CXR-LT 2024: Long-tailed, multi-label, and zero-shot classification on chest X-rays: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

CXR-LT 2024: Long-tailed, multi-label, and zero-shot classification on chest X-rays

Challenge acronym

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

CXR-LT 2024

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.

Chest radiography, like many diagnostic medical exams, produces a long-tailed distribution of clinical findings; while a small subset of diseases is routinely observed, the vast majority of diseases are relatively rare [1]. This poses a challenge for standard deep learning methods, which exhibit bias toward the most common classes at the expense of the important but rare "tail" classes [2]. Many existing methods [3] have been proposed to tackle this specific type of imbalance, though only recently has attention been given to long-tailed medical image recognition problems [4-6]. Diagnosis on chest X-rays (CXRs) is also a multi-label problem, as patients often present with multiple disease findings simultaneously; however, only a select few studies incorporate knowledge of label co-occurrence into the learning process [7-9, 12]. Since most large-scale image classification benchmarks contain single-label images with a mostly balanced distribution of labels, many standard deep learning methods fail to accommodate the class imbalance and co-occurrence problems posed by the long-tailed, multi-label nature of tasks like disease diagnosis on CXRs [2].

In the first iteration of CXR-LT held in 2023, we expanded upon the MIMIC-CXR [10,11] dataset by enlarging the set of target classes from 14 to 26, generating labels for 12 new rare disease findings by parsing radiology reports [13]. While this made for a challenging long-tailed, multi-label disease classification task that attracted 59 teams who contributed over 500 unique submissions, Radiology Gamuts Ontology documents over 4,500 unique radiological image findings. That is, the "true" distribution of all clinical findings on CXR is at least two orders of magnitude longer than what our -- or any existing -- dataset can offer. For this reason, we argue that the only way to truly tackle the long-tail of radiological image findings is to develop a model that can readily generalize to new classes in "zero-shot" fashion [14].

For this year's version of CXR-LT, we extract labels for an additional 19 rare disease findings (for a total of 377,110

CXR images, each with 45 disease labels) and introduce two new challenge tracks, featuring a zero-shot classification task. Our tasks include (i) long-tailed classification on a large, noisy test set, (ii) long-tailed classification in a small, manually annotated test set, and (iii) zero-shot generalization to previously unseen disease findings. For all tracks, participants will be provided with a large, automatically labeled training set of >250,000 CXR images with 40 binary disease labels. Task (i) will be evaluated on a large, automatically labeled test set of >75,000 CXRs from these same 40 labels; task (ii) will be evaluated on a "gold standard" subset of the test set, containing 409 CXRs from 26 of the 40 labels that were manually annotated as described in [14]; task (iii) will be evaluated on the same large test set of images as task (i), but for 5 "held-out" disease findings that have not been encountered during training. While last year's CXR-LT was a success, we hope that CXR-LT 2024 can provide even further meaningful methodological advances toward clinically realistic multi-label, long-tailed, and zero-shot disease classification on CXR.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Long-tailed learning, Zero-shot classification, Chest X-ray, Computer-aided diagnosis

Year

The challenge will take place in 2024

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.

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 estimate that 75-100 teams will register for this challenge. In 2023, a previous version of this challenge was held in conjunction with an ICCV 2023 workshop (https://cvamd2023.github.io), attracting over 200 applicants and 59 registered teams that met dataset access requirements. To enhance participation, we will expand the dataset by including more diseases to form a challenging long-tailed distribution and introduce two additional tasks to the challenge.

Publication and future plans

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

As with the first CXR-LT challenge task held in 2023 (submission under review), we plan to coordinate a publication summarizing the challenge results, inviting top-performing teams to be coauthors.

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 challenge will be hosted online via CodaLab (https://codalab.lisn.upsaclay.fr/) prior to MICCAI 2024. Standard equipment such as a projector, monitor, microphone, and speakers will be needed for invited challenge participants to present their work.

TASK 1: Long-tailed classification in a large, noisy test set

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.

Chest radiography, like many diagnostic medical exams, produces a long-tailed distribution of clinical findings; while a small subset of diseases is routinely observed, the vast majority of diseases are relatively rare [1]. This poses a challenge for standard deep learning methods, which exhibit bias toward the most common classes at the expense of the important but rare "tail" classes [2]. Many existing methods [3] have been proposed to tackle this specific type of imbalance, though only recently has attention been given to long-tailed medical image recognition problems [4-6]. Diagnosis on chest X-rays (CXRs) is also a multi-label problem, as patients often present with multiple disease findings simultaneously; however, only a select few studies incorporate knowledge of label co-occurrence into the learning process [7-9, 12]. Since most large-scale image classification benchmarks contain single-label images with a mostly balanced distribution of labels, many standard deep learning methods fail to accommodate the class imbalance and co-occurrence problems posed by the long-tailed, multi-label nature of tasks like disease diagnosis on CXRs [2]. This task will evaluate a model's ability to perform "in-distribution" long-tailed, multi-label disease classification on CXRs when evaluated on a large test set with noisy, automatically text-mined labels that have been encountered during training.

Keywords

List the primary keywords that characterize the task.

Long-tailed learning, Chest radiography, Computer-aided diagnosis

ORGANIZATION

Organizers

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

Organizing Committee:
Yifan Peng (chair), PhD, Weill Cornell Medicine
Mingquan Lin, PhD, Weill Cornell Medicine
Gregory Holste, The University of Texas at Austin
Song Wang, The University of Texas at Austin
Yiliang Zhou, Weill Cornell Medicine
Hao Chen, Hong Kong University of Science and Technology
Atlas Wang, PhD, The University of Texas at Austin
Steering Committee:
Adam E. Flanders, MD, Thomas Jefferson University
Leo Anthony Celi, MD, MPH, MSc, MIT/Harvard
Zhiyong Lu, PhD, FACMI, NIH

George Shih, MD, Weill Cornell Medicine
Ronald M. Summers, MD, PhD, NIH

b) Provide information on the primary contact person.

Yifan Peng, Assistant Professor of Population Health Sciences, Weill Cornell Medicine, yip4002@med.cornell.edu

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

Challenge venue and platform

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

MICCAI 2024

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

CodaLab (https://codalab.lisn.upsaclay.fr/)

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

TBD. Last year's CXR-LT challenge URL can be found at https://codalab.lisn.upsaclay.fr/competitions/12599.

Participation policies

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

None. Solutions must be 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.

As in the first iteration of CXR-LT, we will allow training on any publicly available data and/or the use of models pretrained on publicly available data. Participants must be careful not to use models that were pretrained on MIMIC-CXR images that are part of the validation or test sets of this challenge; for example, any publicly available MIMIC-CXR-pretrained model is nearly certain to have used different random data splits than we will, meaning this model will have been trained on held-out validation and/or test set cases in this challenge (prohibited).

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.

Members of the organizers' institutes may participate so long as they have no personal relationship or conflict of interest with any of the organizers.

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

1st place: $500, 2nd place: $300, 3rd place: $200. Special recognition may be given to the best-performing team across all three tracks (e.g., of teams who participate in all three tracks, award a bonus prize to the team with the smallest average rank across tracks).

e) Define the policy for result announcement.

Examples:

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

The final ranking of teams will be kept private until MICCAI 2024. However, individual top-performing teams selected to present at MICCAI 2024 will be informed of their results and ranking. Furthermore, the solutions of top-performing teams will be made publicly available in the form of a publication summarizing the challenge.

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).

Up to two members of each of the top-performing teams (TBD precisely how many teams will be invited/how many from each track/etc.) will be invited to coauthor a paper summarizing the CXR-LT challenge. Participating teams may post preprints of their challenge solutions (e.g., on arXiv), which we will cite in our challenge overview paper as appropriate, after the competition. If a team wishes to publish their own paper, they must either waive their right to coauthor the challenge summary paper or wait until after the summary paper is published to submit their own paper.

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.

Teams will submit comma separated value (CSV) files containing their model's predictions on held-out test data via the CodaLab platform. Similar to last year's challenge (https://codalab.lisn.upsaclay.fr/competitions/12599 -> "Learn the Details" -> "Submission Format"), instructions will be given to ensure that the submitted file is in the proper format (e.g., columns correspond to the test set disease classes and there is a row/prediction for every test set image).

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.

As with the prior version of CXR-LT, there will be a months-long "development phase", where participants can submit predictions on a held-out development/validation set (10% of the data). The development phase will feature a live public leaderboard, displaying the performance metrics and ranking of each team's best-performing solution (according to the primary evaluation metric).

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

05/01/2024: Training data released and challenge (development phase) begins
08/01/2024: Test labels released and final evaluation (test phase) begins 08/04/2024: Test phase ends and competition is closed
08/15/2024: Top-performing teams invited to present at MICCAI 2024
10/06/2024: MICCAI 2024 CXR-LT challenge 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).

This challenge uses image data from MIMIC-CXR-JPG v2.0.0 (https://physionet.org/content/mimic-cxr-jpg/2.0.0/), a publicly available dataset for which ethics approval has already been acquired. For full details, see Johnson et al. [15].

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)

Participants are required to sign and abide by the data use agreement for MIMIC-CXR-JPG v2.0.0: https://physionet.org/content/mimic-cxr-jpg/view-dua/2.0.0/. As in last year's challenge, participants are only granted permission to participate if they provide acceptable proof (via Google Form) that they have properly met the registration requirements for MIMIC-CXR-JPG access. Instructions for applying to the prior version of CXR-LT can be found here: https://codalab.lisn.upsaclay.fr/competitions/12599 ("Learn the Details" -> "Terms and Conditions").

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.

Evaluation code will be made available with the release of the training data.

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

All participating teams are encouraged to provide reproducible, open-source training and inference code. On CodaLab, participants will be instructed to submit all relevant code and submission will fail if a "code" directory is not present in their upload.

Conflicts of interest

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

The CXR-LT challenge received funding from the Artificial Intelligence Journal (AIJ). Only the challenge organizers will have access to all test set labels.

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, research

Task category(ies)

State the task category(ies)

Examples:

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

Classification, prediction

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 consists of patients who have undergone chest X-ray imaging.

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

The challenge cohort uses data from MIMIC-CXR, which consists of patients who underwent chest X-ray imaging at the Beth Israel Deaconess Medical Center emergency department between 2011-2016. Full details can be found in Johnson et al. [15].

Imaging modality(ies)

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

X-ray imaging (projectional radiography)

Context information

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

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

For the training set only, disease labels denoting the presence or absence of 40 clinical findings will be provided in CSV format.

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

No patient information will be explicitly provided to participants for this challenge. However, MIMIC-CXR includes basic information such as sex and age, which participants are free to use.

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.

Chest shown in X-ray image

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.

Cardiopulmonary disease present on chest X-ray

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)).

The goal of this task is to accurately identify thoracic diseases present in chest X-rays, where clinical findings follow a long-tailed distribution and can co-occur with one another. The relevant properties to optimize are accuracy, precision, sensitivity, and specificity. This task also involves robustness to noisy labels extracted by automatic text mining of radiology reports.

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).

This information can potentially be found in the DICOM metadata associated with each image in MIMIC-CXR (https://physionet.org/content/mimic-cxr/2.0.0/).

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

The MIMIC-CXR dataset was acquired during routine clinical practice in an emergency department, involving chest X-rays acquired for a variety of reasons, on several different scanners, and potentially from several different views of the patient. Details can be found in Johnson et al. [15].

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.

Beth Israel Deaconess Medical Center Emergency Department

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).

Regarding image data acquisition, this information is not available. The data was collected over years of routine clinical practice in an emergency department by many professionals of many different experience levels.

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.

For training and testing, a case is a chest X-ray image (and, optionally, patient information such as sex and age, when available) with labels describing the presence and absence of 40 clinical findings. These 40 findings are: Adenopathy, Atelectasis, Azygos Lobe, Calcification of the Aorta, Cardiomegaly, Clavicle Fracture, Consolidation, Edema, Emphysema, Enlarged Cardiomediastinum, Fibrosis, Fissure, Fracture, Granuloma, Hernia, Hydropneumothorax, Infarction, Infiltration, Kyphosis, Lobar Atelectasis, Lung Lesion, Lung Opacity, Mass, Normal, Nodule, Pleural Effusion, Pleural Other, Pleural Thickening, Pneumomediastinum, Pneumonia, Pneumoperitoneum, Pneumothorax, Pulmonary Embolism, Pulmonary Hypertension, Rib Fracture, Round Atelectasis, Subcutaneous Emphysema, Support Devices, Tortuous Aorta, Tuberculosis

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

Training: 263,977; Validation/development: 37,711; Test: 75,422

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

The data was randomly partitioned into training (70%), validation/development (10%), and test (20%) sets at the patient level to avoid data leakage. These proportions were chosen such that as much labeled data as possible could be allotted to the training set while still leaving a sufficient number of positive examples for extremely rare classes to facilitate reliable evaluation (at minimum, 30). Critically, the splits are different from those of last year's challenge.

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

The distribution of labels is extremely long-tailed with an "imbalance ratio" (ratio describing the prevalence of the most- to least-common class) over 685, exceeding that of challenge long-tailed benchmark datasets such as ImageNet-LT [16] and iNaturalist2018 [17]. The dataset is also distinct from these "single-label" benchmarks in that labels frequently co-occur in complex patterns that reflect underlying biology.

Also, we replace the "No Finding" class with what we believe to be a more natural "Normal" class. Previously, "No Finding" meant that none of the other clinically relevant findings were present in the study (except for "Support Devices", which is not a clinically meaningful finding). The fact that "No Finding" is defined in relation to other classes included in the set of labels means that "No Finding" will have a different, or ambiguous, meaning when considered in the context of a different set of labels. For example, if participants were given "No Finding" labels defined in relation to the other 39 classes in the training set, what then would "No Finding" mean when evaluated on a new set of 5 unseen diseases? For this reason, we extract new labels for a "Normal" class, which simply means that no cardiopulmonary disease or abnormality was found (again, excluding "Support Devices").

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.

All labels were automatically text-mined from the radiology report associated with each image using MedText (https://github.com/bionlplab/medtext) [18,19]. MedText preprocesses radiology reports through a sequential workflow encompassing de-identification, section segmentation, and sentence splitting. For each label of interest, we manually mined and curated a list of synonyms from various data sources, including RadLex radiology lexicon (https://radlex.org/), Medical Subject Headings (MeSH) (https://www.ncbi.nlm.nih.gov/mesh/), etc. We then utilized rule-based named entity recognition to detect the mentions of our predefined labels from the radiology reports; specifically, spaCy's PhraseMatcher (https://spacy.io/api/phrasematcher) was adopted by MedText as part of this process. Since negative medical findings indicate the absence of findings within the radiology report, identifying the negation status of medical findings is as important as identifying the positive findings. MedText uses NegBio [18] for negation detection, which utilizes universal dependencies for pattern definition and subgraph matching for graph traversal search so that the scope for detecting negation is not limited to the fixed word distance. Eventually, the detected disease findings and their identified assertion/negation status were compiled to produce the labels for the radiology report associated with each image.

The list of newly added clinical findings was chosen from sources including the disease list of the PadChest dataset [20] and Hansell et al. [21], ensuring that a sufficient number of occurrences were observed in the dataset.

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.

N/A

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.

MedText preprocesses radiology reports through a sequential workflow encompassing de-identification, section segmentation, and sentence splitting. For each label of interest, we manually mined and curated a list of synonyms from various data sources, including RadLex radiology lexicon (https://radlex.org/), Medical Subject Headings (MeSH) (https://www.ncbi.nlm.nih.gov/mesh/), etc. We then utilized rule-based named entity recognition to detect the mentions of our predefined labels from the radiology reports; specifically, spaCy's PhraseMatcher (https://spacy.io/api/phrasematcher) was adopted by MedText as part of this process. Since negative medical findings indicate the absence of findings within the radiology report, identifying the negation status of medical findings is as important as identifying the positive findings. MedText uses NegBio [18] for negation detection, which utilizes universal dependencies for pattern definition and subgraph matching for graph traversal search so that the scope for detecting negation is not limited to the fixed word distance.

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

N/A

Data pre-processing method(s)

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

The image data used in this competition comes from MIMIC-CXR-JPG v2.0.0 (https://physionet.org/content/mimic-cxr-jpg/2.0.0/), which contains CXR images in JPEG format. While the original MIMIC-CXR provided data in the standardized DICOM format, MIMIC-CXR-JPG aimed to make the data more accessible (less storage needed) after applying preprocessing steps such as deidentification, histogram equalization, then conversion to JPEG format. Full details can be found in Johnson et al. [22]

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 source of error in automatic labeling is that the curated list of label synonyms may not have fully covered all possible mentions in the radiology reports -- hence, some mentions of certain labels could go undetected. However, to measure the the ability of MedText to accurately detect disease findings, we randomly selected 200 test set reports and manually annotated five disease findings: Calcification of the Aorta, Pneumomediastinum, Pneumoperitoneum, Subcutaneous Emphysema, Tortuous Aorta. Using these manual annotations as a reference, the automated MedText predictions for these five diseases achieved an average precision of 0.91, average recall of 0.94, and average F1 score of 0.92 across the classes. The highest precision (1.0) was achieved for Calcification of the Aorta, the highest recall (1.0) for Pneumomediastinum and Pneumoperitoneum, and the highest F1 score (0.97) for Tortuous Aorta.

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

The annotation methods used in this challenge (automatic or manual) can only, at best, capture the opinion of the radiologist interpreting the CXR imaging, which of course is subject to error. As noted in Holste et al. [14], "Even for highly trained experts, diagnosis from CXR is difficult and complex, leading to high inter-reader variability." [23,24]

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)

Mean average precision (mAP), area under the receiver operating characteristic curve (mAUROC), and F1 score (mF1), and expected calibration error (mECE). Each metric will be computed for each class in the test set and then "macro-averaged" to form overall mAP, mAUROC, mF1, and mECE. The specific implementations of the first three metrics are provided by the scikit-learn library [25]. The primary metric is mAP, and mAUROC will be used to break ties if necessary. F1 is computed by binarizing predicted probabilities with a threshold of 0.5.

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

As explained in the prior version of CXR-LT (https://codalab.lisn.upsaclay.fr/competitions/12599 -> "Learn the Details" -> "Evaluation"), the primary evaluation metric for this challenge will be mAP. While mAUROC is a standard metric for related CXR classification datasets, AUROC can be heavily inflated in the face of strong class imbalance [26,27]. We argue that mAP (summarizing area under the precision-recall curve) is more appropriate for the long-tailed, multi-label setting since it both (i) measures performance across decision thresholds and (ii) does not degrade under imbalance [28]. For thoroughness and accordance with the literature, mAUROC will also be calculated and appear on the leaderboard, only being used in ranking to break ties. For additional evaluation of precision and recall with a fixed decision threshold, mF1 score will also be calculated and appear on the leaderboard. Finally, to provide auxiliary assessment of model calibration (beyond class discrimination), mECE will be computed and appear on the leaderboard.

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.

Ranking will be determined primarily by macro-averaged mAP (higher is better), the average AP value over all target classes present in the test. If ties are present, mAUROC (higher is better) will be used to break ties. *Each task is completely independent with regard to evaluation and ranking.*

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

Submissions with missing results will be discarded (submission will fail). The evaluation code used by CodaLab will ensure that predictions for all test set cases are present in the uploaded CSV file before performance evaluation.

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

Ranking is based primarily on mAP for the reasons outlined above. mAUROC is also included (and used as a tiebreaker) because it is a "standard" metric for related tasks of disease classification on CXR. While we understand the importance of assessing ranking stability via methods like bootstrapping (discussed later), the final ranking of teams will depend on the single scalar metric values obtained on the official test set for this task.

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.

Assessment of variability via bootstrapping will be performed for top-performing solutions as an additional analysis in the challenge overview paper, but final rankings will be determined by the single metric value calculated from the original test set predictions and labels. The evaluation code will use standard third-party Python libraries such as NumPy [29], scikit-learn [25], and pandas [30].

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

Bootstrapping test cases was chosen since it is easy to implement and enables assessment of variability in model performance and ranking stability.

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.

As discussed above, an assessment of variability in predictive performance (for each team/model) and ranking stability (across teams/models) will be performed via bootstrapping. This will be explored in our challenge overview paper, but not used to determine official team rankings.

TASK 2: Long-tailed classification in a small, manually annotated test set

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.

Chest radiography, like many diagnostic medical exams, produces a long-tailed distribution of clinical findings; while a small subset of diseases is routinely observed, the vast majority of diseases are relatively rare [1]. This poses a challenge for standard deep learning methods, which exhibit bias toward the most common classes at the expense of the important but rare "tail" classes [2]. Many existing methods [3] have been proposed to tackle this specific type of imbalance, though only recently has attention been given to long-tailed medical image recognition problems [4-6]. Diagnosis on chest X-rays (CXRs) is also a multi-label problem, as patients often present with multiple disease findings simultaneously; however, only a select few studies incorporate knowledge of label co-occurrence into the learning process [7-9, 12]. Since most large-scale image classification benchmarks contain single-label images with a mostly balanced distribution of labels, many standard deep learning methods fail to accommodate the class imbalance and co-occurrence problems posed by the long-tailed, multi-label nature of tasks like disease diagnosis on CXRs [2]. This task will evaluate a model's ability to perform "in-distribution" long-tailed, multi-label disease classification on CXRs when evaluated on a small test set with more reliable, manually annotated labels that have been encountered during training.

Keywords

List the primary keywords that characterize the task.

Long-tailed learning, Chest radiography, Computer-aided diagnosis

ORGANIZATION

Organizers

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

Organizing Committee:
Yifan Peng (chair), PhD, Weill Cornell Medicine
Mingquan Lin, PhD, Weill Cornell Medicine
Gregory Holste, The University of Texas at Austin
Song Wang, The University of Texas at Austin
Yiliang Zhou, Weill Cornell Medicine
Hao Chen, Hong Kong University of Science and Technology
Atlas Wang, PhD, The University of Texas at Austin
Steering Committee:
Adam E. Flanders, MD, Thomas Jefferson University
Leo Anthony Celi, MD, MPH, MSc, MIT/Harvard
Zhiyong Lu, PhD, FACMI, NIH

George Shih, MD, Weill Cornell Medicine
Ronald M. Summers, MD, PhD, NIH

b) Provide information on the primary contact person.

Yifan Peng, Assistant Professor of Population Health Sciences, Weill Cornell Medicine, yip4002@med.cornell.edu

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

Challenge venue and platform

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

MICCAI 2024

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

CodaLab (https://codalab.lisn.upsaclay.fr/)

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

TBD. Last year's CXR-LT challenge URL can be found at https://codalab.lisn.upsaclay.fr/competitions/12599.

Participation policies

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

None. Solutions must be 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.

As in the first iteration of CXR-LT, we will allow training on any publicly available data and/or the use of models pretrained on publicly available data. Participants must be careful not to use models that were pretrained on MIMIC-CXR images that are part of the validation or test sets of this challenge; for example, any publicly available MIMIC-CXR-pretrained model is nearly certain to have used different random data splits than we will, meaning this model will have been trained on held-out validation and/or test set cases in this challenge (prohibited).

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.

Members of the organizers' institutes may participate so long as they have no personal relationship or conflict of interest with any of the organizers.

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

1st place: $500, 2nd place: $300, 3rd place: $200. Special recognition may be given to the best-performing team across all three tracks (e.g., of teams who participate in all three tracks, award a bonus prize to the team with the smallest average rank across tracks).

e) Define the policy for result announcement.

Examples:

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

The final ranking of teams will be kept private until MICCAI 2024. However, individual top-performing teams selected to present at MICCAI 2024 will be informed of their results and ranking. Furthermore, the solutions of top-performing teams will be made publicly available in the form of a publication summarizing the challenge.

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).

Up to two members of each of the top-performing teams (TBD precisely how many teams will be invited/how many from each track/etc.) will be invited to coauthor a paper summarizing the CXR-LT challenge. Participating teams may post preprints of their challenge solutions (e.g., on arXiv), which we will cite in our challenge overview paper as appropriate, after the competition. If a team wishes to publish their own paper, they must either waive their right to coauthor the challenge summary paper or wait until after the summary paper is published to submit their own paper.

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.

Teams will submit comma separated value (CSV) files containing their model's predictions on held-out test data via the CodaLab platform. Similar to last year's challenge (https://codalab.lisn.upsaclay.fr/competitions/12599 -> "Learn the Details" -> "Submission Format"), instructions will be given to ensure that the submitted file is in the proper format (e.g., columns correspond to the test set disease classes and there is a row/prediction for every test set image).

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.

As with the prior version of CXR-LT, there will be a months-long "development phase", where participants can submit predictions on a held-out development/validation set (10% of the data). The development phase will feature a live public leaderboard, displaying the performance metrics and ranking of each team's best-performing solution (according to the primary evaluation metric).

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

05/01/2024: Training data released and challenge (development phase) begins
08/01/2024: Test labels released and final evaluation (test phase) begins 08/04/2024: Test phase ends and competition is closed
08/15/2024: Top-performing teams invited to present at MICCAI 2024
10/06/2024: MICCAI 2024 CXR-LT challenge 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).

This challenge uses image data from MIMIC-CXR-JPG v2.0.0 (https://physionet.org/content/mimic-cxr-jpg/2.0.0/), a publicly available dataset for which ethics approval has already been acquired. For full details, see Johnson et al. [15]. This challenge also uses data from the Medical Imaging and Data Resource Center (MIDRC) (https://data.midrc.org/), for which ethics approval has already been acquired. MIDRC is an open repository of medical imaging data from many different sources, so each source dataset has undergone its own independent ethics approval process.

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)

Participants are required to sign and abide by the data use agreement for MIMIC-CXR-JPG v2.0.0: https://physionet.org/content/mimic-cxr-jpg/view-dua/2.0.0/. As in last year's challenge, participants are only granted permission to participate if they provide acceptable proof (via Google Form) that they have properly met the registration requirements for MIMIC-CXR-JPG access. Instructions for applying to the prior version of CXR-LT can be found here: https://codalab.lisn.upsaclay.fr/competitions/12599 ("Learn the Details" -> "Terms and Conditions"). Participants also must agree to the MIDRC data use agreement (https://data.midrc.org/dashboard/Public/documentation/DUA.html).

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.

Evaluation code will be made available with the release of the training data.

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

All participating teams are encouraged to provide reproducible, open-source training and inference code. On CodaLab, participants will be instructed to submit all relevant code and submission will fail if a "code" directory is not present in their upload.

Conflicts of interest

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

The CXR-LT challenge received funding from the Artificial Intelligence Journal (AIJ). Only the challenge organizers will have access to all test set labels.

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, research

Task category(ies)

State the task category(ies)

Examples:

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

Classification, prediction

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 consists of patients who have undergone chest X-ray imaging.

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

The challenge cohort uses data from MIMIC-CXR, which consists of patients who underwent chest X-ray imaging at the Beth Israel Deaconess Medical Center emergency department between 2011-2016. Full details can be found in Johnson et al. [15]. This task will also use data from MIDRC, a linked collection of deidentified medical imaging data from hundreds of institutions.

Imaging modality(ies)

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

X-ray imaging (projectional radiography)

Context information

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

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

For the training set only, disease labels denoting the presence or absence of 40 clinical findings will be provided in CSV format.

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

No patient information will be explicitly provided to participants for this challenge. However, MIMIC-CXR includes basic information such as sex and age, which participants are free to use.

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.

Chest shown in X-ray image

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.

Cardiopulmonary disease present on chest X-ray

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)).

The goal of this task is to accurately identify thoracic diseases present in chest X-rays, where clinical findings follow a long-tailed distribution and can co-occur with one another. The relevant properties to optimize are accuracy, precision, sensitivity, and specificity. This task also involves robustness to noisy labels extracted by automatic text mining of radiology reports (during training).

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).

This information can potentially be found in the DICOM metadata associated with each image in MIMIC-CXR (https://physionet.org/content/mimic-cxr/2.0.0/) and MIDRC (https://data.midrc.org/).

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

The MIMIC-CXR dataset was acquired during routine clinical practice in an emergency department, involving chest X-rays acquired for a variety of reasons, on several different scanners, and potentially from several different views of the patient. Details can be found in Johnson et al. [15]. The MIDRC data used in this task comprises chest X-rays acquired during clinical practice; since MIDRC consists of data from hundreds of institutions, it is difficult to summarize the specific data acquisition protocol used.

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.

MIMIC-CXR data comes from the Beth Israel Deaconess Medical Center Emergency Department. Again, MIDRC data has been contributed from hundreds of different institutions (https://data.midrc.org/).

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).

For MIMIC-CXR, information regarding who performed the image acquisition is unavailable. The data was collected over years of routine clinical practice in an emergency department by many professionals of many different experience levels. Similarly, since MIDRC data comprises images collected over decades from hundreds of different institutions, this information cannot be summarized and is assumed to be unavailable in the vast majority of cases.

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.

For training, a case is a chest X-ray image (and, optionally, patient information such as sex and age, when available) with labels describing the presence and absence of 40 clinical findings. These 40 findings are: Adenopathy, Atelectasis, Azygos Lobe, Calcification of the Aorta, Cardiomegaly, Clavicle Fracture, Consolidation, Edema, Emphysema, Enlarged Cardiomediastinum, Fibrosis, Fissure, Fracture, Granuloma, Hernia, Hydropneumothorax, Infarction, Infiltration, Kyphosis, Lobar Atelectasis, Lung Lesion, Lung Opacity, Mass, Normal, Nodule, Pleural Effusion, Pleural Other, Pleural Thickening, Pneumomediastinum, Pneumonia, Pneumoperitoneum, Pneumothorax, Pulmonary Embolism, Pulmonary Hypertension, Rib Fracture, Round Atelectasis, Subcutaneous Emphysema, Support Devices, Tortuous Aorta, Tuberculosis.

For testing, a case is a chest X-ray image (and, optionally, patient information such as sex and age, when available) with labels describing the presence and absence of the 26 clinical findings used in the 2023 version of CXR-LT. These labels were manually annotated from text reports as described earlier and in Holste et al. [14]. These 26 findings are: Atelectasis, Calcification of the Aorta, Cardiomegaly, Consolidation, Edema, Emphysema, Enlarged Cardiomediastinum, Fibrosis, Fracture, Hernia, Infiltration, Lung Lesion, Lung Opacity, Mass, Normal, Nodule, Pleural Effusion, Pleural Other, Pleural Thickening, Pneumomediastinum, Pneumonia, Pneumoperitoneum, Pneumothorax, Subcutaneous Emphysema, Support Devices, Tortuous Aorta.

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

Training: 263,977; Validation/development: 37,711; Test: 609

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

The data was randomly partitioned into training (70%), validation/development (10%), and test (20%) sets at the patient level to avoid data leakage. These proportions were chosen such that as much labeled data as possible could be allotted to the training set while still leaving a sufficient number of positive examples for extremely rare classes to facilitate reliable evaluation (at minimum, 30). Critically, the splits are different from those of last year's challenge. However, we ensured the test set used in this track to be a small subset (N=409), manually annotated as described elsewhere, of the larger test set used in tasks 1 and 3. To address concerns regarding small sample size, we additionally include N=200 images from MIDRC for additional evaluation in this task. Critically, including additional MIDRC data (from a different source institution than MIMIC-CXR) discourages forms of cheating that involve using MIMIC-CXR test set labels or tuning algorithms on MIMIC-CXR test data since such approaches will perform very poorly on external data.

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

The distribution of labels is extremely long-tailed with an "imbalance ratio" (ratio describing the prevalence of the most- to least-common class) over 685, exceeding that of challenge long-tailed benchmark datasets such as ImageNet-LT [16] and iNaturalist2018 [17]. The dataset is also distinct from these "single-label" benchmarks in that labels frequently co-occur in complex patterns that reflect underlying biology.

Also, we replace the "No Finding" class with what we believe to be a more natural "Normal" class. Previously, "No Finding" meant that none of the other clinically relevant findings were present in the study (except for "Support Devices", which is not a clinically meaningful finding). The fact that "No Finding" is defined in relation to other classes included in the set of labels means that "No Finding" will have a different, or ambiguous, meaning when considered in the context of a different set of labels. For example, if participants were given "No Finding" labels defined in relation to the other 39 classes in the training set, what then would "No Finding" mean when evaluated on a new set of 5 unseen diseases? For this reason, we extract new labels for a "Normal" class, which simply means that no cardiopulmonary disease or abnormality was found (again, excluding "Support Devices").

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.

All training set labels were automatically extracted from the radiology report associated with each image using MedText (https://github.com/bionlplab/medtext) [18,19]. Test set labels were manually annotated as described in Holste et al. [14]. Briefly, 6 annotators were asked to identify the presence or absence of each of the 26 disease findings of interest in a set of 409 radiology reports. The same manual annotation workflow will be applied to the 200 radiology reports associated with the newly added MIDRC images.

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.

First, we conducted a meeting with the annotators to familiarize them with the annotation standards (e.g., if a disease is remarked to be "possibly" present, then we will not consider that disease to be present). Next, all annotators practiced the annotation process on a few example reports, with additional instructions through follow-up meetings and a detailed instruction file. Finally, annotation was performed through a customized AWS MTurk framework, which displayed the deidentified radiology report with tick boxes to mark the presence or absence of each of the 26 clinical findings of interest. Prior to annotation, all reports were preprocessed with RadText [19] to identify and highlight all relevant disease mentions in the text to ease the annotation process.

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.

Regarding the acquisition of "gold standard" labels, annotators consisted of an undergraduate, three PhD students, a postdoctoral researcher, and a professor. In this setting, the task of annotation did not require medical expertise, as it involved denoting the presence or absence of a disease in a passage of text (radiology report). As explained above, annotators underwent multiple training sessions to clarify the labeling process and were informed of synonyms for the 26 findings of interest. See Holste et al. [14] for further details.

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.

For manual annotation, each radiology report was annotated by at least two annotators. If the two annotators disagreed on the presence/absence of a particular finding, a third annotator would then reannotate the presence/absence of this finding to rectify the ambiguity (without knowledge of the other two annotators' opinions).

Data pre-processing method(s)

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

The image data used in this competition comes from MIMIC-CXR-JPG v2.0.0 (https://physionet.org/content/mimic-cxr-jpg/2.0.0/), which contains CXR images in JPEG format. While the original MIMIC-CXR provided data in the standardized DICOM format, MIMIC-CXR-JPG aimed to make the data more accessible (less storage needed) after applying preprocessing steps such as deidentification, histogram equalization, then conversion to JPEG format. Full details can be found in Johnson et al. [22]

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 source of error in automatic labeling is that the curated list of label synonyms may not have fully covered all possible mentions in the radiology reports -- hence, some mentions of certain labels could go undetected. However, to measure the the ability of MedText to accurately detect disease findings, we randomly selected 200 test set reports and manually annotated five disease findings: Calcification of the Aorta, Pneumomediastinum, Pneumoperitoneum, Subcutaneous Emphysema, Tortuous Aorta. Using these manual annotations as a reference, the automated MedText predictions for these five diseases achieved an average precision of 0.91, average recall of 0.94, and average F1 score of 0.92 across the classes. The highest precision (1.0) was achieved for Calcification of the Aorta, the highest recall (1.0) for Pneumomediastinum and Pneumoperitoneum, and the highest F1 score (0.97) for Tortuous Aorta.

Regarding manual annotation of the test set, errors in annotation may be caused by failure to notice the mention of a relevant finding, failure to notice the negation of a finding, failure to consistently treat uncertain mentions (e.g., one annotator marks "possible atelectasis" as present but another marks this as negative). However, as reported in Holste et al. [14], the overall agreement rate on all individual binary presence/absence annotations was 93.2% for the first round of labeling and 94.9% for the second round.

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

The annotation methods used in this challenge (automatic or manual) can only, at best, capture the opinion of the radiologist interpreting the CXR imaging, which of course is subject to error. As noted in Holste et al. [14], "Even for highly trained experts, diagnosis from CXR is difficult and complex, leading to high inter-reader variability." [23,24]

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)

Mean average precision (mAP), area under the receiver operating characteristic curve (mAUROC), and F1 score (mF1), and expected calibration error (mECE). Each metric will be computed for each class in the test set and then "macro-averaged" to form overall mAP, mAUROC, mF1, and mECE. The specific implementations of the first three metrics are provided by the scikit-learn library [25]. The primary metric is mAP, and mAUROC will be used to break ties if necessary. F1 is computed by binarizing predicted probabilities with a threshold of 0.5.

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

As explained in the prior version of CXR-LT (https://codalab.lisn.upsaclay.fr/competitions/12599 -> "Learn the Details" -> "Evaluation"), the primary evaluation metric for this challenge will be mAP. While mAUROC is a standard metric for related CXR classification datasets, AUROC can be heavily inflated in the face of strong class imbalance [26,27]. We argue that mAP (summarizing area under the precision-recall curve) is more appropriate for the long-tailed, multi-label setting since it both (i) measures performance across decision thresholds and (ii) does not degrade under imbalance [28]. For thoroughness and accordance with the literature, mAUROC will also be calculated and appear on the leaderboard, only being used in ranking to break ties. For additional evaluation of precision and recall with a fixed decision threshold, mF1 score will also be calculated and appear on the leaderboard. Finally, to provide auxiliary assessment of model calibration (beyond class discrimination), mECE will be computed and appear on the leaderboard.

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.

Ranking will be determined primarily by macro-averaged mAP (higher is better), the average AP value over all target classes present in the test. If ties are present, mAUROC (higher is better) will be used to break ties. *Each task is completely independent with regard to evaluation and ranking.*

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

Submissions with missing results will be discarded (submission will fail). The evaluation code used by CodaLab will ensure that predictions for all test set cases are present in the uploaded CSV file before performance evaluation.

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

Ranking is based primarily on mAP for the reasons outlined above. mAUROC is also included (and used as a tiebreaker) because it is a "standard" metric for related tasks of disease classification on CXR. While we understand the importance of assessing ranking stability via methods like bootstrapping (discussed later), the final ranking of teams will depend on the single scalar metric values obtained on the official test set for this task.

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.

Assessment of variability via bootstrapping will be performed for top-performing solutions as an additional analysis in the challenge overview paper, but final rankings will be determined by the single metric value calculated from the original test set predictions and labels. The evaluation code will use standard third-party Python libraries such as NumPy [29], scikit-learn [25], and pandas [30].

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

Bootstrapping test cases was chosen since it is easy to implement and enables assessment of variability in model performance and ranking stability.

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.

As discussed above, an assessment of variability in predictive performance (for each team/model) and ranking stability (across teams/models) will be performed via bootstrapping. This will be explored in our challenge overview paper, but not used to determine official team rankings.

TASK 3: Zero-shot classification of previously unseen diseases

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.

Chest radiography, like many diagnostic medical exams, produces a long-tailed distribution of clinical findings; while a small subset of diseases is routinely observed, the vast majority of diseases are relatively rare [1]. This poses a challenge for standard deep learning methods, which exhibit bias toward the most common classes at the expense of the important but rare "tail" classes [2]. Many existing methods [3] have been proposed to tackle this specific type of imbalance, though only recently has attention been given to long-tailed medical image recognition problems [4-6]. Diagnosis on chest X-rays (CXRs) is also a multi-label problem, as patients often present with multiple disease findings simultaneously; however, only a select few studies incorporate knowledge of label co-occurrence into the learning process [7-9, 12]. Since most large-scale image classification benchmarks contain single-label images with a mostly balanced distribution of labels, many standard deep learning methods fail to accommodate the class imbalance and co-occurrence problems posed by the long-tailed, multi-label nature of tasks like disease diagnosis on CXRs [2]. This task will evaluate a model's ability to perform "out-of-distribution" long-tailed, multi-label disease classification on CXRs when evaluated on a large test set with noisy, automatically text-mined labels that have not been encountered during training.

Keywords

List the primary keywords that characterize the task.

Long-tailed learning, Zero-shot classification, Chest radiography, Computer-aided diagnosis

ORGANIZATION

Organizers

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

Organizing Committee:
Yifan Peng (chair), PhD, Weill Cornell Medicine
Mingquan Lin, PhD, Weill Cornell Medicine
Gregory Holste, The University of Texas at Austin
Song Wang, The University of Texas at Austin
Yiliang Zhou, Weill Cornell Medicine
Hao Chen, Hong Kong University of Science and Technology
Atlas Wang, PhD, The University of Texas at Austin
Steering Committee:
Adam E. Flanders, MD, Thomas Jefferson University
Leo Anthony Celi, MD, MPH, MSc, MIT/Harvard
Zhiyong Lu, PhD, FACMI, NIH

George Shih, MD, Weill Cornell Medicine
Ronald M. Summers, MD, PhD, NIH

b) Provide information on the primary contact person.

Yifan Peng, Assistant Professor of Population Health Sciences, Weill Cornell Medicine, yip4002@med.cornell.edu

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

Challenge venue and platform

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

MICCAI 2024

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

CodaLab (https://codalab.lisn.upsaclay.fr/)

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

TBD. Last year's CXR-LT challenge URL can be found at https://codalab.lisn.upsaclay.fr/competitions/12599.

Participation policies

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

None. Solutions must be 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.

As in the first iteration of CXR-LT, we will allow training on any publicly available data and/or the use of models pretrained on publicly available data. Participants must be careful not to use models that were pretrained on MIMIC-CXR images that are part of the validation or test sets of this challenge; for example, any publicly available MIMIC-CXR-pretrained model is nearly certain to have used different random data splits than we will, meaning this model will have been trained on held-out validation and/or test set cases in this challenge (prohibited).

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.

Members of the organizers' institutes may participate so long as they have no personal relationship or conflict of interest with any of the organizers.

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

1st place: $500, 2nd place: $300, 3rd place: $200. Special recognition may be given to the best-performing team across all three tracks (e.g., of teams who participate in all three tracks, award a bonus prize to the team with the smallest average rank across tracks).

e) Define the policy for result announcement.

Examples:

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

The final ranking of teams will be kept private until MICCAI 2024. However, individual top-performing teams selected to present at MICCAI 2024 will be informed of their results and ranking. Furthermore, the solutions of top-performing teams will be made publicly available in the form of a publication summarizing the challenge.

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).

Up to two members of each of the top-performing teams (TBD precisely how many teams will be invited/how many from each track/etc.) will be invited to coauthor a paper summarizing the CXR-LT challenge. Participating teams may post preprints of their challenge solutions (e.g., on arXiv), which we will cite in our challenge overview paper as appropriate, after the competition. If a team wishes to publish their own paper, they must either waive their right to coauthor the challenge summary paper or wait until after the summary paper is published to submit their own paper.

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.

Teams will submit comma separated value (CSV) files containing their model's predictions on held-out test data via the CodaLab platform. Similar to last year's challenge (https://codalab.lisn.upsaclay.fr/competitions/12599 -> "Learn the Details" -> "Submission Format"), instructions will be given to ensure that the submitted file is in the proper format (e.g., columns correspond to the test set disease classes and there is a row/prediction for every test set image).

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.

As with the prior version of CXR-LT, there will be a months-long "development phase", where participants can submit predictions on a held-out development/validation set (10% of the data). The development phase will feature a live public leaderboard, displaying the performance metrics and ranking of each team's best-performing solution (according to the primary evaluation metric).

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

05/01/2024: Training data released and challenge (development phase) begins
08/01/2024: Test labels released and final evaluation (test phase) begins 08/04/2024: Test phase ends and competition is closed
08/15/2024: Top-performing teams invited to present at MICCAI 2024
10/06/2024: MICCAI 2024 CXR-LT challenge 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).

This challenge uses image data from MIMIC-CXR-JPG v2.0.0 (https://physionet.org/content/mimic-cxr-jpg/2.0.0/), a publicly available dataset for which ethics approval has already been acquired. For full details, see Johnson et al. [15].

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)

Participants are required to sign and abide by the data use agreement for MIMIC-CXR-JPG v2.0.0: https://physionet.org/content/mimic-cxr-jpg/view-dua/2.0.0/. As in last year's challenge, participants are only granted permission to participate if they provide acceptable proof (via Google Form) that they have properly met the registration requirements for MIMIC-CXR-JPG access. Instructions for applying to the prior version of CXR-LT can be found here: https://codalab.lisn.upsaclay.fr/competitions/12599 ("Learn the Details" -> "Terms and Conditions").

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.

Evaluation code will be made available with the release of the training data.

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

All participating teams are encouraged to provide reproducible, open-source training and inference code. On CodaLab, participants will be instructed to submit all relevant code and submission will fail if a "code" directory is not present in their upload.

Conflicts of interest

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

The CXR-LT challenge received funding from the Artificial Intelligence Journal (AIJ). Only the challenge organizers will have access to all test set labels.

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, research

Task category(ies)

State the task category(ies)

Examples:

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

Classification, prediction

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 consists of patients who have undergone chest X-ray imaging.

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

The challenge cohort uses data from MIMIC-CXR, which consists of patients who underwent chest X-ray imaging at the Beth Israel Deaconess Medical Center emergency department between 2011-2016. Full details can be found in Johnson et al. [15].

Imaging modality(ies)

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

X-ray imaging (projectional radiography)

Context information

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

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

For the training set only, disease labels denoting the presence or absence of 40 clinical findings will be provided in CSV format.

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

No patient information will be explicitly provided to participants for this challenge. However, MIMIC-CXR includes basic information such as sex and age, which participants are free to use.

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.

Chest shown in X-ray image

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.

Cardiopulmonary disease present on chest X-ray

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)).

The goal of this task is to accurately identify thoracic diseases present in chest X-rays *that have not been encountered during training*, where clinical findings follow a long-tailed distribution and can co-occur with one another. The relevant properties to optimize are accuracy, precision, sensitivity, and specificity. This task also involves robustness to noisy labels extracted by automatic text mining of radiology reports.

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).

This information can potentially be found in the DICOM metadata associated with each image in MIMIC-CXR (https://physionet.org/content/mimic-cxr/2.0.0/).

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

The MIMIC-CXR dataset was acquired during routine clinical practice in an emergency department, involving chest X-rays acquired for a variety of reasons, on several different scanners, and potentially from several different views of the patient. Details can be found in Johnson et al. [15].

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.

Beth Israel Deaconess Medical Center Emergency Department

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).

Regarding image data acquisition, this information is not available. The data was collected over years of routine clinical practice in an emergency department by many professionals of many different experience levels.

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.

For training, a case is a chest X-ray image (and, optionally, patient information such as sex and age, when available) with labels describing the presence and absence of 40 clinical findings. These 40 findings are: Adenopathy, Atelectasis, Azygos Lobe, Calcification of the Aorta, Cardiomegaly, Clavicle Fracture, Consolidation, Edema, Emphysema, Enlarged Cardiomediastinum, Fibrosis, Fissure, Fracture, Granuloma, Hernia, Hydropneumothorax, Infarction, Infiltration, Kyphosis, Lobar Atelectasis, Lung Lesion, Lung Opacity, Mass, Normal, Nodule, Pleural Effusion, Pleural Other, Pleural Thickening, Pneumomediastinum, Pneumonia, Pneumoperitoneum, Pneumothorax, Pulmonary Embolism, Pulmonary Hypertension, Rib Fracture, Round Atelectasis, Subcutaneous Emphysema, Support Devices, Tortuous Aorta, Tuberculosis.

For testing, a case is a chest X-ray image (and, optionally, patient information such as sex and age, when available) with labels describing the presence and absence of 5 clinical findings that have *not* been encountered in the training set. These labels were automatically text-mined from radiology reports as the other 40 findings. These 5 findings are: Bulla, Cardiomyopathy, Hilum, Osteopenia, Scoliosis.

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

Training: 263,977; Validation/development: 37,711; Test: 75,422

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

The data was randomly partitioned into training (70%), validation/development (10%), and test (20%) sets at the patient level to avoid data leakage. These proportions were chosen such that as much labeled data as possible could be allotted to the training set while still leaving a sufficient number of positive examples for extremely rare classes to facilitate reliable evaluation (at minimum, 30). Critically, the splits are different from those of last year's challenge.

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

The distribution of labels is extremely long-tailed with an "imbalance ratio" (ratio describing the prevalence of the most- to least-common class) over 685, exceeding that of challenge long-tailed benchmark datasets such as ImageNet-LT [16] and iNaturalist2018 [17]. The dataset is also distinct from these "single-label" benchmarks in that labels frequently co-occur in complex patterns that reflect underlying biology.

Also, we replace the "No Finding" class with what we believe to be a more natural "Normal" class. Previously, "No Finding" meant that none of the other clinically relevant findings were present in the study (except for "Support Devices", which is not a clinically meaningful finding). The fact that "No Finding" is defined in relation to other classes included in the set of labels means that "No Finding" will have a different, or ambiguous, meaning when considered in the context of a different set of labels. For example, if participants were given "No Finding" labels defined in relation to the other 39 classes in the training set, what then would "No Finding" mean when evaluated on a new set of 5 unseen diseases? For this reason, we extract new labels for a "Normal" class, which simply means that no cardiopulmonary disease or abnormality was found (again, excluding "Support Devices").

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.

All labels were automatically text-mined from the radiology report associated with each image using MedText (https://github.com/bionlplab/medtext) [18,19]. MedText preprocesses radiology reports through a sequential workflow encompassing de-identification, section segmentation, and sentence splitting. For each label of interest, we manually mined and curated a list of synonyms from various data sources, including RadLex radiology lexicon (https://radlex.org/), Medical Subject Headings (MeSH) (https://www.ncbi.nlm.nih.gov/mesh/), etc. We then utilized rule-based named entity recognition to detect the mentions of our predefined labels from the radiology reports; specifically, spaCy's PhraseMatcher (https://spacy.io/api/phrasematcher) was adopted by MedText as part of this process. Since negative medical findings indicate the absence of findings within the radiology report, identifying the negation status of medical findings is as important as identifying the positive findings. MedText uses NegBio [18] for negation detection, which utilizes universal dependencies for pattern definition and subgraph matching for graph traversal search so that the scope for detecting negation is not limited to the fixed word distance. Eventually, the detected disease findings and their identified assertion/negation status were compiled to produce the labels for the radiology report associated with each image.

The list of newly added clinical findings was chosen from sources including the disease list of the PadChest dataset [20] and Hansell et al. [21], ensuring that a sufficient number of occurrences were observed in the dataset.

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.

N/A

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.

MedText preprocesses radiology reports through a sequential workflow encompassing de-identification, section segmentation, and sentence splitting. For each label of interest, we manually mined and curated a list of synonyms from various data sources, including RadLex radiology lexicon (https://radlex.org/), Medical Subject Headings (MeSH) (https://www.ncbi.nlm.nih.gov/mesh/), etc. We then utilized rule-based named entity recognition to detect the mentions of our predefined labels from the radiology reports; specifically, spaCy's PhraseMatcher (https://spacy.io/api/phrasematcher) was adopted by MedText as part of this process. Since negative medical findings indicate the absence of findings within the radiology report, identifying the negation status of medical findings is as important as identifying the positive findings. MedText uses NegBio [18] for negation detection, which utilizes universal dependencies for pattern definition and subgraph matching for graph traversal search so that the scope for detecting negation is not limited to the fixed word distance.

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

N/A

Data pre-processing method(s)

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

The image data used in this competition comes from MIMIC-CXR-JPG v2.0.0 (https://physionet.org/content/mimic-cxr-jpg/2.0.0/), which contains CXR images in JPEG format. While the original MIMIC-CXR provided data in the standardized DICOM format, MIMIC-CXR-JPG aimed to make the data more accessible (less storage needed) after applying preprocessing steps such as deidentification, histogram equalization, then conversion to JPEG format. Full details can be found in Johnson et al. [22]

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 source of error in automatic labeling is that the curated list of label synonyms may not have fully covered all possible mentions in the radiology reports -- hence, some mentions of certain labels could go undetected. However, to measure the the ability of MedText to accurately detect disease findings, we randomly selected 200 test set reports and manually annotated five disease findings: Calcification of the Aorta, Pneumomediastinum, Pneumoperitoneum, Subcutaneous Emphysema, Tortuous Aorta. Using these manual annotations as a reference, the automated MedText predictions for these five diseases achieved an average precision of 0.91, average recall of 0.94, and average F1 score of 0.92 across the classes. The highest precision (1.0) was achieved for Calcification of the Aorta, the highest recall (1.0) for Pneumomediastinum and Pneumoperitoneum, and the highest F1 score (0.97) for Tortuous Aorta.

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

The annotation methods used in this challenge (automatic or manual) can only, at best, capture the opinion of the radiologist interpreting the CXR imaging, which of course is subject to error. As noted in Holste et al. [14], "Even for highly trained experts, diagnosis from CXR is difficult and complex, leading to high inter-reader variability." [23,24]

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)

Mean average precision (mAP), area under the receiver operating characteristic curve (mAUROC), and F1 score (mF1), and expected calibration error (mECE). Each metric will be computed for each class in the test set and then "macro-averaged" to form overall mAP, mAUROC, mF1, and mECE. The specific implementations of the first three metrics are provided by the scikit-learn library [25]. The primary metric is mAP, and mAUROC will be used to break ties if necessary. F1 is computed by binarizing predicted probabilities with a threshold of 0.5.

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

As explained in the prior version of CXR-LT (https://codalab.lisn.upsaclay.fr/competitions/12599 -> "Learn the Details" -> "Evaluation"), the primary evaluation metric for this challenge will be mAP. While mAUROC is a standard metric for related CXR classification datasets, AUROC can be heavily inflated in the face of strong class imbalance [26,27]. We argue that mAP (summarizing area under the precision-recall curve) is more appropriate for the long-tailed, multi-label setting since it both (i) measures performance across decision thresholds and (ii) does not degrade under imbalance [28]. For thoroughness and accordance with the literature, mAUROC will also be calculated and appear on the leaderboard, only being used in ranking to break ties. For additional evaluation of precision and recall with a fixed decision threshold, mF1 score will also be calculated and appear on the leaderboard. Finally, to provide auxiliary assessment of model calibration (beyond class discrimination), mECE will be computed and appear on the leaderboard.

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.

Ranking will be determined primarily by macro-averaged mAP (higher is better), the average AP value over all target classes present in the test. If ties are present, mAUROC (higher is better) will be used to break ties. *Each task is completely independent with regard to evaluation and ranking.*

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

Submissions with missing results will be discarded (submission will fail). The evaluation code used by CodaLab will ensure that predictions for all test set cases are present in the uploaded CSV file before performance evaluation.

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

Ranking is based primarily on mAP for the reasons outlined above. mAUROC is also included (and used as a tiebreaker) because it is a "standard" metric for related tasks of disease classification on CXR. While we understand the importance of assessing ranking stability via methods like bootstrapping (discussed later), the final ranking of teams will depend on the single scalar metric values obtained on the official test set for this task.

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.

Assessment of variability via bootstrapping will be performed for top-performing solutions as an additional analysis in the challenge overview paper, but final rankings will be determined by the single metric value calculated from the original test set predictions and labels. The evaluation code will use standard third-party Python libraries such as NumPy [29], scikit-learn [25], and pandas [30].

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

Bootstrapping test cases was chosen since it is easy to implement and enables assessment of variability in model performance and ranking stability.

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.

As discussed above, an assessment of variability in predictive performance (for each team/model) and ranking stability (across teams/models) will be performed via bootstrapping. This will be explored in our challenge overview paper, but not used to determine official team rankings.

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] Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A, Prince JL, Rueckert D, Summers RM. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE. 2021 Feb 26;109(5):820-38.

[2] Holste G, Wang S, Jiang Z, Shen TC, Shih G, Summers RM, Peng Y, Wang Z. Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study. In Data Augmentation, Labelling, and Imperfections: Second MICCAI Workshop, DALI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings 2022 Sep 16 (pp. 22-32). Cham: Springer Nature Switzerland.

[3] Zhang Y, Kang B, Hooi B, Yan S, Feng J. Deep long-tailed learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023 Apr 19.

[4] Zhang R, Haihong E, Yuan L, He J, Zhang H, Zhang S, Wang Y, Song M, Wang L. MBNM: multi-branch network based on memory features for long-tailed medical image recognition. Computer Methods and Programs in Biomedicine. 2021 Nov 1;212:106448.

[5] Ju L, Wang X, Wang L, Liu T, Zhao X, Drummond T, Mahapatra D, Ge Z. Relational subsets knowledge distillation for long-tailed retinal diseases recognition. In Medical Image Computing and Computer Assisted Intervention--MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part VIII 24 2021 (pp. 3-12). Springer International Publishing.

[6] Yang Z, Pan J, Yang Y, Shi X, Zhou HY, Zhang Z, Bian C. ProCo: Prototype-Aware Contrastive Learning for Long-Tailed Medical Image Classification. In Medical Image Computing and Computer Assisted Intervention--MICCAI 2022: 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part VIII 2022 Sep 16 (pp. 173-182). Cham: Springer Nature Switzerland.

[7] Chen H, Miao S, Xu D, Hager GD, Harrison AP. Deep hierarchical multi-label classification of chest X-ray images. In International Conference on Medical Imaging with Deep Learning 2019 May 24 (pp. 109-120). PMLR.

[8] Wang G, Wang P, Cong J, Liu K, Wei B. BB-GCN: A Bi-modal Bridged Graph Convolutional Network for Multi-label Chest X-Ray Recognition. arXiv preprint arXiv:2302.11082. 2023 Feb 22.

[9] Chen B, Li J, Lu G, Yu H, Zhang D. Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. IEEE Journal of Biomedical and Health Informatics. 2020 Jan 16;24(8):2292-302.

[10] Johnson AE, Pollard TJ, Greenbaum NR, Lungren MP, Deng CY, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042. 2019 Jan 21.

[11] PhysioNet. MIMIC-CXR-JPG - chest radiographs with structured labels [Internet]. Available from: https://physionet.org/content/mimic-cxr-jpg/2.0.0/.

[12] Moukheiber D, Mahindre S, Moukheiber L, Moukheiber M, Wang S, Ma C, Shih G, Peng Y, Gao M. Few-Shot

Learning Geometric Ensemble for Multi-label Classification of Chest X-Rays. In Data Augmentation, Labelling, and Imperfections: Second MICCAI Workshop, DALI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings 2022 Sep 16 (pp. 112-122). Cham: Springer Nature Switzerland.

[13] CodaLab. CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays [Internet]. Available from: https://codalab.lisn.upsaclay.fr/competitions/12599.

[14] Holste G, Zhou Y, Wang S, Jaiswal A, Lin M, Zhuge S, Yang Y, Kim D, Nguyen-Mau TH, Tran MT, Jeong J. Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge. arXiv preprint arXiv:2310.16112. 2023 Oct 24.

[15] Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data. 2019 Dec 12;6(1):317.

[16] Liu Z, Miao Z, Zhan X, Wang J, Gong B, Yu SX. Large-scale long-tailed recognition in an open world. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 2537-2546).

[17] Van Horn G, Mac Aodha O, Song Y, Cui Y, Sun C, Shepard A, Adam H, Perona P, Belongie S. The inaturalist species classification and detection dataset. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 8769-8778).

[18] Peng, Yifan, et al. "NegBio: a high-performance tool for negation and uncertainty detection in radiology reports." AMIA Summits on Translational Science Proceedings 2018 (2018): 188.

[19] Wang, Song, et al. "Radiology Text Analysis System (RadText): Architecture and Evaluation." IEEE Int Conf Healthc Inform. 2022 Jun; 2022: 288-296.

[20] Bustos A, Pertusa A, Salinas JM, De La Iglesia-Vaya M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis. 2020 Dec 1;66:101797.

[21] Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. Radiology. 2008 Mar;246(3):697-722.

[22] Johnson AE, Pollard TJ, Greenbaum NR, Lungren MP, Deng CY, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042. 2019 Jan 21.

[23] Hopstaken RM, Witbraad T, Van Engelshoven JM, Dinant GJ. Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections. Clinical radiology. 2004 Aug 1;59(8):743-52.

[24] Sakurada S, Hang NT, Ishizuka N, Toyota E, Hung LD, Chuc PT, Lien LT, Thuong PH, Bich PT, Keicho N, Kobayashi N. Inter-rater agreement in the assessment of abnormal chest X-ray findings for tuberculosis between

two Asian countries. BMC infectious diseases. 2012 Dec;12(1):1-8.

[25] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Nov 1;12:2825-30.

[26] Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Cham: Springer; 2018 Oct 22.

[27] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning 2006 Jun 25 (pp. 233-240).

[28] Rethmeier N, Augenstein I. Long-tail zero and few-shot learning via contrastive pretraining on and for small data. In Computer Sciences & Mathematics Forum 2022 May 20 (Vol. 3, No. 1, p. 10). MDPI.

[29] Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R. Array programming with NumPy. Nature. 2020 Sep 17;585(7825):357-62.

[30] McKinney W. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference 2010 Jun 28 (Vol. 445, No. 1, pp. 51-56).

Further comments

Further comments from the organizers.

N/A