE) exemplifies these challenges. Subtle neoplastic changes in BE often go unnoticed during routine endoscopic surveillance, yet early detection is vital. Timely identification enables curative treatment through endoscopic mucosal resection, with long-term remission rates exceeding 90%. In contrast, missed lesions that progress to
advanced cancer result in dire prognoses, with a five-year survival rate of approximately 15%. Despite these high stakes, the prevalence of early neoplasia during surveillance is exceptionally low, complicating data collection and model training. This gap underscores the need for CADe systems that effectively address extreme class imbalance while being rigorously evaluated.

Keywords

List the primary keywords that characterize the task.

Computer-aided detection, Low-prevalence, Recognition, Early-stage cancer detection

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

Tim Jaspers, Cris Claessens, Francisco Caetano, Koen Kusters, Tim Boers, Nikoo Dehghani, Fons van der Sommen (Eindhoven University of Technology, Department of Electrical Engineering, VCA Research Group)

b) Provide information on the primary contact person.

Tim Jaspers (t.j.m.jaspers@tue.nl)

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Repeated event with annual fixed conference submission deadline

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI 2025/EndoVis

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

TBA

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Publicly available data is allowed

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard

d) Define the award policy. In particular, provide details with respect to challenge prizes.

We will provide several awards depending on the availability of sponsoring.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

A live leaderboard will be available throughout the challenge. The top three methods will be publicly announced, and the organization team will also recognize innovative solutions.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Team member will be asked to participate in a shared publication and can be listed as author with sufficient contributions. Participating teams may publish their own results after the publication of the shared publication.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Algorithm container submission (type 2) on Grand Challenge.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

A small validation set is provided, each team has unlimited submission on this set. On the final test set each teams can only submit their final solution.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Release of training data: 1st of May 2025
Intention to submit:15st of July 2025

Validation phase: 1th of August 2025 (unlimited submission on the validation set)
Final Test phase: Starting 1th of September 2025 (1 week to submit final model on the test set)
Challenge Day: Day of Endovis 2025

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

All data is acquired under Dutch Trial Register number NL8411. Data collection from each center involved review of the data collection plan by the local institutional review boards or medical ethics committee. To maintain privacy and comply with data protection regulations, strict anonymization of images was carried out as described in the previous paragraphs. This process ensured that the data was no longer considered 'personal data' as defined by the General Data Protection Regulation (GDPR).

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY-NC (Attribution-NonCommercial)

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

All evaluation code will be open-sourced on the challenge GitHub page

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

All code of the participating teams need to be shared with the organizers together with their methodology report.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

The availability of the test set labels will be limited to the organizing team from the Eindhoven University of Technology and the Amsterdam Medical Center.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Diagnosis,Screening

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Classification

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Barrett's esophagus surveillance patients.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Endoscopist performing Barrett's esophagus surveillance.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Endoscopy

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

class label (neoplasic lesion, non-dysplasic lesion)

b) ... to the patient in general (e.g. sex, medical history).

No additional patient information will be provided

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Endoscopic image of Barrett's esophagus surveillance patients

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

neoplastic lesions in Barrett's esophagus patients

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Classification of lesions in Barrett's esophagus patients. These lesions can be either non dysplastic or neoplastic.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Exera II, Exera III, and EVIS X1 production lines (Olympus Corp., Tokyo, Japan)

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Training data:
The data was acquired retrospectively using an automated query from EndoBase, designed as a broad retrieval process to collect all stored images within the system. From this dataset, all Barrett's esophagus-related images were manually identified and extracted. No standardized data acquisition protocol was applied during the original imaging process, meaning that imaging parameters, techniques, and
device settings may have varied depending on clinical practices and operator preferences at the time of acquisition.

Test data:
The test data was acquired through two methods. A portion of the data was retrospectively retrieved from archives (EndoBase) using patient lists, which also included pathology results (PA) for each patient. Images were then manually reviewed to determine whether abnormalities were present in each case. The other portion of the test data was collected prospectively, adhering to our standardized prospective acquisition protocol. For further details on the test data acquisition process, please refer to Fockens et al. (2023).

Fockens KN, Jong MR, Jukema JB, et al. A deep learning system for detection of early Barrett's neoplasia: a model development and validation study. Lancet Digital Health. 2023;5(12):e905-e916. doi:10.1016/S2589-7500(23)00199-1.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

The training set is constructed with data from: Medical Center Spectrum Twente, Medical Center Spaarne,

The test set, on the other hand, includes data from 12 medical centers that are part of the BONSAI consortium (Fockens et al. (2023)).

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

The training data was acquired during routine endoscopic surveillance in two medical centers (Medical Center Spaarne, Medical Center Spectrum Twente) . No specific group of experts was designated for the data acquisition process, as it involved standard clinical practice.
Consequently, we cannot provide detailed information regarding the expertise level of the medical professionals involved in acquiring the data.

The test data is acquired by expert endoscopist specifically focused on Barrett's esophagus screening.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case in this challenge represents a single endoscopic image with the respective label nondysplastic or neoplastic for the training, validation and test set.

b) State the total number of training, validation and test cases.

• Training set: 3,226 nondysplastic images and 100 neoplasia images, representing a 3% neoplasia ratio.
• Validation set: 1000 nondysplastic images and 100 neoplasia images. This small set is designed solely for participants to verify that their submission process is working. It is not intended to serve as a reliable indicator of model performance. The evaluation methodology remains consistent with the test set.
• Test set: +-20,000 nondysplastic images and 3,339 neoplasia images.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The training set contains predominantly non-dysplastic lesions due to their availability and prevalence, reflecting real-world scenarios. The test set mimics real-world prevalence but includes enough neoplastic images to evaluate model performance adequately. The validation set has the same ratio of neoplastic to NDBE cases as the test set but is smaller in size.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

The class balance in the training set reflects the real-world distribution. However, the test set contains a higher proportion of neoplastic lesions to better evaluate the models' ability to differentiate between non-dysplastic and neoplastic lesions.

e) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

However for evaluation, we aim to simulate a real-world class distribution using a 1:100 neoplasia-to-nondysplastic ratio. To achieve this, we will:

• Select 300 neoplasia cases for each evaluation iteration.
• Sample 100 times as many nondysplastic cases (30,000 images) with replacement to match the target ratio.
• Repeat this process 1,000 times, ensuring that the full extent of the test set is utilized.
• Calculate the median PPV@RECALL=0.9

This approach maintains the test set's full resolution while effectively reflecting a real-world deployment scenario.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

For training data, class label ground truth was based on expert assessment. Each image was evaluated by at least 1 assessor. In
case of doubt, a second assessor was consulted. Imagery was labeled as neoplastic if there was a visible lesion suspicious for
early neoplasia, or as NDBE if there was no visible lesion.

For test data, class label ground truth was two-fold. First, similarly to the training data by visual inspection. In addition, histopathological assessment was included. Imagery was labeled as neoplastic if there was a visible lesion and histology of the endoscopic resection specimen revealed high-grade dysplasia or adenocarcinoma, or as NDBE if there was no visible lesion and all random biopsies were negative for neoplasia. All included patients were treatment-naïve for Barrett's esophagus.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

All annotators had extensive prior experience in data annotation through their involvement with a research consortium dedicated to developing AI for Barrett's esophagus patients for over a decade. Consequently, no additional instructions were deemed necessary.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

All annotators have extensive clinical experience (4+ years) in the field of Barrett's esophagus endoscopy. Edge cases were evaluated by an expert endoscopist with over 20 years of experience with the diagnosis and treatment of early Barrett's neoplasia.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

N/A

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

The raw training data undergoes minimal pre-processing before being provided to the participating teams. Specifically, black edges are cropped from the images to remove extraneous borders. No additional pre-processing techniques are applied to the data. If necessary, the same approach is consistently applied across training, validation, and test cases to ensure uniformity.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

In the training set class labels are based on visual examination by three experience Barrett's esophagus researchers. Each neoplastic case is also evaluated by a endoscopic working a Barrett's expert center. In the test and validation set all class labels are based on pathological assessment and no errors are expected.

b) In an analogous manner, describe and quantify other relevant sources of error.

N/A

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

The primary and sole ranking metric will be the prevalence corrected formula regarding class imbalance: PPV_corrected = (Recall * Prevalence) / (Recall * Prevalence + (1 - Specificity)*(1 - Prevalence))

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The primary goal of the system is to detect neoplastic lesions in images, where sensitivity (recall) is critical to avoid missing tumors, which could lead to undiagnosed or untreated cancers. Focusing on PPV at recall = 0.9 ensures high sensitivity while maintaining precision, reducing the risk of false positives and unnecessary tests.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are

aggregated to arrive at a final score/ranking.

Average PPV@RECALL=0.9 value over 1000 Bootstrap samples.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Any missing results will lead to the metric being set to 0.

c) Justify why the described ranking scheme(s) was/were used.

Since the test set is still relatively small we opted to use a bootstrapping approach for generating the rankings across the methods.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

The analysis utilized mean and confidence intervals of the AUC (area under the curve) over bootstrap sampling (95% CI, 2-sigma) to estimate model performance variability and robustness. Missing data was not applicable in this challenge, and rankings variability was assessed using the same bootstrap approach.

b) Justify why the described statistical method(s) was/were used.

Bootstrap sampling was chosen as it provides a non-parametric way to assess variability and confidence in the rankings without relying on assumptions about the data distribution, making it particularly suited for performance metrics like AUC.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Additionally, a novel analysis could evaluate the ability of models to localize anomalies, though this would require designing models explicitly for anomaly identification, which may be a future challenge focus.

TASK 3: STIR: Surgical Tissue Tracking Using the STIR (Surgical Tattoos in Infrared) Dataset

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

This challenge shall quantify efficient methods for tracking and reconstruction of surgical tissues in videos. This challenge will help enable robust quantification of tracking and mapping methods over many scenes. This is essential to verify methods for use in image guidance and automation. Datasets that have been developed thus far either use rigid environments (not general), visible markers (visible to algorithms), or require annotators to label salient points in videos after collection (costly).
To address this gap, we would like to use a withheld component of our dataset (Surgical Tattoos in Infrared (STIR), https://dx.doi.org/10.21227/w8g4-g548) for a tissue tracking challenge. This component is separate from the STIR 2024 withheld data.
STIR includes labels that are persistent but invisible to visible spectrum algorithms and comprise hundreds of stereo video clips in both in vivo and ex vivo scenes with start and end points labelled in the IR spectrum. Submissions including 2D tracking, 3D tracking, and dynamic neural radiance fields are welcomed. Submissions shall take in input locations for a video and output the final location of estimated location of points in each video. Error will be calculated between the estimated end positions, and that of the ground truth.

Keywords

List the primary keywords that characterize the task.

tracking, deformable, reconstruction

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

Adam Schmidt (Intuitive), Mert Karaoglu (ImFusion), Omid Mohareri (Intuitive)

b) Provide information on the primary contact person.

Adam Schmidt (Intuitive): adam.schmidt@intusurg.com

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Repeated event with annual fixed conference submission deadline

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI 2025/EndoVis

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

synapse for the challenge, IEEE DataPort for dataset hosting

c) Provide the URL for the challenge website (if any).

The sub-challenge site will be on synapse. It will be similar to last years (https://www.synapse.org/Synapse:syn54126082/wiki/626617)

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Publicly available data is allowed

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Organization members may participate, but they will not be eligible for awards. They will be exempt from ranking, although listed on the leaderboard without numbering.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

The challenge will include a financial prize which is dependent on sponsors. A team can win multiple awards but may only have one submission per category.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

The methods and results will be announced publicly at the EndoVis Challenge at MICCAI 2025

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All the methods and results will be included in a joint publication, with all team members as authors. Participants may publish their challenge results only after the challenge is completed with citation of the STIR dataset.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participants must be provide a link to a docker image which uses our evaluation pipeline (available on github, see item 9). For submission, teams will upload their algorithm that they integrate with our evaluation pipeline. The metrics will be evaluated automatically by our evaluation script.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Members can only assess their methods on the openly available validation set. This can be done through the synapse website with a live leaderboard. Docker instructions and an example script will be provided for both the results submission and the benchmarking submission.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

March 31st: Participant Registration, and initial website release. Challenge details uploaded to website, along with an updated test evaluation script to github and tracker for the currently available STIR dataset. STIR 2024's test set will be released for use as a validation set.

April 15th: Submission instructions will be posted.

May 30th: Validation runs will be opened for submission on synapse.

August 20th: Submission deadline.

August 27th: Report Submission Deadline

Sept 23-27: 2025 EndoVis Challenge at MICCAI, results will be released.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

N/A

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY (Attribution)

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The evaluation code is publicly available on github along with an example tracker.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Participants must share their code and models with the challenge organizers. Participants must agree to publish their code online after the challenge.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Sponsorship will be provided by the organizers. Only the organizers will have access to the test data. Participants will not be able to access said data. The authors have no conflicts to report.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Intervention planning,Research,Assistance

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Tracking

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The target cohort is humans with a focus on applications in tracking mapping and diagnostics in surgery. Additional details for any of the following answers can be found at: https://dx.doi.org/10.21227/w8g4-g548

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The challenge cohort comprises in vivo and ex vivo stereo footage from porcine labs with many different tissue types.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Rectified stereo RGB videos from a da Vinci Xi endoscope are used for algorithm testing. Labelled infrared stereo pairs are used for quantification.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Each stereo video pair includes calibration, and image IR labels as described in https://dx.doi.org/10.21227/w8g4-g548

b) ... to the patient in general (e.g. sex, medical history).

N/A

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The target is surgical tissue in the abdomen, with final applications being in minimally invasive surgery. The challenge cohort only includes in vivo and ex vivo animal tissues, while the target cohort is human.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The challenge has been designed to target performance on deformable surgical tissue of any type.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

We would like to assess robustness and accuracy of algorithms under many different difficult scenes which include discontinuity of movement and tissue occlusion. We additionally will emphasize the importance of computational efficiency.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

For the data provided, the stereo video is captured using a da Vinci Xi endoscope (see https://dx.doi.org/10.21227/w8g4-g548)

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

The data acquisition is detailed in depth at https://dx.doi.org/10.21227/w8g4-g548 A da Vinci Xi endoscope is used to collect calibrated stereo data along with infrared images for the ground truth labels.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Intuitive Surgical

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Refer to 21a for camera characteristics. As for the level of expertise, this data is collected by clinician engineers with hundreds of hours of experience using da Vinci robots.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case is a video clip with ground truth infrared start and end points. Training and test cases are the same in structure apart from being captured on different scenes. Test cases will only be usable through the grand challenge interface.

b) State the total number of training, validation and test cases.

There are >1000 training cases, and ~60 test cases. Validation will have ~60 cases (STIR 2024 Test set).

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The training and test cases are separated to provide a large amount for training while still leaving a reasonable size for robust evaluation in testing and validation. The total number of cases is limited by the number of labelling experiments we performed.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

The class distributions of training and test are the same since they are sampled equally. We would like to quantify the performance of methods in environments like those they are trained for.

e) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

We used one test split for the STIR challenge in 2024 that was unseen and unpublished. For STIR 2025 we will be using another unseen and unpublished test split that has zero intersection with either of: the original validation data that was released as a TMI paper; or the STIR challenge 2024 test data that should be released in the coming months.

The test set for STIR 2025 will be ~60 videos of unseen and unpublished videos, different from, and not in the 2024 dataset, or the public validation dataset. The number of cases and characteristics will be like those in STIR 2024:
· Similar in/ex vivo distribution to STIR 2024
· Similar types of movements to STIR 2024

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

No annotators were used per-se. One person is used for filtering outlier data (IR frames that did not include labels/etc). Test cases will be verified again by the challenge organizers for validity.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

Detailed instructions and protocol are in the dataset paper: https://dx.doi.org/10.21227/w8g4-g548

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

The data was filtered by Adam Schmidt who has worked with da Vinci systems for over 5 years.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

N/A

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

The image filtering methodology that processes Infrared labels into center labels is described in https://dx.doi.org/10.21227/w8g4-g548

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

The primary possible error is bulk tissue movement between the IR frame capture and visible light start, as mentioned in the paper at https://dx.doi.org/10.21227/w8g4-g548 . This possible error is the same for each split and should be <1mm.

b) In an analogous manner, describe and quantify other relevant sources of error.

No other error is expected.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

We will use <delta_avg^x as is introduced in the TAP-Vid Benchmark [3] which is used for evaluation of many modern point trackers [2, 4, 5]. This is chosen rather than multi-object-tracking metrics [1] which focus on multiple object regions. <delta_avg^x measures the fraction of points closer than a specified threshold distance and is averaged over multiple thresholds. Similar to popular methods [2, 3, 4, 5], we choose thresholds of 4, 8, 16, 32, and 64 pixels.

Median trajectory error [5] and mean trajectory error will also be reported, but not used for ranking. 2D tracking methods will be evaluated in pixels, and 3D will be in millimetres.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

These were chosen as we are interested in physical accuracy for tissue tracking, so metrics such as IOU or association are not useful for points. We desire metrics that are commonly used and standard for point tracking. <δ_avg^xhas become the most used metric for this purpose from our survey of the point tracking literature.

[1] J. Luiten et al., “HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking,” Int J Comput Vis, vol. 129, no. 2, pp. 548–578, Feb. 2021, doi: 10.1007/s11263-020-01375-2.
[2] M. Neoral, J. Šerých, and J. Matas, “MFT: Long-Term Tracking of Every Pixel,” presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 6837–6847
[3] C. Doersch et al., “TAP-Vid: A Benchmark for Tracking Any Point in a Video,” in Advances in Neural Information Processing Systems, Dec. 2022, pp. 13610–13626.
[4] Q. Wang et al., “Tracking Everything Everywhere All at Once,” presented at the Proceedings of the IEEE International Conference on Computer Vision, 2023.
[5] Y. Zheng, A. W. Harley, B. Shen, G. Wetzstein, and L. J. Guibas, “PointOdyssey: A Large-Scale Synthetic Dataset for Long-Term Point Tracking,” presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 19855–19865.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

Algorithms will be ranked based on their performance on <δ_avg^x. There will be separate rankings for best 2D and 3D algorithms. Aggregation will be performed over all points over all images. That means a video clip with 15 points will contribute 50% more to the metric than one with 10 points.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Submissions with missing results will be excluded from evaluation.

c) Justify why the described ranking scheme(s) was/were used.

The ranking by the metric <δ_avg^x is motivated in 26(a)

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,

- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

We plan to perform a statistical analysis to determine a level of uncertainty of metrics such as the median trajectory error and <δ_avg^x by bootstrapping the data samples. We will use this reference https://arxiv.org/abs/1811.12808 as a guide.

b) Justify why the described statistical method(s) was/were used.

Bootstrapping provides a simple way to estimate variance of population level statistics.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

N/A

TASK 4: OSS: Open Suturing Skills

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Efficient and precise surgical skills are essential in ensuring positive patient outcomes. Whereas machine learning-based surgical skill assessment is gaining traction for minimally invasive techniques, this cannot be said for open surgery skills. Open surgery generally has more degrees of freedom when compared to minimally invasive surgery, making it more difficult to interpret. By continuously providing real-time, data driven, and objective evaluation of surgical performance, automated skill assessment has the potential to greatly improve surgical skill training.

Keywords

List the primary keywords that characterize the task.

Skill assessment, open surgery, suturing, artificial intelligence, surgical tracking

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

NCT Dresden: Hanna Hoffmann, Sebastian Bodenstedt, Stefanie Speidel
Essen: Jan Egger
RWTH Aachen: Frank Hölzle, Rainer Röhrig, Behrus Puladi

b) Provide information on the primary contact person.

Hanna Hoffmann (hanna.hoffmann@nct-dresden.de)
Behrus Puladi (bpuladi@ukaachen.de)

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

One-time event with fixed conference submission deadline

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI 2025/EndoVis

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

synapse.org

c) Provide the URL for the challenge website (if any).

TBA

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Publicly available data is allowed

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate and be visible on leaderboard, but will not be eligible for awards

d) Define the award policy. In particular, provide details with respect to challenge prizes.

There will be three tasks in the challenge, the winner of each task will be awarded a prize, if at least 3 teams submit a result for the task. If at least 5 teams participate, the runner-up will also be awarded a prize.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Results of all teams will be first presented at the Endoscopic Vision Challenge meeting at MICCAI 2025. Afterwards, the information will be made available to all participating teams. The results will be made publically available in the form of a joint paper.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

The first author (or shared first authors) of each team will be listed as authors in alphabetical order on the joint challenge paper. Before publication of the joint paper, no results may be published.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Docker container on the Synapse platform. Link to submission instructions: TBA

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Only the final submission of each team will be evaluated

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

June, 1st Release of training data
August, 1st Start of evaluation
September, 1st 11:59pm GMT Submission deadline
September 15th 11:59 PM GMT Write-up and presentation submission deadline
September, 2xth Challenge Day

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

The data used were anonymized, and the collection and publication of the data was authorized by informed consent of each subject. The data collection and the conduct of the original study leading to the dataset were approved by the local ethics committee of the University Hospital RWTH Aachen (approval code EK 352/21 and EK 22-329) and registered, including the study protocol, in the German Clinical Trials Register (DRKS00029307).

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY-NC (Attribution-NonCommercial)

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The script(s) for computing metrics and rankings will be made available with the release of the training dataset.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

To participate in the challenge, each team has to submit a Docker image capable of producing results on the testing examples. The Docker images will not be shared by the organizers. Each team can choose to provide their source code, though they are not required to. Only a paper describing their method is required.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Sponsorship will be provided by the organizers. Only the organizers will have access to the test data. Participants will not be able to access said data. The authors have no conflicts to report.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis

- Research
- Screening
- Training
- Cross-phase
-

Surgical Skill Assessment,Training

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Classification,Prediction,Regression,Detection,Tracking

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Medical and dental students and residents undergoing open surgical suturing training

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Medical and dental students, surgical residents, and specialist participating in training

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Birds-eye-view video stream

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

None

b) ... to the patient in general (e.g. sex, medical history).

None

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Videos in a simulated training setting

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Surgical skill

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Surgical skill classification task 1: Surgical skill classification algorithm with a low expected cost and a high F1 score, classifying the global rating score (GRS) intro four classes (novice: 8-15, intermediate: 16-23, proficient: 24-31, expert: 32-40) Surgical skill classification task 2: Surgical skill classification algorithm with a low expected cost and a high F1 score, classifying the five different scores in the eight different objective structured assessment of technical skill (OSATS) categories task 3: key-point tracking algorithm with a high HOTA,Algorithms will be required to run on the following hardware specifications: GPU RTX A5000, RAM 300 GB, storage < 20GB

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g.

tracking system used in a surgical setting).

Go Pro Hero 5

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Videos are of approximate 5 min in length each and show participants suturing until the time limit has been reached. Two videos are taken from each participant: once before theoretical training and once after training. In addition, a single video was recorded from surgical residents and specialists to expand the skills spectrum.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

University Hospital RWTH Aachen, Department of Oral and Maxillofacial Surgery & Institute of Medical Informatics

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Medical and dental students, surgical residents, and specialists

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

One case is a participant performing one or multiple sutures before the time limit. Each student has a unique identifier and each video as well. Each video is annotated with a Global Rating Score (GRS). Annotations are performed by three raters. For tracking the videos are annotated at 1 fpm (training) or 1 fps (test) by one annotator and verified by a second annotator with segmentation masks of hands and tools as well as keypoints.

b) State the total number of training, validation and test cases.

Skill Assessment: A least 314 training cases and 30 test cases from the previous challenge (MICCAI24) + at least 10 new training and 10 new test cases
Tracking: at least 50 training cases and 5 test cases

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The total number was chosen due to annotation effort, the total number of test cases was chosen to maximize the ability the generalize and evaluate while maintaining a large enough training set.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Task 1 & 2: Details to the dataset composition of the OSS 2025 Challenge can be found in the benchmark paper https://link.springer.com/article/10.1007/s11548-024-03093-3. This is also the same data used in the 2024 OSS Challenge https://www.synapse.org/Synapse:syn54123724/wiki/626561. Further data is still being collected. Therefore, there is no information regarding the exact dataset composition. However, it is expected to have a similar distribution as the previous dataset.
Task 3: This data is also still being collected and annotated; wherefore, there is no exact composition information available.

e) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

Skill Assessment: at least 10 new training and 10 new test cases
Tracking: at least 50 training cases and 5 test cases

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

To account for interrater variability, three human annotators were involved.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

The annotators were trained in the evaluation of the GRS on the basis of several test cases. The GRS is the summary of eight items (respect for tissue, time and motion, instrument handling, suture technique, procedure of surgery and advance planning, knowledge of specific procedure, quality of final product, and overall performance) of the Objective Structured Assessment of Technical Skills (OSATS) and has been used since the late 1990s:
Datta V, Bann S, Mandalia M, et al. The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. Am J Surg 2006; 192:372–8.
Hatala R, Cook DA, Brydges R, et al. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. Adv Health Sci Educ Theory Pract 2015;20:1149–75.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

All raters worked in a blinded fashion and were experienced surgical residents or specialist in oral and maxillofacial surgery with a dual degree in medicine and dentistry.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Annotations from all three raters are given, and it is up to the challenge participants how to merge them for training. For the evaluation, the annotations will be averaged.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

None

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Disagreement on rating - inter-rater agreement of >0.8 measured with pairwise Pearson correlation coefficient

b) In an analogous manner, describe and quantify other relevant sources of error.

None

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Classification task 1: expected cost and average F1
Classification task 2: expected cost and average F1
Tracking task 3: HOTA

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The average F1-score for each class was selected as the F1-score combines both precision and recall and is more useful than accuracy given uneven class distribution.

Expected cost chosen to consider ordinality of classes.

HOTA was chosen to balance accurate object detection and correct localization and tracking in a single metric

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

As each task consists of a scalar metric, the entries will be sorted in ascending order. Rankings for individual metrics will be determined given the order and averaged to one rank per task. Expected cost will be used as a tie breaker for task 1&2. As task 3 only consists of one metric, no aggregation is necessary for ranking.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Only full submissions for each task will be considered

c) Justify why the described ranking scheme(s) was/were used.

As each task is ranked separately, no other scheme seemed sensible

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

For the paper a Wilcoxon Signed Rank test will be performed to determine significance in metrics. We will also examine whether leaving out the cases for each task with the overall best/worst performance will influence ranking

b) Justify why the described statistical method(s) was/were used.

The Wilcoxon Signed Rank test is a nonparametric test that can be used to determine whether two dependent samples were selected from populations, which cannot be assumed to be normally distributed, having the same distribution.
Analysis will be performed to estimate significance via bootstrap sampling and by checking ranking stability on randomly sampled subsets of the test set.
OSS: We will be conducting a Wilcoxson Signed Rank test for each task to verify the significance of the teams' performance rankings on the test set. We will calculate the rank test for each task separately. The Wilcoxson signed rank test is a pairwise test, so we will be comparing each team with all other teams individually. The rank test requires at least 8 test cases, so we will be applying the test on a video level. This means that we will calculate a metric score, if applicable, for each video of the test set and use this metric in the rank test calculation. The metric basis upon which the test will be performed depends on the task
Task 1: The teams are compared pairwise based on each GRS classification given to the video. No further metric can be calculated, so the GRS classification will be directly used in the rank test calculation.
Task 2: Each video is given a score on the OSATS scoring table which consists of eight different categories. We will combine the scores from these eight categories into one metric and use this in the rank test calculation. The test will be performed for each metric, F1 and EC, separately.
Task 3: Each video is scored based on its tracking performance using the HOTA metric. This metric will be used for the rank test calculation.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

The submitted methods will be analyzed for common problems and biases.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

N/A

Further comments

Further comments from the organizers.

N/A