

Challenge for Vision-Language Modeling in 3D Medical Imaging (VLM3D): Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Challenge for Vision-Language Modeling in 3D Medical Imaging (VLM3D)

Challenge acronym

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

VLM3D

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.

Three-dimensional (3D) medical imaging, particularly chest computed tomography (CT), plays a vital role in diagnosing thoracic abnormalities by offering detailed insights into complex anatomical structures. However, interpreting 3D CT data is a time-consuming and challenging process, especially given the increasing global demand for CT scans. While artificial intelligence (AI) has demonstrated significant potential in automating radiological tasks such as report generation and abnormality detection, its application to 3D medical imaging remains limited due to the lack of large-scale, paired datasets and the computational challenges associated with processing 3D data.

To address these challenges, we introduce the Challenge for Vision-Language Modeling in 3D Medical Imaging (VLM3D), built around the open-source CT-RATE dataset. CT-RATE pairs over 50,000 3D chest CT volumes with corresponding radiology reports, annotated for 18 clinically significant abnormalities. The VLM3D Challenge presents participants with four tasks: (1) radiology report generation from chest CT volumes, (2) multi-abnormality classification, (3) self-supervised multi-abnormality localization, and (4) text-conditional 3D chest CT generation.

These tasks address key aspects of radiological diagnostics and treatment planning. Radiology report generation automates the creation of accurate and detailed reports from CT volumes, reducing radiologists' workloads. Multi-abnormality classification enables quicker and more precise detection of pathologies in CT scans. The self-supervised multi-abnormality localization task focuses on identifying specific pathological regions, including pericardial effusion, pleural effusion, consolidation, lung opacity, and lung nodules, without the use of ground-truth labels during training. This task will be evaluated using a manually labeled dataset to assess how effectively models can localize abnormalities in a self-supervised manner. Text-conditional 3D chest CT generation supports the creation of realistic CT volumes from textual descriptions, with applications in data augmentation,

education, and generative modeling research.

The challenge evaluation will utilize two datasets: an internal closed test set containing 2,000 cases and an external closed test set from Boston University Hospital with 1,024 cases. This dual evaluation framework ensures that models are assessed on diverse and clinically relevant data, supporting their real-world applicability. The training and validation datasets are derived from the CT volumes in the CT-RATE dataset.

The VLM3D Challenge aims to advance AI in 3D medical imaging by providing a benchmark for vision-language models. Biomedically, it seeks to improve diagnostic accuracy, streamline radiological workflows, and enhance patient care. Technically, it encourages the development of scalable and generalizable AI systems that integrate visual and textual modalities, fostering innovation and collaboration in research and clinical applications.

Challenge keywords

List the primary keywords that characterize the challenge.

3D Chest CT, Radiology Report, Multimodal AI

Year

2025

Novelty of the challenge

Briefly describe the novelty of the challenge.

This challenge introduces a structured approach to 3D medical imaging by integrating multiple interrelated tasks into a unified framework. These tasks include radiology report generation, volume abnormality classification, abnormality localization, and report-to-volume generation. Leveraging the CT-RATE dataset—the largest publicly available collection of 3D chest CT volumes paired with corresponding radiology reports—the challenge fosters the development of models that effectively bridge the gap between imaging and reporting in clinical practice.

The tasks are designed to address critical aspects of radiological diagnostics and treatment planning. Radiology report generation focuses on automating the creation of detailed and clinically relevant reports from CT volumes, reducing radiologists' workloads. Volume abnormality classification facilitates the detection of pathologies in CT scans, enabling quicker and more accurate diagnoses. Abnormality localization targets the identification of specific pathological regions, such as pericardial effusion, pleural effusion, consolidation, lung opacity, and lung nodules, enhancing the interpretability and reliability of AI models. Report-to-volume generation supports the synthesis of realistic 3D CT images from textual descriptions, with potential applications in data augmentation, education, and generative modeling.

To ensure robustness and generalizability, the challenge will be evaluated on two datasets: an internal closed test set and an external test set from Boston University Hospital. This dual evaluation approach ensures that models are tested on diverse and clinically relevant data, promoting their applicability to real-world clinical settings.

By integrating these tasks, the challenge aims to advance methods in 3D medical imaging, encouraging innovation and collaboration in developing AI-driven solutions.

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.

N/A

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.

We propose a full-day slot for our challenge due to the comprehensive nature of the event, which includes four distinct tasks: 3D CT Generation, Report Generation, Organ Segmentation, and Visual Question Answering (VQA). Each task addresses a critical aspect of 3D medical imaging and involves unique challenges requiring detailed analysis and discussion.

A full-day slot will allow sufficient time to:

1. Present Each Task: Introduce the background, objectives, and dataset details for all four tasks to ensure participants have a clear understanding.
2. Highlight Methodologies: Showcase diverse approaches and solutions proposed by the participating teams for each task.
3. Discuss Results: Analyze the performance of submitted algorithms, focusing on strengths, limitations, and insights specific to each task.
4. Engage in Discussions: Facilitate interactive discussions and Q&A sessions with participants and panelists to explore future directions for research and development in 3D medical imaging.
5. Recognize Excellence: Allocate time to announce and celebrate the top-performing teams for each task.

The extended format will ensure an engaging, in-depth exploration of the tasks and their impact on advancing the field of multimodal 3D medical imaging.

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.

The expected number of participants for this challenge is based on the strong interest already shown in the CT-RATE dataset and the benchmark tasks. The CT-RATE dataset, which comprises 11TB of data, has been accessed by over 1,000 researchers within six months of its release. Furthermore, it has been cited in more than 30 papers during the same period, demonstrating its relevance and impact in the field.

Over 10 research groups have already expressed explicit interest in participating in the benchmark, including several leading teams in AI and medical imaging. Considering this early enthusiasm and the growing adoption of the CT-RATE dataset, we conservatively estimate that at least 25 research groups will participate in the challenge. This includes teams that have previously utilized the dataset and those motivated by the novel tasks and evaluation framework presented in this benchmark.

The diverse and clinically significant tasks, coupled with the opportunity to test models on internal and external datasets, are likely to attract widespread participation from both academia and industry, fostering innovation and collaboration in 3D medical imaging.

Publication and future plans

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

We plan to release an arXiv paper before the announcement of the challenge results, serving as an initial reference for the dataset and its associated tasks. Subsequently, all results will be consolidated and published in a final challenge journal paper, targeting Medical Image Analysis (MedIA) or IEEE Transactions on Medical Imaging (IEEE TMI). All participants with viable submissions will be included as co-authors in the journal paper, recognizing their contributions to the challenge. This approach ensures broad dissemination of the challenge's outcomes while fostering collaboration and acknowledgment of participants' efforts.

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.

The evaluation for the challenge will be conducted on grand-challenge.org.

Presentations will be conducted on-site, requiring the following technical equipment to ensure smooth execution.

- Microphones: For clear audio during presentations and audience interactions.
- Projector and Screen: To display presentation slides, visual results, and rankings.
- Speakers: For enhanced audio during video demonstrations and panel discussions.
- Laptops/Computers: For seamless integration with the projector and to run any necessary software for live demonstrations.

We will coordinate with the event organizers to ensure the availability of these resources and proper setup for an engaging on-site experience.

TASK 1: Radiology Report Generation

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.

Automating radiology report generation for 3D medical imaging is a pivotal step toward advancing diagnostic workflows. While significant progress has been made in report generation for 2D modalities, such as chest X-rays, applying similar methodologies to 3D medical imaging remains underexplored. Chest computed tomography (CT), a cornerstone of 3D imaging, provides detailed insights into a patient's condition but requires labor-intensive, time-consuming manual reporting, often subject to variability. The complexity of volumetric data and the scarcity of paired datasets have been major barriers to developing automated solutions for this critical task.

The radiology report generation task, a key component of the Challenge for Vision-Language Modeling in 3D Medical Imaging (VLM3D), leverages the open-source CT-RATE dataset, which pairs over 50,000 reconstructed chest CT volumes with corresponding radiology reports. This task challenges participants to develop models capable of generating clinically accurate and descriptive radiology reports directly from 3D chest CT volumes. Participants are encouraged to explore innovative approaches to overcome challenges such as processing volumetric data, capturing clinical nuances, and ensuring that generated reports meet diagnostic standards.

This task aims to drive innovation with profound impacts on both biomedical and technical domains.

Biomedically, automated report generation has the potential to enhance diagnostic efficiency, reduce reporting variability, and alleviate radiologists' workloads, ultimately improving patient outcomes. Technically, the task provides a platform to advance state-of-the-art AI methodologies, fostering the development of robust, scalable, and generalizable vision-language models for 3D medical imaging. By addressing these challenges, the task paves the way for transformative research and clinical applications in radiology.

Keywords

List the primary keywords that characterize the task.

Automated Reporting, Natural Language Generation, Radiology Report

ORGANIZATION

Organizers

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

University of Zurich, Switzerland

- Ibrahim Ethem Hamamci

- Suprosanna Shit

- Anjany Sekuboyina

- Murong Xu

- Chinmay Prabhakar
- Bjoern Menze

University Hospital Zurich, Switzerland

- Christian Bluethgen

Istanbul Medipol University, Turkey

- Sezgin Er
- Ayse Gulnihan Simsek
- Omer Faruk Durugol
- Seval Nil Esirgun
- Muhammed Furkan Dasdelen
- Neslihan Simsek
- Gulhan Ertan Akan

Boston University, USA

- Chenyu Wang
- Weicheng Dai
- Kayhan Batmanghelich

Harvard University, USA

- Xiaoman Zhang
- Pranav Rajpurkar

Johns Hopkins University, USA

- Pedro R. A. S. Bassi
- Wenxuan Li
- Alan Yuille
- Zongwei Zhou

Imperial College London, UK

- Hadrien Reynaud
- Bernhard Kainz

Shanghai Jiao Tong University, China

- Chaoyi Wu
- Weidi Xie

National Institutes of Health (NIH), USA

- Benjamin Hou
- Zhiyong Lu

NVIDIA, USA

- Daguang Xu
- Dong Yang
- Pengfei Guo

b) Provide information on the primary contact person.

Ibrahim Ethem Hamamci, MD - University of Zurich, Switzerland (ibrahim.hamamci@uzh.ch)

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 as open call challenge

Challenge venue and platform

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

N/A

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).

N/A

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.

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

Will be announced later.

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.

All results from all teams will be made public. The top-performing teams will be recognized and rewarded at MICCAI.

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 team members who contributed to the design of the algorithm will be named co-authors in the final challenge paper (unless the results are retracted, see above). Every participant can publish their algorithms and results independently (in fact, they are encouraged to do so). They can only refer to the quantitative results of other challenge participants after the challenge results are published officially as an arXiv paper draft.

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.

Submission instructions will be on the webpage.

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.

Instructions will be on the webpage.

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

1. 30 March 2025 (12:00 AM EST): Launch of challenge registration.
2. 15 June 2025 (12:00 AM EST): Opening of submissions for the internal and external test sets.
3. 15 August 2025 (12:00 AM EST): Deadline for testing submissions.
4. 20 August 2025 (12:00 AM EST): Invite top-performing teams to prepare presentations and participate in the MICCAI25 Satellite Event.
5. September 2025: Presentation of top teams at the MICCAI25 Satellite Event.

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).

Ethical approval is granted with ethical approval from the Clinical Research Ethics Committee at Istanbul Medipol University (E-10840098-772.02-6841, 27/10/2023) for open sourcing CT-RATE and associated models built on the dataset.

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-SA (Attribution-NonCommercial-ShareAlike)

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 will be made public before the system is open for submission.

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

Not required, however, we encourage the participants to publish their code on GitHub or the challenge platform. Also, the final Docker containers should be available on request for research purposes.

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.

Currently, there is no explicit sponsoring of the challenge. Access to test cases will only be given to individual members of the organizers involved in the evaluation process.

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

Decision support,Diagnosis

Task category(ies)

State the task category(ies)

Examples:

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

- Segmentation
- Tracking

Report Generation

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 of the task includes patients ranging in age from 18 to 102 years, drawn from Istanbul Medipol University. These patients were included in the internal validation and training sets.

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

All patients above 12 years of age are included in the data acquisition process. To ensure patient privacy and confidentiality, patients were randomly selected from the hospital's database without considering or utilizing any personal information such as name, gender, age, address, etc. One or more CT scans from each unique patient is included in the data acquisition process. More information about the demographics of randomly chosen patients cannot be shared due to the hospital's privacy rules.

Imaging modality(ies)

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

Computed tomography (CT)

Context information

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

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

In addition to radiology reports, all DICOM metadata—including patient positions, resolutions, spacings, CT machine details, reconstruction kernel, and more—was provided alongside the images. Additionally, 18 abnormality labels were included.

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

The patient's age and sex are provided alongside the images.

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 thoracic region, including the lungs and mediastinum, is depicted in the computed tomography (CT) data.

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 algorithm target is the radiology report, which describes abnormalities and normal findings in the chest CT volume.

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)).

Accuracy, Precision, Sensitivity, Specificity, Runtime

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).

61.5% of the scans in the training set were acquired using Philips CT scanners, 30.1% using Siemens scanners, and 8.4% using PNMS (Philips-Neusoft Medical Systems) scanners.

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

The CT chest (non-contrast) protocol serves as an outline for the acquisition of a chest CT without the use of an intravenous contrast medium.

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 dataset, validation dataset, and internal test dataset were acquired from Istanbul Medipol University Mega Hospital. The external test dataset was acquired from Boston University Hospital.

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).

CT volumes were acquired by experienced CT technicians using Philips, Siemens, or PNMS CT scanners. All associated reports were written by senior radiologists.

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.

The training, validation, internal test, and external test sets each consist of a CT image and the corresponding radiology report. The desired algorithm output is the radiology report for the given CT images.

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

The CT-RATE dataset includes 50,188 reconstructed CT volumes from 25,692 distinct CT experiments conducted on 21,304 unique patients. We divided the cohort into two groups: 20,000 patients are allocated to the training set, and 1,304 patients are allocated to the validation set. We also acquired 2,000 unique CT scans from the same hospital for the internal test set. Additionally, we included 1,024 unique CT scans from Boston University Hospital as the external 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 validation sets are chosen as described in the main paper where the CT-RATE dataset is introduced. Internal test cases were selected to be 10% of the training dataset, and the external test set comprises 5%.

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.

To ensure that the dataset aligns with real-world clinical scenarios, the CT volumes are randomly selected from a diverse clinical population. This random selection strategy is designed to maintain the natural distribution of abnormalities as reflected in the corresponding radiology reports. Such an approach ensures that the class distribution in the dataset mirrors the actual prevalence and diversity of conditions observed in clinical practice. By preserving the real-world class distribution, the training process better reflects the challenges of clinical diagnosis, improving the robustness and reliability of the model in real-world applications.

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.

2,000 and 1,024 new unique CT scans are used for the internal test dataset and external dataset, respectively. The training and validation sets are from the open-source CT-RATE dataset.

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.

Radiology reports are generated by senior radiologists at Istanbul Medipol University Hospital as part of daily clinical practice for the training, validation, and internal test sets. For the external test set, radiologists at Boston University Hospital generated the radiology reports.

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.

Annotators are not provided with any prior information for report generation, mirroring the standard practice followed by senior radiologists in their daily workflow. This ensures that the report generation process remains consistent with real-world clinical conditions, where radiologists must interpret and generate reports based solely on the available imaging data.

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.

Reports were generated by radiologists with at least 10 years of experience in the field.

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.

Chest CT volumes are provided directly from the PACS server without preprocessing. All relevant information, including rescale slope, rescale intercept, and spacings, is provided for preprocessing.

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.

Errors might be caused by the translation of reports from Turkish to English. To prevent this, translations are checked by four bilingual final-year medical students.

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)

Natural Language Generation (NLG) Metrics:

- BLEU (Bilingual Evaluation Understudy): Evaluates n-gram overlap between the generated and reference text to measure fluency and relevance. BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores are reported to capture performance at varying n-gram lengths.
- METEOR (Metric for Evaluation of Translation with Explicit Ordering): Assesses semantic similarity by considering exact matches, synonyms, stemming, and word order alignment.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Measures recall and overlap, including n-grams and longest common subsequences, to evaluate content completeness and coherence.

Clinical Accuracy Metrics:

- Precision: Evaluates the proportion of correctly predicted abnormality labels among all predicted positives, reflecting the system's ability to minimize false positives.
- Recall (Sensitivity): Measures the proportion of actual abnormality labels correctly predicted, assessing the ability to detect true abnormalities.
- F1 Score: The harmonic mean of precision and recall, balancing both metrics to provide a single measure of accuracy.

These metrics collectively assess the algorithm's ability to generate clinically relevant, high-quality reports, aligning with the objectives of both linguistic fluency and clinical correctness.

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

These metrics are widely adopted for evaluating the accuracy of radiology report generation. While alternative approaches based on large language models (LLMs) exist, they typically demand substantial computational resources without delivering a proportionate improvement in performance [cite]. Relying solely on natural language generation (NLG) metrics is insufficient, as these scores can be high even if the generated radiology reports lack clinical accuracy. This issue arises when models overfit to specific sentence structures rather than understanding and accurately generating clinically relevant content.

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.

Inspired by the methodologies employed in the VerSe'19, VerSe'20, and Brain Tumor Segmentation (BraTS) challenges, we will implement a point-based ranking system as described in VerSe.

The ranking process will follow these steps:

Metric Computation: Calculate the relevant evaluation metrics for all test images to assess the performance of each algorithm.

Statistical Analysis: For each metric, conduct a two-sided unpaired permutation test with 10,000 permutations per test image. This statistical approach evaluates the differences in performance between teams.

Point Allocation: Assign a "total point count" to each team based on the number of pairwise comparisons in which they outperform other teams.

Final Ranking: Use the total point counts to determine the final ranking of the participating teams.

This structured approach ensures a fair and statistically robust evaluation, enabling an accurate comparison of algorithm performance.

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

Missing predictions from the algorithm (e.g., not providing a predicted radiology report) will implicitly penalize the NLG and clinical accuracy metrics.

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

Similar ranking methods used in BraTS, Medical Segmentation Decathlon, VerSe'19, and VerSe'20 have received positive feedback from participants for their stability in handling outlier performances.

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.

Note that our ranking scheme is inherently based on statistically significant differences between the algorithm's metric values, reflecting the team's performance on the given task (cf. Section 27a). We employ a two-sided unpaired permutation test to determine the statistical significance of these differences.

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

Calculating the corresponding points by usual metrics such as mean or median does not allow us to consider the entire distribution of the performance metric values of the algorithms in all cases. Using statistical significance for evaluating points, on the other hand, allows us to consider case-level performance as a sample of the distribution in which we can compare not only the metric values but also their distributions. We believe such an evaluation would be more robust and stable.

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 ensuing journal article about the challenge will have a detailed analysis of inter-algorithm variability, algorithm-human variability, and an evaluation of the ensemble of algorithms.

TASK 2: Multi-Abnormality Classification

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.

The classification of abnormalities in chest computed tomography (CT) is a crucial component of radiological diagnostics, directly influencing the speed and accuracy of clinical decision-making. As a 3D imaging modality, chest CT offers detailed insights into thoracic anatomy and pathology, making it a cornerstone for diagnosing complex conditions. However, manually analyzing these images to classify multiple abnormalities is time-intensive and prone to variability, especially given the increasing global demand for CT imaging. The need for automated, reliable solutions to detect and classify multiple pathologies simultaneously is more pressing than ever.

As part of the Challenge for Vision-Language Modeling in 3D Medical Imaging (VLM3D), the multi-abnormality classification task utilizes the CT-RATE dataset, which includes over 50,000 reconstructed chest CT volumes annotated for 18 clinically significant abnormalities. Participants are tasked with developing AI models to classify multiple abnormalities from 3D chest CT volumes, addressing real-world scenarios where scans often exhibit several co-occurring conditions, such as pericardial effusion, pleural effusion, consolidation, lung opacity, and lung nodules.

The multi-abnormality classification task aims to enhance diagnostic efficiency and accuracy by automating the detection of thoracic pathologies. From a biomedical perspective, this task supports faster diagnosis and reduces the workload on radiologists, enabling them to focus on complex interpretative tasks. From a technical standpoint, it challenges participants to design robust and scalable models capable of handling multi-label classification in complex 3D datasets. By addressing these challenges, the task contributes to advancing AI methods for clinical applications, paving the way for improved diagnostic workflows and better patient outcomes.

Keywords

List the primary keywords that characterize the task.

Chest CT, Abnormality Classification, Multi-Label Classification

ORGANIZATION

Organizers

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

University of Zurich, Switzerland

- Ibrahim Ethem Hamamci

- Suprosanna Shit

- Anjany Sekuboyina

- Murong Xu

- Chinmay Prabhakar
- Bjoern Menze

University Hospital Zurich, Switzerland

- Christian Bluethgen

Istanbul Medipol University, Turkey

- Sezgin Er
- Ayse Gulnihhan Simsek
- Omer Faruk Durugol
- Seval Nil Esirgun
- Muhammed Furkan Dasdelen
- Neslihan Simsek
- Gulhan Ertan Akan

Boston University, USA

- Chenyu Wang
- Weicheng Dai
- Kayhan Batmanghelich

Harvard University, USA

- Xiaoman Zhang
- Pranav Rajpurkar

Johns Hopkins University, USA

- Pedro R. A. S. Bassi
- Wenxuan Li
- Alan Yuille
- Zongwei Zhou

Imperial College London, UK

- Hadrien Reynaud
- Bernhard Kainz

Shanghai Jiao Tong University, China

- Chaoyi Wu
- Weidi Xie

National Institutes of Health (NIH), USA

- Benjamin Hou
- Zhiyong Lu

NVIDIA, USA

- Daguang Xu
- Dong Yang
- Pengfei Guo

b) Provide information on the primary contact person.

Ibrahim Ethem Hamamci, MD - University of Zurich, Switzerland (ibrahim.hamamci@uzh.ch)

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 as open call challenge

Challenge venue and platform

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

N/A

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).

N/A

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.

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

Will be announced later.

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.

All results from all teams will be made public. The top-performing teams will be recognized and rewarded at MICCAI.

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 team members who contributed to the design of the algorithm will be named co-authors in the final challenge paper (unless the results are retracted, see above). Every participant can publish their algorithms and results independently (in fact, they are encouraged to do so). They can only refer to the quantitative results of other challenge participants after the challenge results are published officially as an arXiv paper draft.

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.

Submission instructions will be on the webpage.

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.

Instructions will be on the webpage.

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

1. 30 March 2025 (12:00 AM EST): Launch of challenge registration.
2. 15 June 2025 (12:00 AM EST): Opening of submissions for the internal and external test sets.
3. 15 August 2025 (12:00 AM EST): Deadline for testing submissions.
4. 20 August 2025 (12:00 AM EST): Invite top-performing teams to prepare presentations and participate in the MICCAI25 Satellite Event.
5. September 2025: Presentation of top teams at the MICCAI25 Satellite Event.

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).

Ethical approval is granted with ethical approval from the Clinical Research Ethics Committee at Istanbul Medipol University (E-10840098-772.02-6841, 27/10/2023) for open sourcing CT-RATE and associated models built on the dataset.

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-SA (Attribution-NonCommercial-ShareAlike)

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 will be made public before the system is open for submission.

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

Not required, however, we encourage the participants to publish their code on GitHub or the challenge platform. Also, the final Docker containers should be available on request for research purposes.

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.

Currently, there is no explicit sponsoring of the challenge. Access to test cases will only be given to individual members of the organizers involved in the evaluation process.

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.

The target cohort of the task includes patients ranging in age from 18 to 102 years, drawn from Istanbul Medipol University. These patients were included in the internal validation and training sets.

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

All patients above 12 years of age are included in the data acquisition process. To ensure patient privacy and confidentiality, patients were randomly selected from the hospital's database without considering or utilizing any personal information such as name, gender, age, address, etc. One or more CT scans from each unique patient is included in the data acquisition process. More information about the demographics of randomly chosen patients cannot be shared due to the hospital's privacy rules.

Imaging modality(ies)

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

Computed tomography (CT)

Context information

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

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

In addition to radiology reports, all DICOM metadata—including patient positions, resolutions, spacings, CT machine details, reconstruction kernel, and more—was provided alongside the images. Additionally, 18 abnormality labels were included.

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

The patient's age and sex are provided alongside the images.

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 thoracic region, including the lungs and mediastinum, is depicted in the computed tomography (CT) data.

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 target of the algorithms in this challenge is to generate binary labels indicating the presence or absence of 18 clinically significant abnormalities for each chest CT scan. These abnormalities include critical thoracic pathologies such as pericardial effusion, pleural effusion, lung consolidation, lung opacity, lung nodules, and other key conditions.

The algorithms are designed to analyze the entire 3D CT scan and output a set of binary labels corresponding to each abnormality. This task reflects real-world clinical scenarios where multiple co-occurring conditions are often present in a single scan, necessitating accurate and simultaneous multi-label classification.

The challenge cohort is derived from the CT-RATE dataset, which provides a comprehensive collection of annotated 3D chest CT scans. The dataset's annotations are based on radiologists' assessments and encompass diverse clinical cases, ensuring that the cohort represents a wide range of abnormalities commonly encountered in clinical practice. By focusing on this multi-abnormality classification task, the challenge aims to advance the development of robust algorithms capable of performing reliable and scalable analyses of 3D chest CT data.

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)).

Accuracy, Precision, Sensitivity, Specificity

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).

61.5% of the scans in the training set were acquired using Philips CT scanners, 30.1% using Siemens scanners, and 8.4% using PNMS (Philips-Neusoft Medical Systems) scanners.

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

The CT chest (non-contrast) protocol serves as an outline for the acquisition of a chest CT without the use of an intravenous contrast medium.

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 dataset, validation dataset, and internal test dataset were acquired from Istanbul Medipol University Mega Hospital. The external test dataset was acquired from Boston University Hospital.

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).

CT volumes were acquired by experienced CT technicians using Philips, Siemens, or PNMS CT scanners. All associated reports, written by senior radiologists, were labeled for each abnormality using an automated text classifier (LLM-based). The metrics for the automated classifier demonstrate that it operates with exceptionally high accuracy.

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.

Training, internal validation, internal test, and external test sets represent a CT image and corresponding 18 abnormalities. The desired algorithm output is the binary labels of 18 abnormalities for each CT scan.

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

The CT-RATE dataset includes 50,188 reconstructed CT volumes from 25,692 distinct CT experiments conducted on 21,304 unique patients. We divided the cohort into two groups: 20,000 patients are allocated to the training set, and 1,304 patients are allocated to the validation set. We also acquired 2,000 unique CT scans from the same hospital for the internal test set. Additionally, we included 1,024 unique CT scans from Boston University Hospital as the external 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 validation sets are chosen as described in the main paper where the CT-RATE dataset is introduced. Internal test cases were selected to be 10% of the training dataset, and the external test set comprises 5%.

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.

To ensure that the dataset aligns with real-world clinical scenarios, the CT volumes are randomly selected from a diverse clinical population. This random selection strategy is designed to maintain the natural distribution of abnormalities as reflected in the corresponding radiology reports. Such an approach ensures that the class

distribution in the dataset mirrors the actual prevalence and diversity of conditions observed in clinical practice. By preserving the real-world class distribution, the training process can better reflect the challenges of clinical diagnosis, improving the robustness and reliability of the model in real-world applications.

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.

2,000 and 1,024 new unique CT scans are used for the internal test dataset and external dataset, respectively. The training and validation sets are from the open-source CT-RATE dataset.

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.

Abnormality labels were extracted from radiology reports by four final-year medical students for 1,000 reports. Subsequently, an LLM-based abnormality classifier was trained to label the remaining radiology reports with near-perfect accuracy. Of these 1,000 reports, 800 were used to train the model, and 200 were used for evaluation.

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.

Experienced radiologists provided example sentences for specific abnormalities to guide the final-year medical students.

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.

Final-year medical students manually generated 1,000 labels based on radiology reports authored by experienced radiologists.

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.

Chest CT volumes are provided directly from the PACS server without preprocessing. All relevant information, including rescale slope, rescale intercept, and spacings, is provided for preprocessing.

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.

Errors might arise from the accuracy of the label classifier. To mitigate this, we evaluated the classifier on 200 independent radiology reports, demonstrating its near-perfect accuracy on the validation set.

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)

Abnormality classification metrics, including AUROC, F1 Score, Precision, and Accuracy, are utilized to evaluate the performance of models in detecting and categorizing abnormalities in clinical data. These metrics are calculated by comparing the predicted abnormality labels with the ground truth labels, providing a comprehensive assessment of model performance.

- Area Under the ROC Curve (AUROC): AUROC measures the model's ability to differentiate between normal and abnormal cases across varying decision thresholds. By capturing the trade-off between the True Positive Rate and False Positive Rate, AUROC provides a robust indicator of the model's discriminatory power. A higher AUROC reflects superior performance in distinguishing abnormalities.
- F1 Score: The F1 Score, calculated as the harmonic mean of Precision and Recall, balances both metrics to offer a single, unified measure of performance. This is particularly important in clinical contexts where both false positives and false negatives carry significant implications, ensuring a thorough evaluation of the model's effectiveness.
- Precision: Precision evaluates the proportion of correctly identified abnormalities among all instances predicted as abnormal. This metric is critical for minimizing false positives and reflects the model's reliability in accurately confirming true abnormal cases.
- Accuracy: Accuracy assesses the overall correctness of the model's predictions by calculating the proportion of all correct predictions, both normal and abnormal, out of the total predictions. While effective for balanced datasets, Accuracy is complemented by other metrics, such as AUROC and F1 Score, in scenarios involving class imbalance.

Together, these metrics provide a multi-faceted evaluation of model performance, ensuring reliability and clinical relevance.

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

These metrics are widely used for evaluating the performance of abnormality classification models in medical imaging and diagnostic tasks. Relying solely on a single metric, such as accuracy, is insufficient, as it can be misleading in cases of class imbalance, where high scores may result from the overrepresentation of normal

cases. By incorporating a combination of metrics, including AUROC, Precision, Recall, and F1 Score, the evaluation framework ensures a comprehensive assessment of the model's ability to detect clinically relevant abnormalities while balancing sensitivity, specificity, and overall reliability.

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.

Inspired by the methodologies employed in the VerSe'19, VerSe'20, and Brain Tumor Segmentation (BraTS) challenges, we will implement a point-based ranking system as described in VerSe.

The ranking process will follow these steps:

Metric Computation: Calculate the relevant evaluation metrics for all test images to assess the performance of each algorithm.

Statistical Analysis: For each metric, conduct a two-sided unpaired permutation test with 10,000 permutations per test image. This statistical approach evaluates the differences in performance between teams.

Point Allocation: Assign a "total point count" to each team based on the number of pairwise comparisons in which they outperform other teams.

Final Ranking: Use the total point counts to determine the final ranking of the participating teams.

This structured approach ensures a fair and statistically robust evaluation, enabling an accurate comparison of algorithm performance.

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

Missing predictions from the algorithm (e.g., failing to provide predicted abnormality labels) will automatically result in a penalty to the classification metrics.

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

Similar ranking methods used in BraTS, Medical Segmentation Decathlon, VerSe'19, and VerSe'20 have received positive feedback from participants for their stability in handling outlier performances.

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.

Note that our ranking scheme is inherently based on statistically significant differences between the algorithm's metric values, reflecting the team's performance on the given task (cf. Section 27a). We employ a two-sided unpaired permutation test to determine the statistical significance of these differences.

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

Calculating the corresponding points by usual metrics such as mean or median does not allow us to consider the entire distribution of the performance metric values of the algorithms in all cases. Using statistical significance for evaluating points, on the other hand, allows us to consider case-level performance as a sample of the distribution in which we can compare not only the metric values but also their distributions. We believe such an evaluation would be more robust and stable.

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 ensuing journal article about the challenge will have a detailed analysis of inter-algorithm variability, algorithm-human variability, and an evaluation of the ensemble of algorithms.

TASK 3: Self-Supervised Multi-Abnormality Localization

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.

Localizing multiple abnormalities in 3D medical imaging is a critical challenge for improving diagnostic accuracy and fostering explainability in AI-driven healthcare systems. While substantial progress has been made in abnormality detection for 2D modalities such as chest X-rays, the application of similar techniques to 3D imaging remains underexplored. Chest computed tomography (CT), a key modality in 3D imaging, provides detailed insights into a patient's condition. However, manually identifying and localizing abnormalities within volumetric data is labor-intensive, prone to inter-observer variability, and limits scalability. The complexity of 3D data and the scarcity of labeled datasets further hinder the development of automated, clinically applicable solutions.

The Self-Supervised Multi-Abnormality Localization Task, part of the Challenge for Vision-Language Modeling in 3D Medical Imaging (VLM3D), aims to overcome these challenges by leveraging the CT-RATE dataset. Participants are tasked with developing models capable of unsupervised localization of abnormalities, identifying the spatial regions of pathology without requiring extensive annotated data. Abnormality localization is particularly crucial for enhancing the explainability of AI systems, offering clinicians intuitive visualizations of pathological regions and bolstering trust in automated solutions.

This task has significant implications for both biomedical and technical fields. Biomedically, precise localization improves diagnostic workflows by enabling clinicians to quickly identify regions of interest, enhancing efficiency and consistency. Technically, the task pushes the boundaries of self-supervised learning, allowing models to autonomously discover and localize abnormalities within high-dimensional 3D data. By addressing both localization and explainability, this task provides a foundation for transformative advancements in medical AI, enhancing clinical decision-making and patient outcomes.

Keywords

List the primary keywords that characterize the task.

Self-Supervised Learning, Multi-Abnormality Localization, Explainability

ORGANIZATION

Organizers

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

University of Zurich, Switzerland

- Ibrahim Ethem Hamamci

- Suprosanna Shit

- Anjany Sekuboyina

- Murong Xu
- Chinmay Prabhakar
- Bjoern Menze

University Hospital Zurich, Switzerland

- Christian Bluethgen

Istanbul Medipol University, Turkey

- Sezgin Er
- Ayse Gulnihan Simsek
- Omer Faruk Durugol
- Seval Nil Esirgun
- Muhammed Furkan Dasdelen
- Neslihan Simsek
- Gulhan Ertan Akan

Boston University, USA

- Chenyu Wang
- Weicheng Dai
- Kayhan Batmanghelich

Harvard University, USA

- Xiaoman Zhang
- Pranav Rajpurkar

Johns Hopkins University, USA

- Pedro R. A. S. Bassi
- Wenxuan Li
- Alan Yuille
- Zongwei Zhou

Imperial College London, UK

- Hadrien Reynaud
- Bernhard Kainz

Shanghai Jiao Tong University, China

- Chaoyi Wu
- Weidi Xie

National Institutes of Health (NIH), USA

- Benjamin Hou
- Zhiyong Lu

NVIDIA, USA

- Daguang Xu

- Dong Yang

- Pengfei Guo

b) Provide information on the primary contact person.

Ibrahim Ethem Hamamci, MD - University of Zurich, Switzerland (ibrahim.hamamci@uzh.ch)

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 as open call challenge

Challenge venue and platform

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

N/A

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).

N/A

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.

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

Will be announced later.

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.

All results from all teams will be made public. The top-performing teams will be recognized and rewarded at MICCAI.

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 team members who contributed to the design of the algorithm will be named co-authors in the final challenge paper (unless the results are retracted, see above). Every participant can publish their algorithms and results independently (in fact, they are encouraged to do so). They can only refer to the quantitative results of other challenge participants after the challenge results are published officially as an arXiv paper draft.

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.

Submission instructions will be on the webpage.

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.

Instructions will be on the webpage.

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

1. 30 March 2025 (12:00 AM EST): Launch of challenge registration.
2. 15 June 2025 (12:00 AM EST): Opening of submissions for the internal and external test sets.
3. 15 August 2025 (12:00 AM EST): Deadline for testing submissions.
4. 20 August 2025 (12:00 AM EST): Invite top-performing teams to prepare presentations and participate in the MICCAI25 Satellite Event.
5. September 2025: Presentation of top teams at the MICCAI25 Satellite Event.

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).

Ethical approval is granted with ethical approval from the Clinical Research Ethics Committee at Istanbul Medipol University (E-10840098-772.02-6841, 27/10/2023) for open sourcing CT-RATE and associated models built on the dataset.

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-SA (Attribution-NonCommercial-ShareAlike)

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 will be made public before the system is open for submission.

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

Not required, however, we encourage the participants to publish their code on GitHub or the challenge platform. Also, the final Docker containers should be available on request for research purposes.

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.

Currently, there is no explicit sponsoring of the challenge. Access to test cases will only be given to individual members of the organizers involved in the evaluation process.

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, Treatment planning

Task category(ies)

State the task category(ies)

Examples:

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

- Segmentation
- Tracking

Localization

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 of the task includes patients ranging in age from 18 to 102 years, drawn from Istanbul Medipol University. These patients were included in the internal validation and training sets.

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

All patients above 12 years of age are included in the data acquisition process. To ensure patient privacy and confidentiality, patients were randomly selected from the hospital's database without considering or utilizing any personal information such as name, gender, age, address, etc. One or more CT scans from each unique patient is included in the data acquisition process. More information about the demographics of randomly chosen patients cannot be shared due to the hospital's privacy rules.

Imaging modality(ies)

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

Computed tomography (CT)

Context information

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

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

In addition to radiology reports, all DICOM metadata—including patient positions, resolutions, spacings, CT machine details, reconstruction kernel, and more—was provided alongside the images. Additionally, 18 abnormality labels were included.

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

The patient's age and sex are provided alongside the images.

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 thoracic region, including the lungs and mediastinum, is depicted in the computed tomography (CT) data.

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 Self-Supervised Multi-Abnormality Localization Task focuses on identifying and localizing five specific pathological regions within chest CT scans: pericardial effusion, pleural effusion, consolidation, lung opacity, and lung nodules. The task challenges participating algorithms to autonomously detect these abnormalities without the use of ground-truth labels during training. Algorithms must differentiate between pathological and healthy scans, accurately delineating the regions of interest associated with each abnormality. By leveraging self-supervised learning, the task aims to minimize reliance on extensive manual annotations, advancing automated localization techniques in medical imaging.

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)).

Precision, Accuracy, User satisfaction

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).

61.5% of the scans in the training set were acquired using Philips CT scanners, 30.1% using Siemens scanners, and 8.4% using PNMS (Philips-Neusoft Medical Systems) scanners.

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

The CT chest (non-contrast) protocol serves as an outline for the acquisition of a chest CT without the use of an intravenous contrast medium.

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 dataset, validation dataset, and internal test dataset were acquired from Istanbul Medipol University Mega Hospital. There will be no external test dataset for this task.

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).

CT volumes were acquired by experienced technicians using Philips, Siemens, or PNMS CT scanners. The corresponding reports were written by senior radiologists and annotated for each abnormality using an automated text classifier (LLM-based). In the training dataset, localization labels are not available, while in the test dataset, these labels are provided by senior radiologists.

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.

As the task is designed to be self-supervised, both the training and internal validation sets consist solely of CT volumes and corresponding binary pathology labels, without any localization information. The test set, however, includes both CT volumes and pathology localizations. The desired output from the algorithm is the rough localization of abnormalities, achieved using a self-supervised model.

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

The CT-RATE dataset includes 50,188 reconstructed CT volumes from 25,692 distinct CT experiments conducted on 21,304 unique patients. We divided the cohort into two groups: 20,000 patients are allocated to the training set, and 1,304 patients are allocated to the validation set. We also acquired 2,000 unique CT scans from the same hospital for the test set. There will be only one test set for this task.

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

The training and validation sets are chosen as described in the main paper where the CT-RATE dataset is introduced. Test cases were selected to be 10% of the training dataset.

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.

To ensure that the dataset aligns with real-world clinical scenarios, the CT volumes are randomly selected from a diverse clinical population. This random selection strategy is designed to maintain the natural distribution of abnormalities as reflected in the corresponding radiology reports. Such an approach ensures that the class distribution in the dataset mirrors the actual prevalence and diversity of conditions observed in clinical practice. By preserving the real-world class distribution, the training process can better reflect the challenges of clinical diagnosis, improving the robustness and reliability of the model in real-world applications.

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.

2,000 new unique CT scans are used for the test dataset. The training and validation sets are from the open-source CT-RATE dataset.

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.

Abnormality localizations in the test set are labeled by four final-year medical students, who undergo thorough training under the supervision of an experienced radiologist. Following the labeling process, all annotations are reviewed and verified by experienced radiologists to ensure accuracy.

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.

Before labeling the abnormality localizations, final-year medical students receive comprehensive training on the identified pathologies in chest CT volumes, provided by senior radiologists.

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.

For the test set, the abnormality localizations were labeled by final-year medical students, under the supervision and training of experienced radiologists. The labels for the test set were also cross-checked by senior radiologists to ensure accuracy and consistency.

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.

Chest CT volumes are provided directly from the PACS server without preprocessing. All relevant information, including rescale slope, rescale intercept, and spacings, is provided for preprocessing.

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.

Errors may arise from incorrect labeling by the medical students. To mitigate this risk, all labels are carefully reviewed and verified by experienced radiologists.

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)

Localization metrics such as Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Hausdorff Distance, and Sensitivity (Recall) are used to evaluate the performance of models tasked with identifying and delineating abnormal regions without extensive annotated data. These metrics are computed by comparing the predicted abnormal regions to the ground truth.

- Dice Similarity Coefficient (DSC): DSC measures the spatial overlap between the predicted and ground truth abnormal regions by calculating the ratio of their intersection to the total combined area, reflecting the model's accuracy in predicting the lesion area.
- Intersection over Union (IoU) / Jaccard Index: IoU quantifies how well the predicted abnormalities align with the actual lesion areas. It is calculated by comparing the intersection of the predicted and ground truth regions to their union, giving a sense of overlap and model precision.
- Hausdorff Distance: Hausdorff Distance calculates the maximum boundary discrepancy between the predicted and ground truth lesion contours, providing a metric for the model's ability to capture the shapes and edges of lesions accurately.
- Sensitivity (Recall): Sensitivity measures the proportion of correctly identified abnormal voxels among all actual abnormal voxels, highlighting the model's ability to detect true lesions and minimizing false negatives.

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

These metrics are widely recognized and adopted for evaluating the performance of multi-abnormality localization models in medical imaging tasks. Relying on a single metric alone is inadequate, as it may fail to capture the full spectrum of spatial precision, overlap accuracy, and detection reliability necessary for clinical applicability. By incorporating a combination of metrics, such as Dice Similarity Coefficient, Intersection over Union, Hausdorff Distance, and Sensitivity, the evaluation framework offers a comprehensive assessment of the model's ability to detect, delineate, and localize abnormalities. This multifaceted approach ensures that both spatial accuracy and clinical relevance are thoroughly considered, providing a more robust and clinically meaningful evaluation of the model's performance.

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.

Inspired by the methodologies employed in the VerSe'19, VerSe'20, and Brain Tumor Segmentation (BraTS) challenges, we will implement a point-based ranking system as described in VerSe.

The ranking process will follow these steps:

Metric Computation: Calculate the relevant evaluation metrics for all test images to assess the performance of each algorithm.

Statistical Analysis: For each metric, conduct a two-sided unpaired permutation test with 10,000 permutations per test image. This statistical approach evaluates the differences in performance between teams.

Point Allocation: Assign a "total point count" to each team based on the number of pairwise comparisons in which they outperform other teams.

Final Ranking: Use the total point counts to determine the final ranking of the participating teams.

This structured approach ensures a fair and statistically robust evaluation, enabling an accurate comparison of algorithm performance.

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

Missing predictions from the algorithm (e.g., failure to provide predicted localizations) will result in an implicit penalty on the localization metrics.

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

Similar ranking methods used in BraTS, Medical Segmentation Decathlon, VerSe'19, and VerSe'20 have received positive feedback from participants for their stability in handling outlier performances.

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.

Note that our ranking scheme is inherently based on statistically significant differences between the algorithm's metric values, reflecting the team's performance on the given task (cf. Section 27a). We employ a two-sided unpaired permutation test to determine the statistical significance of these differences.

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

Calculating the corresponding points by usual metrics such as mean or median does not allow us to consider the entire distribution of the performance metric values of the algorithms in all cases. Using statistical significance for evaluating points, on the other hand, allows us to consider case-level performance as a sample of the distribution in which we can compare not only the metric values but also their distributions. We believe such an evaluation would be more robust and stable.

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 ensuing journal article about the challenge will have a detailed analysis of inter-algorithm variability, algorithm-human variability, and an evaluation of the ensemble of algorithms.

TASK 4: Text-Conditional CT Generation

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.

The text-conditional generation of 3D chest computed tomography (CT) volumes marks a significant breakthrough in medical imaging, offering the ability to synthesize realistic, high-fidelity 3D CT scans based on free-form medical text prompts. Chest CT is essential in diagnosing a wide range of thoracic conditions, providing in-depth 3D views of anatomy and pathology. However, challenges such as patient privacy concerns, imbalanced class distributions, limited large-scale annotated datasets, and the high costs of manual data curation hinder progress. These barriers highlight the growing need for synthetic data generation, which can revolutionize medical research by providing scalable, diverse datasets for clinical and research applications.

As part of the Vision-Language Modeling in 3D Medical Imaging (VLM3D) Challenge, the text-conditional 3D Chest CT Generation task utilizes the CT-RATE dataset. Participants are tasked with developing generative models that produce clinically realistic chest CT volumes conditioned on medical text descriptions. This task involves not only the accurate generation of anatomically correct 3D CT scans but also the precise alignment of generated images with the input text, making them suitable for downstream applications such as data augmentation, multi-abnormality classification, and model pretraining.

The challenge addresses several key issues in text-conditional 3D medical image generation, including the efficient handling of high-dimensional data, the encoding and decoding of variable-depth 3D CT volumes, and the alignment between text and image modalities. Clinically, it provides a pathway to alleviate data scarcity by generating synthetic data to enhance model training and improve diagnostic workflows. Technically, it encourages innovations in generative architectures and efficient 3D representation learning.

By advancing research in text-conditional 3D image generation, this task lays the groundwork for future developments in medical imaging AI, enabling the creation of robust, scalable, and clinically impactful generative models. Through this effort, the VLM3D Challenge aims to accelerate progress in the field, ultimately enhancing diagnostic capabilities and improving patient outcomes.

Keywords

List the primary keywords that characterize the task.

Text-to-CT, Generative Models, Multimodality

ORGANIZATION

Organizers

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

University of Zurich, Switzerland

- Ibrahim Ethem Hamamci
- Suprosanna Shit
- Anjany Sekuboyina
- Murong Xu
- Chinmay Prabhakar
- Bjoern Menze

University Hospital Zurich, Switzerland

- Christian Bluethgen

Istanbul Medipol University, Turkey

- Sezgin Er
- Ayse Gulnihan Simsek
- Omer Faruk Durugol
- Seval Nil Esirgun
- Muhammed Furkan Dasdelen
- Neslihan Simsek
- Gulhan Ertan Akan

Boston University, USA

- Chenyu Wang
- Weicheng Dai
- Kayhan Batmanghelich

Harvard University, USA

- Xiaoman Zhang
- Pranav Rajpurkar

Johns Hopkins University, USA

- Pedro R. A. S. Bassi
- Wenxuan Li
- Alan Yuille
- Zongwei Zhou

Imperial College London, UK

- Hadrien Reynaud
- Bernhard Kainz

Shanghai Jiao Tong University, China

- Chaoyi Wu
- Weidi Xie

National Institutes of Health (NIH), USA

- Benjamin Hou
- Zhiyong Lu

NVIDIA, USA

- Daguang Xu
- Dong Yang
- Pengfei Guo

b) Provide information on the primary contact person.

Ibrahim Ethem Hamamci, MD - University of Zurich, Switzerland (ibrahim.hamamci@uzh.ch)

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 as open call challenge

Challenge venue and platform

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

N/A

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).

N/A

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.

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

Will be announced later.

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.

All results from all teams will be made public. The top-performing teams will be recognized and rewarded at MICCAI.

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 team members who contributed to the design of the algorithm will be named co-authors in the final challenge paper (unless the results are retracted, see above). Every participant can publish their algorithms and results independently (in fact, they are encouraged to do so). They can only refer to the quantitative results of other challenge participants after the challenge results are published officially as an arXiv paper draft.

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.

Submission instructions will be on the webpage.

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.

Instructions will be on the webpage.

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

1. 30 March 2025 (12:00 AM EST): Launch of challenge registration.
2. 15 June 2025 (12:00 AM EST): Opening of submissions for the internal and external test sets.
3. 15 August 2025 (12:00 AM EST): Deadline for testing submissions.
4. 20 August 2025 (12:00 AM EST): Invite top-performing teams to prepare presentations and participate in the MICCAI25 Satellite Event.
5. September 2025: Presentation of top teams at the MICCAI25 Satellite Event.

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).

Ethical approval is granted with ethical approval from the Clinical Research Ethics Committee at Istanbul Medipol University (E-10840098-772.02-6841, 27/10/2023) for open sourcing CT-RATE and associated models built on the dataset.

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-SA (Attribution-NonCommercial-ShareAlike)

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 will be made public before the system is open for submission.

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

Not required, however, we encourage the participants to publish their code on GitHub or the challenge platform. Also, the final Docker containers should be available on request for research purposes.

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.

Currently, there is no explicit sponsoring of the challenge. Access to test cases will only be given to individual members of the organizers involved in the evaluation process.

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

Education, Research, Training

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling

- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Image Generation

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 of the task includes patients ranging in age from 18 to 102 years, drawn from Istanbul Medipol University. These patients were included in the internal validation and training sets.

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

All patients above 12 years of age are included in the data acquisition process. To ensure patient privacy and confidentiality, patients were randomly selected from the hospital's database without considering or utilizing any personal information such as name, gender, age, address, etc. One or more CT scans from each unique patient is included in the data acquisition process. More information about the demographics of randomly chosen patients cannot be shared due to the hospital's privacy rules.

Imaging modality(ies)

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

Computed tomography (CT)

Context information

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

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

In addition to radiology reports, all DICOM metadata—including patient positions, resolutions, spacings, CT machine details, reconstruction kernel, and more—was provided alongside the images. Additionally, 18 abnormality labels were included.

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

The patient's age and sex are provided alongside the images.

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 thoracic region, including the lungs and mediastinum, is depicted in the computed tomography (CT) data.

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 algorithms for the Text-Conditional 3D Chest CT Generation task are designed to generate anatomically accurate and clinically relevant 3D chest CT volumes based on free-form medical text prompts (radiology reports). The primary focus includes key thoracic structures, such as the lungs, heart, and surrounding tissues, as well as pathologies specified in the input text, such as lung nodules, pleural effusion, atelectasis, and other thoracic abnormalities. The goal is for the generated CT volumes to faithfully reflect both the anatomical and pathological features described in the input prompts, ensuring that the visual representation aligns precisely with the semantic content. This task challenges models to capture the complexity of 3D chest CT imaging, ensuring the synthesis of realistic and clinically interpretable volumes that can support downstream tasks such as data augmentation, multi-abnormality classification, and model pretraining.

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)).

Consistency, Usability, Runtime

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).

61.5% of the scans in the training set were acquired using Philips CT scanners, 30.1% using Siemens scanners, and 8.4% using PNMS (Philips-Neusoft Medical Systems) scanners.

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

The CT chest (non-contrast) protocol serves as an outline for the acquisition of a chest CT without the use of an intravenous contrast medium.

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 dataset, validation dataset, and internal test dataset were acquired from Istanbul Medipol University Mega Hospital. The external test dataset was acquired from Boston University Hospital.

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).

CT volumes were acquired by experienced CT technicians using Philips, Siemens, or PNMS CT scanners. All associated reports were written by senior radiologists.

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.

The training, internal validation, internal test, and external test sets each consist of a radiology report paired with its corresponding CT image. The desired output for the algorithm is the generation of realistic 3D chest CT images that accurately align with the information provided in the conditioned radiology reports. This alignment ensures that the generated CT volumes faithfully represent both the anatomical structures and pathological features described in the text, supporting the model's ability to generate clinically relevant and high-fidelity imaging based on textual input.

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

The CT-RATE dataset includes 50,188 reconstructed CT volumes from 25,692 distinct CT experiments conducted on 21,304 unique patients. We divided the cohort into two groups: 20,000 patients are allocated to the training set, and 1,304 patients are allocated to the validation set. We also acquired 2,000 unique CT scans from the same hospital for the internal test set. Additionally, we included 1,024 unique CT scans from Boston University Hospital as the external 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 validation sets are chosen as described in the main paper where the CT-RATE dataset is introduced. Internal test cases were selected to be 10% of the training dataset, and the external test set comprises 5%.

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.

To ensure that the dataset aligns with real-world clinical scenarios, the CT volumes are randomly selected from a diverse clinical population. This random selection strategy is designed to maintain the natural distribution of abnormalities as reflected in the corresponding radiology reports. Such an approach ensures that the class distribution in the dataset mirrors the actual prevalence and diversity of conditions observed in clinical practice. By preserving the real-world class distribution, the training process can better reflect the challenges of clinical diagnosis, improving the robustness and reliability of the model in real-world applications.

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.

2,000 and 1,024 new unique CT scans are used for the internal test dataset and external dataset, respectively. The training and validation sets are from the open-source CT-RATE dataset.

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.

Radiology reports are generated by senior radiologists at Istanbul Medipol University Hospital as part of daily clinical practice for the training, validation, and internal test sets. For the external test set, radiologists at Boston University Hospital generated the radiology reports.

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.

Annotators are not provided with any prior information for report generation, mirroring the standard practice followed by senior radiologists in their daily workflow. This ensures that the report generation process remains consistent with real-world clinical conditions, where radiologists must interpret and generate reports based solely on the available imaging data.

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.

Reports were generated by radiologists with at least 10 years of experience in the field.

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.

Chest CT volumes are provided directly from the PACS server without preprocessing. All relevant information, including rescale slope, rescale intercept, and spacings, is provided for preprocessing.

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.

Errors might be caused by the translation of reports from Turkish to English. To prevent this, translations are checked by four bilingual final-year medical students.

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)

To assess the performance of text-conditional 3D chest CT generation algorithms, a combination of visual fidelity, semantic alignment, and realism metrics is employed. These metrics collectively evaluate the anatomical accuracy, alignment with text prompts, and overall visual realism of the generated 3D chest CT volumes.

Visual Fidelity Metrics:

- FVDI3D (Fréchet Video Distance): Originally designed for video generation tasks, this metric is adapted for 3D chest CT evaluation. It measures the temporal and spatial coherence of generated volumes by treating CT slices as sequential frames, comparing the feature distribution of the generated volumes to that of real CT data. This metric emphasizes how well the generated volumes maintain continuity and anatomical structure across slices.
- FVDCT-Net: A variant of the Fréchet distance, this task-specific metric is trained directly on chest CT volumes to provide a clinically relevant measure of fidelity. It focuses on the anatomical accuracy and clinical feature alignment of the generated CT volumes, ensuring that generated data matches real-world diagnostic criteria.

Semantic Alignment Metric:

- CT-CLIP: This vision-language model metric evaluates how well the generated 3D CT volumes align with their corresponding text prompts. Fine-tuned specifically for CT data, CT-CLIP measures the semantic consistency between the generated images and the descriptive medical text, ensuring that the images accurately reflect the pathological and anatomical features described in the prompts.

Realism Metric:

- FID (Fréchet Inception Distance): FID quantifies the distributional similarity between real and generated CT volumes, providing an overall measure of the realism and visual quality of the generated images. A lower FID score indicates that the generated CT volumes closely resemble real clinical data, ensuring that the synthesized images are indistinguishable from actual scans in terms of visual fidelity.

Together, these metrics offer a comprehensive evaluation of the algorithms' ability to generate high-fidelity, anatomically accurate, semantically aligned, and visually realistic 3D chest CT volumes based on free-form medical text prompts.

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

imaging and related fields. Relying on a single metric, such as visual realism (FID), is insufficient as it may neglect critical aspects like clinical relevance and semantic alignment necessary for medical applications. By integrating a combination of metrics—such as FVDI3D, FVDCT-Net, CT-CLIP, and FID—the evaluation framework ensures a holistic assessment of generated 3D chest CT volumes. Together, these metrics evaluate visual fidelity, semantic alignment, and anatomical accuracy, offering a comprehensive framework that supports the biomedical objective of generating clinically meaningful and realistic 3D CT volumes.

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.

Inspired by the methodologies employed in the VerSe'19, VerSe'20, and Brain Tumor Segmentation (BraTS) challenges, we will implement a point-based ranking system as described in VerSe.

The ranking process will follow these steps:

Metric Computation: Calculate the relevant evaluation metrics for all test images to assess the performance of each algorithm.

Statistical Analysis: For each metric, conduct a two-sided unpaired permutation test with 10,000 permutations per test image. This statistical approach evaluates the differences in performance between teams.

Point Allocation: Assign a "total point count" to each team based on the number of pairwise comparisons in which they outperform other teams.

Final Ranking: Use the total point counts to determine the final ranking of the participating teams.

This structured approach ensures a fair and statistically robust evaluation, enabling an accurate comparison of algorithm performance.

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

In cases where the algorithm fails to provide predicted CT volumes (e.g., missing predictions), a black volume will be used as a placeholder for evaluation. This approach will implicitly penalize the generation metrics, reflecting the absence of valid predictions.

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

Similar ranking methods used in BraTS, Medical Segmentation Decathlon, VerSe'19, and VerSe'20 have received positive feedback from participants for their stability in handling outlier performances.

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.

Note that our ranking scheme is inherently based on statistically significant differences between the algorithm's metric values, reflecting the team's performance on the given task (cf. Section 27a). We employ a two-sided unpaired permutation test to determine the statistical significance of these differences.

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

Calculating the corresponding points by usual metrics such as mean or median does not allow us to consider the entire distribution of the performance metric values of the algorithms in all cases. Using statistical significance for evaluating points, on the other hand, allows us to consider case-level performance as a sample of the distribution in which we can compare not only the metric values but also their distributions. We believe such an evaluation would be more robust and stable.

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 ensuing journal article about the challenge will have a detailed analysis of inter-algorithm variability, algorithm-human variability, and an evaluation of the ensemble of algorithms.

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