

AMPLIFAI - Annotated Multi-Phase Liver Imaging For Artificial Intelligence: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

AMPLIFAI - Annotated Multi-Phase Liver Imaging For Artificial Intelligence

Challenge acronym

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

AMPLIFAI

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.

Hepatocellular carcinoma (HCC) is the most common primary liver cancer and the third leading cause of cancer-related mortality worldwide, with early detection improving survival from <20% up to >70%.[1,2] Abdominal computed tomography (CT) is one of the primary imaging modalities for HCC diagnosis, with the standardized Liver Imaging Reporting and Data System (LI-RADS) defining specific diagnostic criteria that directly impact clinical decision-making and downstream treatment pathways.[3,4] Specifically, LI-RADS category 5 (LR-5) is considered “definitely HCC” on a scale from LR-1 through LR-5, and does not require further tissue biopsy for diagnosis [5-7] – a fully radiology imaging-based diagnostic pathway, where artificial intelligence (AI) tools can promote early diagnosis and affect patient outcomes.[8]

Compared to other simpler medical imaging classification tasks, HCC diagnosis on abdominal CT presents a unique challenge based on established clinical LI-RADS criteria: The need for multi-phase acquisition requirements with different enhancement patterns (i.e. for arterial and washout assessment). Specifically, LI-RADS characterization depends on the presence of four imaging features: 1) Arterial phase hyperenhancement (APHE), where the lesion appears brighter than the surrounding liver tissue in the early arterial phase and serves as the primary entry point of HCC evaluation; 2) Size of the lesion; 3) Non-peripheral washout, where the lesion becomes darker than the liver in the portal venous or delayed phases; and 4) Enhancing capsule, where a thin, peripheral rim remains enhanced in the portal venous or delayed phases due to contrast retention.

From a technical standpoint, these clinical requirements translate into distinct machine learning challenges: effective LI-RADS characterization for HCC requires spatio-temporal reasoning across contrast phases to capture dynamic enhancement patterns, robust multi-task learning architectures capable of modeling phase-specific lesion characteristics and cross-phase feature interaction, handling substantial class imbalance between LI-RADS categories, and strong out-of-distribution generalization across heterogeneous acquisition protocols.

Dataset: To potentially train AI tools to automate HCC characterization on CT, public multi-phase liver CT datasets exist, however have historically been fragmented, heterogeneous, and limited in size. We have selectively harmonized four public CT datasets [9-13] into a unified, analysis-ready resource for HCC AI research. The dataset consists of 668 cases, including: 83 normal livers with no liver lesions, and 585 livers with HCC lesions. All HCC lesions were re-annotated with clinical LI-RADS categories and voxel-level segmentation masks were created for non-rim APHE, non-peripheral washout, and enhancing capsule by three board-certified abdominal radiologists (F.D., B.L., J.W.).

Task: We propose a challenge focused on advancing AI-based characterization of HCC lesions in multi-phase abdominal CT scans using clinical imaging-based LI-RADS categories. Additionally, all submissions will be evaluated on a held-out private institutional test set to assess generalizability. This challenge aims to provide a standardized benchmark for clinically grounded liver lesion AI and drive methodological advances in multi-phase feature integration, multi-task learning for hierarchical diagnostic criteria, and domain generalization under varied real-world protocol heterogeneity – ultimately accelerating translation of AI tools into clinical workflows, where early, accurate HCC characterization can meaningfully improve patient outcomes.

Challenge keywords

List the primary keywords that characterize the challenge.

Hepatocellular carcinoma, LI-RADS, liver lesion, artificial intelligence, computed tomography

Year

2026

Novelty of the challenge

Briefly describe the novelty of the challenge.

The AMPLIFAI challenge aims at advancing AI-based characterization of HCC lesions using clinical imaging-based LI-RADS categories. Accurate lesion characterization requires spatio-temporal reasoning across multiple contrast phases to capture dynamic enhancement patterns, making the task substantially more complex than single-phase classification problems.

While MRI remains a highly sensitive reference modality for HCC characterization, CT-based imaging continues to play a critical role in routine clinical practice. CT is frequently used in patients who are ineligible for MRI (e.g., due to implanted devices, renal insufficiency, or intolerance), in post-treatment and longitudinal surveillance settings, in resource-limited environments where MRI access is constrained, and in high-volume clinical workflows where report burden and inter-reader variability remain substantial. In these scenarios, reliable CT-based decision support has the potential to improve consistency, reduce diagnostic variability, and support earlier or more confident clinical decision-making.

To achieve this goal, we release the largest publicly available dataset of HCC lesions with corresponding LI-RADS categories, and the first dataset to provide granular, voxel-level segmentation masks for non-rim APHE, non-peripheral washout, and enhancing capsule. These annotations enable the development and evaluation of AI models for LI-RADS categorization on multi-phase abdominal CT.

Task description and application scenarios

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

The purpose of the challenge is to develop an AI model that captures the spatio-temporal relationship between lesion morphology and dynamic enhancement patterns in multi-phase abdominal CT to characterize HCC lesions with clinical LI-RADS categories (LR-1 through LR-5).

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?

Half day

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

Expanded program including dataset overview, method presentations, clinical panel, and future directions.

This is the intended order of proceedings:

1. Opening & Motivation (15 min)

- Welcome and workshop overview
- Clinical motivation for LI-RADS benchmarking
- Positioning CT-based LI-RADS within current clinical workflows

2. Dataset & Evaluation Overview (20 min)

- Dataset construction and harmonization
- Preprocessing and phase handling
- Evaluation protocol, metrics, and ranking

3. Top Challenge Submissions – Method Talks (45–60 min)

Short oral presentations from top 3–5 teams

Format:

8–10 minutes per team + 2 min Q&A;

4. Coffee / Poster & Networking Break (30 min)

Poster viewing (all accepted teams)

Informal discussion with participants and clinicians

5. Invited Talk or Panel: Clinical & Regulatory Perspectives (30–40 min)

Option A – Invited Talk: Radiologist or clinical AI expert on LI-RADS variability, AI readiness, and unmet needs
or

Option B – Panel Discussion:

Topics:

- Interpretability vs performance
- CT vs MRI workflows
- Regulatory and deployment considerations

6. Awards, Future Directions & Closing (15 min)

- Distribution of Awards
- Discussion of future challenge iterations and dataset evolution

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.

We anticipate between 20-30 participating teams based on prior liver CT challenges (e.g., LiTS [17], ATLAS [18], autoPET/CT [19], and TrialS [20]).

Publication and future plans

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

We intend to publish a joint challenge paper describing the task, dataset, and results, in a peer-reviewed journal (e.g., Medical Image Analysis). A 12-month embargo period will apply only to publications that directly disclose or duplicate challenge-level evaluation results, including rankings or comparative performance on the private held-out test set.

During the embargo period, participating teams retain full ownership of their methods and are free to continue developing, extending, and internally disseminating their work. Publications that substantially extend the submitted method or evaluate it outside the official challenge test set are permitted once the embargo period concludes.

This policy is intended to enable a coherent and archival challenge summary publication without restricting independent methodological research or scientific ownership.

MICCAI LNCS proceedings

Indicate if you want to offer MICCAI Springer LNCS proceedings to the participants. Publishing a proceedings volume is optional and at the discretion of each challenge's organizers. At a minimum, organizers must ensure that a description of each participant's submission is publicly available. Organizers who wish to publish MICCAI Springer LNCS proceedings must adhere to the MICCAI Satellite events publication process.

Yes

Collaboration with European Society of Radiology (ESR)

In collaboration with European Society of Radiology (ESR), we announce special clinical interest topics with associated clinicians who can help with the preparation of the proposals; the best 3 challenge proposals on these topics will get the opportunity to present their challenges at the European Congress of Radiology (ECR) 2027 in a special session. If you want to organize a challenge in collaboration with ESR on one of these topics, please reach out to the MICCAI Challenges Team (miccai-challenges-2026@dkfz-heidelberg.de) and we will put you in contact with the corresponding clinician.

Challenge in collaboration with ESR. Ticking 'Yes' implies that the challenge has been prepared in collaboration with the clinical contact point.

No

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.

We will host the competition on CodaLab, utilizing a private competition worker deployed on our institutional on-premises infrastructure. All submissions will be executed on a dedicated node equipped with dual NVIDIA A6000 GPUs. This architecture ensures that the private test dataset remains strictly within institutional premises and never leaves our secure environment.

TASK 1: LI-RADS Characterization

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 AMPLIFAI challenge focuses on automated characterization of hepatocellular carcinoma (HCC) lesions in multi-phase abdominal CT using standardized, clinical imaging-based LI-RADS categories. On a scale from LR-1 to LR-5, LI-RADS categories describe the increasing probability of HCC based on lesion morphology and dynamic contrast enhancement patterns. Specifically, LR-1 indicates a lesion is “definitely benign”, while LR-5 is considered “definitely HCC” and does not require further tissue biopsy for diagnosis. This framework represents a fully imaging-based diagnostic pathway, where AI models have the potential to support early diagnosis and improve patient outcomes.

The LI-RADS framework follows a strict clinical logic (<https://radiologyassistant.nl/assets/tab-1b-overview.png>) consisting of temporal analysis in multi-phase CT, where multiple contrast phases are acquired: before contrast injection (non-contrast) and at successive time points after injection (arterial, portal venous, and delayed) to capture dynamic enhancement patterns in the lesion. The primary gating feature is arterial phase hyperenhancement (APHE), where the lesion appears brighter than the surrounding healthy liver tissue in the early arterial phase. It is considered “non-rim” if the enhancement spreads internally and is not limited to the outer edge of the lesion. Only lesions exhibiting unequivocal non-rim APHE are considered for the highest probability categories (LR-3 to LR-5). However, if the enhancement is limited to the outer edge, the lesion indicates rim APHE and cannot be considered for LR-5.

If APHE is present, the LI-RADS category is determined by a combination of the lesion’s size and how APHE interacts with two major features: First, non-peripheral washout, where the lesion becomes darker than the liver in the portal venous or delayed phases. Second, enhancing capsule, where a thin, peripheral rim remains enhanced in the portal venous or delayed phases due to contrast retention. The lesion size and presence of one or both features increases the probability for HCC. For example, a lesion with non-rim APHE that is 2.3 cm and has one additional feature (washout or capsule) is categorized as LR-5 (“definitely HCC”) while a lesion without non-rim APHE can only reach a maximum of LR-4 (“probable HCC”), even if it is large and shows washout.

The purpose of the challenge is to develop an AI model that captures the spatio-temporal relationship between lesion morphology and dynamic enhancement patterns to characterize HCC lesions from LR-1 through LR-5. Participants will be provided with lesion-level LI-RADS categories and voxel-level segmentations for non-rim APHE, non-peripheral washout, and enhancing capsule. While building sub-classifiers for “APHE” or “Washout” specifically is not required, successful methods will likely need to prioritize modeling phase-specific lesion characteristics and cross-phase feature interaction to replicate the clinical logic.

Keywords

List the primary keywords that characterize the task.

Hepatocellular carcinoma, LI-RADS, liver lesion, artificial intelligence, computed tomography

ORGANIZATION

Organizers

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

Nikhil Shah, University of Maryland Institute for Health Computing, North Bethesda, MD;

Pranav Kulkarni, University of Maryland Institute for Health Computing, North Bethesda, MD;

Amritansh Suryavanshi, University of Maryland Institute for Health Computing, North Bethesda, MD;

Junfeng Guo, University of Maryland Institute for Health Computing, North Bethesda, MD;

Jana Delfino, U.S. Food and Drug Administration, Silver Spring, MD;

Barton F. Lane, University of Maryland School of Medicine, Baltimore, MD;

Jade J. Wong-You-Cheong, University of Maryland School of Medicine, Baltimore, MD;

Heng Huang, University of Maryland Institute for Health Computing, North Bethesda, MD;

Florence X. Doo, University of Maryland Institute for Health Computing, North Bethesda, MD;

b) Provide information on the primary contact person.

Florence X. Doo (fdoo@som.umaryland.edu)

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

The organizing team includes three board-certified radiologists, who have advanced subspecialty training specifically in abdominal imaging:

Florence X. Doo (Senior Clinical/Challenge Lead), Barton F. Lane, and Jade J. Wong-You-Cheong.

Clinical leadership & challenge framework (F.D.): As the senior clinical/challenge lead, Dr. Doo spearheads the intellectual design of the challenge, ensuring the task definition, data selection, and evaluation metrics are rigorously aligned with clinical utility and real-world diagnostic workflows.

Expert annotation for LI-RADS ground truth (F.D., B.L., J.W.): All three radiologists provide the expert "ground truth" for the dataset. This involves complex LI-RADS characterization and high-precision voxel-level segmentation for key imaging features, including Arterial Phase Hyperenhancement (APHE), non-peripheral washout, and enhancing capsule.

Validation: The clinical team will oversee the qualitative assessment of the winning algorithms to ensure the results are interpretatively sound in a radiological context.

Life cycle type

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

Examples:

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

One-time event with fixed conference submission deadline

Challenge venue and platform

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

MICCAI

b) Report the platform used to run the challenge.

The challenge will be hosted via a private CodaBench worker integrated with our institutional compute infrastructure. This setup enables fully containerized evaluation of submissions, ensuring that the private test set remains entirely within our secure environment with zero external data transfer.

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

N/A

Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

Fully automatic, i.e., no user interaction is allowed at any step

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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

Publicly available data is allowed

Publicly available data and pretrained models allowed with mandatory disclosure.

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

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

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

The exact award policy will be published later as we are currently exploring multiple funding avenues.

Our tentative policy will provide top 3 participating teams with cash prizes:

1st: \$1000

2nd: \$500

3rd: \$250

In case of a tie, the position's cash prize will be divided equally between the tied teams.

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.

Top 3 performing methods will be announced publicly.

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).

We intend to publish challenge paper describing the task, dataset, and results, in a peer-reviewed journal (e.g., Medical Image Analysis). While teams are encouraged to independently publish their methods, a 12-month embargo period will be applied starting from September 2026, lasting until the joint challenge paper is published. The 12-month embargo applies exclusively to publications that would disclose challenge-level evaluation results, including leaderboard rankings or direct comparisons on the private test set.

Participants retain full ownership of their methods. Independent publications that extend the submitted approach or evaluate it on external or non-challenge datasets are permitted after the embargo period.

This policy is designed to protect the integrity of the challenge evaluation while remaining consistent with community norms for participant scholarship.

Submission method

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

Examples:

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

Docker container submission via CodaBench. Containers are executed centrally by organizers against the private test set.

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.

Participating teams are permitted up to five development submissions. Each team must designate one final submission, and only this final submission will be used for official ranking and awards.

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

Challenge website and registration opens – April 1, 2026

Training dataset release – May 1, 2026

Team registration closes – September 1, 2026

Submission deadline – September 15, 2026

Challenge day – October 3 or 8, 2026

Ethics approval

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

An IRB exemption was granted to conduct this retrospective study by the University of Maryland Institutional Review Board (HP-00117149). The training dataset consists of only publicly available datasets. The private held-out institutional test set consists of studies acquired within the University of Maryland Medical System. All studies were fully anonymized and processed in accordance with institutional policy.

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)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

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 code for model evaluation will be available on our GitHub. Additionally, we will provide two trained baseline models and the code for building the Docker submission container.

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

We encourage all participating teams to publish their code.

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.

No conflict of interests. The private institutional test set will only be accessible to the challenge organizers, and we do not intend to publicly release this dataset.

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

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 this challenge is patients with suspected HCC lesions. Patients may present with multiple liver lesions clinically; however, each challenge case is defined by a single target lesion per scan selected and annotated for LI-RADS characterization. The indented models should be able to characterize these lesions with clinical LI-RADS categories.

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

The public dataset comprises 668 cases aggregated from four publicly available sources (TCGA-LIHC, WAW-TACE, HCC-TACE-SEG, and PLC-CECT). This includes 83 cases with normal livers and no liver lesions, and 585 cases containing hepatocellular carcinoma (HCC) lesions. Approximately 10% of the public data will be reserved as a labeled validation set, with the remaining cases used for training.

In addition, we curated a held-out private institutional test set at the University of Maryland consisting of 257 cases spanning LI-RADS categories LR-1 through LR-5. This cohort includes 197 male patients (77%; mean age 65 ± 10 years) and 60 female patients (23%; mean age 69 ± 10 years). In accordance with institutional and regulatory constraints, this test set will remain fully private and will not be released or made directly accessible during or after the challenge.

Imaging modality(ies)

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

Multi-phase abdominal CT acquired before and after contrast injection in the arterial, portal venous, and delayed phases.

Context information

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

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

The public training dataset will include lesion-level LI-RADS categories (LR-1 to LR-5), size (in mm), and voxel-level segmentation masks for non-rim APHE, non-peripheral washout, and enhancing capsule.

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

We will provide metadata for each volume, including contrast phase and acquisition protocol. Patient demographics will be provided if they are available.

Target entity(ies)

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

The data for this challenge originates from multi-phase contrast-enhanced CT imaging of the liver, focusing on the liver parenchyma and surrounding structures. The imaging encompasses the entire liver volume across arterial, portal venous, and delayed phases, providing comprehensive spatio-temporal enhancement patterns critical for assessing lesion characterization.

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 to accurately characterize HCC lesions within multi-phase contrast-enhanced CT images of the liver. Reliable lesion characterization is essential for evaluating disease staging, treatment planning, and therapeutic response monitoring, providing the clinical insights necessary for informed decision-making.

Participants' algorithms are designed to tackle the challenges of temporal analysis across multiple contrast phases, heterogeneous lesion presentations, and variable image quality, ensuring precise lesion characterization with clinical LI-RADS categories.

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, Agreement

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).

Challenge data was primarily acquired on CT scanners from GE Healthcare, Philips, and Siemens with some cases from Canon and Toshiba scanners.

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

All images were acquired as multi-phase abdominal CT consisting of before (non-contrast) and after intravenous contrast was injected (arterial, portal venous, and delayed phases).

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.

Training Data: The public training dataset was curated from four publicly accessible datasets and re-annotated by three board-certified abdominal radiologists. The contributing centers and institutions for each dataset are listed below.

WAW-TACE (Medical University of Warsaw)

HCC-TACE-SEG (The University of Texas MD Anderson Cancer Center)

TCGA-LIHC (Mayo Clinic, Rochester; University of North Carolina, Chapel Hill; Alberta Health Services; Lahey Hospital & Medical Center)

PLC-CECT (Chongqing Yubei District People's Hospital)

Testing Data: The held-out private institutional dataset was curated within the University of Maryland Medical System. It was annotated following similar protocol as the training dataset. However, we provide no additional information to ensure the integrity of the challenge.

The held-out test set is drawn from three different imaging centers within the University of Maryland Medical System, rather than a single site.

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

The held-out private institutional test set was acquired by specialized team of radiologists, physicians, and technologists. Manual annotation of LI-RADS categories and voxel-level segmentation mask was performed by three board-certified abdominal 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 and test datasets all consist of multi-phase contrast-enhanced CT scans. All cases in these datasets will be annotated by three board-certified abdominal radiologists who provided lesion-level LI-RADS category assignments and voxel-level annotations for non-rim APHE, non-peripheral washout, and enhancing capsule for HCC lesions.

A case refers to a single patient containing all four contrast phases (non-contrast, arterial, portal venous, and delayed). All CT volumes and segmentations were processed to NIfTI format for accessibility.

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

Training: 601

Validation: 67

Note: The validation set comprises approximately 10% of the public data and is constructed to ensure equitable representation across all contributing datasets.

Testing: 257 (held-out, private)

c) How much of the data are already annotated (stratified by train test in percentage)?

Training and validation cases are partially annotated if the source dataset provided them. However, all cases (training and testing) will be re-annotated by three board-certified radiologists for the purpose of the challenge.

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

We decided to include all publicly available cases within the training and validation datasets, while keeping all test cases private to ensure the integrity of the challenge.

The total number of training cases was determined by the number of cases containing HCC suspicious lesions within the four harmonized public datasets: TCGA-LIHC, WAW-TACE, HCC-TACE-Seg, and PLC-CECT.

The total number of testing cases was determined by the availability of multi-phase abdominal CT exams indicated with LR-1 through LR-5 within the University of Maryland Medical System.

e) 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.

All training and test cases were drawn from real-world distributions of patients with suspected HCC lesions.

f) 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.

All models will be tested on a held-out private institutional dataset curated at the University of Maryland Medical System and following all institutional protocols. It consists of 257 cases acquired between February 2023 and December 2025. We do not intend to release the test set to the public nor provide any additional details to ensure integrity of the challenge.

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.

3

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

All studies were randomly shuffled and pre-loaded into a custom 3D Slicer module prior to annotation. Board-certified abdominal subspecialty radiologists used this module to manually segment liver lesions and record the required annotations using standard Slicer tools integrated into the custom interface. A brief onboarding demonstrated the workflow and annotation fields, after which the same protocol was applied uniformly across training, validation, and test cases. Radiologists were blinded to dataset splits and model outputs, and no separate annotation procedures were used for different splits.

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

All cases were annotated by board-certified radiologists with substantial clinical experience in abdominal imaging. The annotators included: Dr. Barton F. Lane, in practice since 2002 (24 years of experience); Dr. Jade Wong-You-Cheong, in practice since 1987 (39 years of experience); and Dr. Florence X. Doo, in practice since 2018

(8 years of experience). All annotations were performed manually by these medically trained experts using the custom 3D Slicer module. The same group of radiologists and annotation processes were used across the training, validation, and test cases.

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.

Multiple annotations were not algorithmically merged. After all lesion labeling was completed, a cross-validation review process was performed among the three radiologists: annotations produced by Radiologist 1 were validated by Radiologists 2 and 3, annotations by Radiologist 2 were validated by Radiologists 1 and 3, and annotations by Radiologist 3 were validated by Radiologists 1 and 2. During this process, segmentations and labels were reviewed and corrected as needed, and the consensus-validated annotation was treated as the final reference standard. The same review and validation procedure was applied consistently across the training, validation, and test cases.

Data pre-processing method(s)

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

The following pre-processing steps were applied to standardize the raw training data across all source datasets:

1. Removed irrelevant DICOM and DICOM-SEG series

TCGA-LIHC contained irrelevant series for our dataset (e.g., Scout, Dose Reports, etc.) which were removed. HCC-TACE-SEG contained combined lesion segmentations that were not compatible with our per-lesion annotation schema; these were removed.

2. Converted DICOM datasets to NIfTI volumes

TCGA-LIHC and HCC-TACE-SEG were provided in DICOM format and were converted to NIfTI using dcm2niix for standardization across the dataset.

3. Converted lesion segmentations from NRRD to NIfTI

WAW-TACE provided lesion segmentations in NRRD format; these were converted to NIfTI to maintain format consistency across all segmentation masks.

4. Phase Detection

We detected the contrast phase (non-contrast, arterial, portal venous, delayed) of each volume using imaging metadata and temporal acquisition patterns using Comp2Comp [14,15]. Phase labels were stored in the metadata of the dataset to enable phase-specific analysis and model development.

5. Image Registration

All multi-phase CT studies underwent image registration to align arterial, portal venous, and delayed phase acquisitions to a common spatial reference. The selection of the key (reference) image for registration was determined hierarchically:

- Primary criterion: The phase containing the highest count of annotated lesions was selected as the key image to maximize lesion visibility during alignment.
- Default criterion: If lesion count information was not available, the arterial phase image was selected as the default key image, as this phase typically provides optimal lesion conspicuity for HCC characterization

The preprocessing pipeline will be implemented centrally and applied uniformly to the training, validation, and test sets. The complete preprocessing code will be made publicly available to ensure transparency and

reproducibility.

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.

One possible error source for image annotation is uncertainty and inter-annotator variability. Prior literature suggests radiologist segmentation performance can vary 10-30% [16]

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

Another potential error source is image artifacts in CT that may result in suboptimal diagnostic quality.

ASSESSMENT METHODS

Metric(s)

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

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

We will use Quadratic Weighted Kappa (QWK) to evaluate ordinal classification performance. QWK measures agreement between predicted and ground truth LI-RADS categories while penalizing predictions based on their distance from the true category. This is clinically appropriate because predicting LI-RADS 4 for a true LI-RADS 5 lesion is less severe than predicting LI-RADS 2.

LI-RADS categories for classification: LR-1 ("definitely benign"), LR-2 ("probably benign"), LR-3 ("intermediate probability"), LR-4 ("probably HCC"), and LR-5 ("definitely HCC"). Lesions classified as LR-M are excluded from this task to maintain the ordinal nature of the classification.

QWK retained is the primary ranking metric; secondary metrics reported for analysis only.

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

QWK was chosen because it is specifically designed for ordinal classification tasks where the magnitude of disagreement matters. Unlike standard accuracy, QWK appropriately penalizes predictions based on their clinical significance. This metric was successfully employed in DRAC 2022 (Diabetic Retinopathy Analysis Challenge) for grading image quality, demonstrating its suitability for ordinal medical imaging classification tasks.

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. Ideally, provide the ranking scheme as a concrete pseudo code.

Only one submission per participating team, designated as the final submission, will be considered for ranking. For each test case, a single predicted LI-RADS category is generated for the pre-defined target lesion. The QWK

score is computed across all test cases by comparing predictions against ground-truth labels. Final score is determined by normalized QWK on scale 0-1.

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

All submitted algorithms are required to produce a LI-RADS prediction for the single pre-defined target lesion in every test case. If an algorithm fails to generate a valid prediction for a test case, that case is excluded from the QWK calculation for that algorithm, and a completion rate (percentage of test cases with valid predictions) is reported separately. Methods below 95% completeness may be deemed non-compliant and excluded from awards.

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

The weighted aggregation scheme was chosen to reflect the primary clinical objective of this challenge: accurate LI-RADS category assignment for HCC characterization. In clinical practice, the LI-RADS category directly determines patient management – whether a lesion requires continued surveillance, further diagnostic workup, or treatment. Accurate categorization is therefore the most clinically impactful outcome of an AI system in this domain.

Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

The statistical evaluation framework for this challenge employs bootstrap resampling methods to assess algorithm performance, precision, and ranking stability for the ordinal classification task. Quadratic Weighted Kappa (QWK) is used as the primary evaluation metric because it appropriately measures agreement in ordinal classifications while penalizing predictions proportionally to their distance from the true LI-RADS category – reflecting the clinical severity of misclassifications.

Bootstrap resampling is performed at the patient level (resampling patients with replacement, including the single annotated target lesion per patient) to account for potential intra-patient correlation between multiple lesions per patient. This hierarchical resampling approach preserves the dependency structure in the data while providing robust confidence intervals and significance testing. The bootstrap method is chosen because it makes minimal distributional assumptions and is appropriate for the ordinal nature of LI-RADS categories.

Pairwise significance testing between algorithms uses paired bootstrap comparisons with Bonferroni correction for multiple comparisons to control the family-wise error rate. Ranking variability is assessed through bootstrap-based rank distributions to determine whether performance differences are robust or due to sampling variability.

All analyses assume that test cases are representative of real-world clinical scenarios and that lesions within the test set adequately represent the spectrum of LI-RADS categories (LR-1 through LR-5). The appropriateness of bootstrap confidence intervals is validated by ensuring sufficient bootstrap iterations (N=10,000) for stable empirical distributions.

Prior to challenge execution, we will perform preliminary distributional analyses comparing the public-source data and the private held-out test data, including scanner/vendor characteristics, acquisition parameters, voxel spacing, intensity statistics, and lesion characteristics, to contextualize generalization performance.

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

The precision of the performance estimates for each algorithm is assessed by computing 95% confidence intervals for the Quadratic Weighted Kappa (QWK) score on the held-out test set using percentile bootstrap resampling with 10,000 iterations. Bootstrap samples are generated by resampling patients with replacement, and all lesions from the resampled patients are included in each bootstrap iteration. This patient-level resampling approach accounts for potential intra-patient correlation between lesions. For each bootstrap iteration, QWK is recomputed on the resampled lesions, and the empirical distribution of QWK values is used to derive confidence intervals from the 2.5th and 97.5th percentiles. Bootstrap resampling is justified because it provides robust confidence intervals without requiring distributional assumptions about the QWK statistic, which is appropriate for ordinal categorical data.

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

Variability of algorithm performance is assessed using lesion-level ordinal error measures derived from LI-RADS predictions. For each case-defining lesion, the absolute difference between the predicted and ground-truth LI-RADS category is computed to quantify ordinal misclassification severity. The distribution of these errors across all test lesions is summarized using descriptive statistics including median, interquartile range (IQR), and standard deviation to characterize the spread of prediction errors. Performance variability is further analyzed using confusion matrices stratified by ground-truth LI-RADS category to identify systematic patterns of misclassification, and by visualizing error distributions to identify outlier lesions with large ordinal disagreement. This approach is justified because ordinal error distances directly reflect the clinical significance of misclassifications in the LI-RADS framework, and non-parametric summary statistics (median, IQR) are appropriate for ordinal data that may not follow normal distributions.

Provide a description of how variability of rankings is assessed.

Variability of algorithm rankings is assessed using bootstrap resampling of the test set with 10,000 iterations. For each bootstrap iteration, patients are resampled with replacement (with all their lesions included), and the Quadratic Weighted Kappa (QWK) score is recomputed for all submitted algorithms on the resampled lesions. Algorithms are re-ranked based on their QWK scores in each iteration. The distribution of ranks across bootstrap iterations is used to quantify ranking stability, including the frequency with which each algorithm attains each possible rank. This approach assesses whether observed differences in rankings are robust or driven by sampling variability. Bootstrap-based ranking variability is justified because it directly quantifies the uncertainty in relative algorithm performance while maintaining the hierarchical structure of the data (patients containing multiple lesions).

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

Statistical significance of performance differences between algorithms is assessed using paired bootstrap resampling on the test set with 10,000 iterations. For each bootstrap iteration, patients are resampled with replacement (with all their lesions included), and the Quadratic Weighted Kappa (QWK) score is recomputed for all algorithms on the same resampled cohort of lesions. Pairwise differences in QWK between algorithms are calculated across bootstrap iterations, and the 95% bootstrap confidence interval of the QWK difference is computed. A difference in performance is considered statistically significant if the confidence interval does not include zero. For multiple pairwise comparisons between algorithms, p-values are adjusted using the Bonferroni correction to control the family-wise error rate and reduce the risk of false positive findings. Paired bootstrap testing is justified because it accounts for case-to-case variability while maintaining the paired structure of algorithm comparisons on identical test data, and the bootstrap approach avoids parametric assumptions about the distribution of QWK differences.

Provide a description of the missing data handling.

All test cases included in the evaluation contain valid ground-truth LI-RADS labels (LR 1 to LR 5) for all lesions. All submitted algorithms are required to produce a LI-RADS prediction for every eligible lesion in the test set. If an algorithm fails to generate a valid prediction for any lesion, that lesion is excluded from the QWK calculation for that algorithm, and a completion rate (percentage of lesions with valid predictions) is reported separately for each algorithm. Algorithms with completion rates below 95% will be flagged in the results to highlight potential robustness issues. No statistical imputation of missing ground-truth labels is performed. This approach is justified because it separates the evaluation of prediction accuracy from algorithm robustness/failure rates, ensuring that QWK scores reflect true classification performance rather than being confounded by arbitrary missing data penalties.

Indicate any software product that is used for all data analysis methods.

All statistical analyses, including computation of Quadratic Weighted Kappa, bootstrap confidence intervals, ranking variability, and significance testing, are performed using Python (version 3.10 or higher) with libraries including scikit-learn (for QWK computation), scipy (for statistical tests), numpy (for numerical operations), pandas (for data handling), and matplotlib/seaborn (for visualization). Custom scripts are implemented to perform patient-level bootstrap resampling with lesion-level QWK computation, ordinal error analyses, and ranking stability assessments. These tools provide comprehensive, validated support for statistical computations required for the challenge evaluation.

Further analyses

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

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

N/A

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.

- [1] Singal AG, Pillai A, Tiro J. Early Detection, Curative Treatment, and Survival Rates for Hepatocellular Carcinoma Surveillance in Patients with Cirrhosis: A Meta-analysis. *PLOS Medicine*. Public Library of Science; 2014;11(4):e1001624. doi: 10.1371/journal.pmed.1001624.
- [2] Llovet JM, Kelley RK, Villanueva A, et al. Hepatocellular carcinoma. *Nat Rev Dis Primers*. 2021;7(1):6. doi: 10.1038/s41572-020-00240-3.
- [3] Chernyak V, Fowler KJ, Kamaya A, et al. Liver Imaging Reporting and Data System (LI-RADS) Version 2018: Imaging of Hepatocellular Carcinoma in At-Risk Patients. *Radiology*. 2018;289(3):816–830. doi: 10.1148/radiol.2018181494.
- [4] Lee S, Kim Y-Y, Shin J, et al. Percentages of Hepatocellular Carcinoma in LI-RADS Categories with CT and MRI: A Systematic Review and Meta-Analysis. *Radiology*. Radiological Society of North America; 2023;307(1):e220646. doi: 10.1148/radiol.220646.
- [5] Goins SM, Jiang H, van der Pol CB, et al. Individual Participant Data Meta-Analysis of LR-5 in LI-RADS Version 2018 versus Revised LI-RADS for Hepatocellular Carcinoma Diagnosis. *Radiology*. Radiological Society of North America; 2023;309(3):e231656. doi: 10.1148/radiol.231656.
- [6] Adamo RG, van der Pol CB, Alabousi M, et al. Diagnostic Performance of CT/MRI LI-RADS Version 2018 Major Feature Combinations: Individual Participant Data Meta-Analysis. *Radiology*. 2025;315(3):e243450. doi: 10.1148/radiol.243450.
- [7] Goins SM, Adamo RG, Lam E, et al. Conversion Strategy for LI-RADS Category 5 Observations across Versions 2014, 2017, and 2018. *Radiology*. Radiological Society of North America; 2023;307(4):e222971. doi: 10.1148/radiol.222971.
- [8] Ying H, Liu X, Zhang M, et al. A multicenter clinical AI system study for detection and diagnosis of focal liver lesions. *Nat Commun*. Nature Publishing Group; 2024;15(1):1131. doi: 10.1038/s41467-024-45325-9.
- [9] Bartnik K, Bartczak T, Krzyżiński M, et al. WAW-TACE: A Hepatocellular Carcinoma Multiphase CT Dataset with Segmentations, Radiomics Features, and Clinical Data. *Radiol Artif Intell*. 2024;6(6):e240296. doi: 10.1148/ryai.240296.
- [10] Erickson B, Kirk S, Lee Y, et al. The Cancer Genome Atlas Liver Hepatocellular Carcinoma Collection (TCGA-LIHC) (Version 5). 2016. doi: <https://doi.org/10.7937/K9/TCIA.2016.IMMQW8UQ>.
- [11] Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat Genet*. 2013;45(10):1113–1120. doi: 10.1038/ng.2764.
- [12] Moawad AW, Morshid A, Khalaf AM, et al. Multimodality annotated hepatocellular carcinoma data set including pre- and post-TACE with imaging segmentation. *Sci Data*. Nature Publishing Group; 2023;10(1):33. doi: 10.1038/s41597-023-01928-3.
- [13] Luo J, Wan X, Du J, et al. Comprehensive multi-phase 3D contrast-enhanced CT imaging for primary liver cancer. *Sci Data*. Nature Publishing Group; 2025;12(1):768. doi: 10.1038/s41597-025-05125-2.
- [14] Wasserthal J, Breit HC, Meyer MT, Pradella M, Hinck D, Sauter AW, Heye T, Boll DT, Cyriac J, Yang S, Bach M. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*. 2023;5(5):e230024. doi: 10.1148/ryai.230024
- [15] Blankemeier L, Desai A, Chaves JM, Wentland A, Yao S, Reis E, Jensen M, Bahl B, Arora K, Patel BN, Lenchik L. Comp2Comp: Open-source body composition assessment on computed tomography. *arXiv preprint arXiv:2302.06568*. 2023.
- [16] van der Loo I, Bucho TM, Hanley JA, Beets-Tan RG, Imholz AL, Trebeschi S. Measurement variability of radiologists when measuring brain tumors. *European Journal of Radiology*. 2025;183:111874. doi:

10.1016/j.ejrad.2024.111874

[17] Bilic P, Christ P, Li HB, Vorontsov E, Ben-Cohen A, Kaissis G, Szeskin A, Jacobs C, Mamani GE, Chartrand G, Lohöfer F. The liver tumor segmentation benchmark (lits). Medical image analysis. 2023 Feb 1;84:102680. doi: 10.1016/j.media.2022.102680

[18] Félix Quinton, Sarah Leclerc, Benoît Presles, Fabrice Meriaudeau, Arnaud Boucher, Dominique Ginhac, et al. A tumor and liver automatic segmentation challenge. Zenodo; 2023. doi: 10.5281/zenodo.7835370

[19] Küstner T, Gatidis S, Megne O, Ingris M, Fabritius M, Dextl J, et al. Automated Lesion Segmentation in Whole-Body PET/CT and Longitudinal (autoPET/CT IV). Zenodo; 2025. doi: 10.5281/zenodo.15045096

[20] Peeters D, Obreja B, Antonissen N, Jacobs C. Benchmarking of Artificial Intelligence and Radiologists for Lung Cancer Screening in CT: The LUNA25 Challenge. Zenodo; 2025. doi: 10.5281/zenodo.15094631

Further comments

Further comments from the organizers.

N/A