Medical Image Analysis 97 (2024) 103224

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge

Gregory Holste [a], Yiliang Zhou [b], Song Wang [a], Ajay Jaiswal [a], Mingquan Lin [b], Sherry Zhuge [c], Yuzhe Yang [d], Dongkyun Kim [e], Trong-Hieu Nguyen-Mau [f], Minh-Triet Tran [f], Jaehyup Jeong [g], Wongi Park [h], Jongbin Ryu [h], Feng Hong [i], Arsh Verma [j], Yosuke Yamagishi [k], Changhyun Kim [l], Hyeryeong Seo [m], Myungjoo Kang [n], Leo Anthony Celi [o,p,q], Zhiyong Lu [r], Ronald M. Summers [s], George Shih [t], Zhangyang Wang [a], Yifan Peng [b,*]

[a] Department of Electrical and Computer Engineering, The University of Texas at Austin, 78712, Austin, TX, USA
[b] Department of Population Health Sciences, Weill Cornell Medicine, 10065, New York, NY, USA
[c] School of Information Systems, Carnegie Mellon University, 15213, Pittsburgh, PA, USA
[d] Department of Electrical Engineering and Computer Science, Massachussetts Institute of Technology, 02139, Cambridge, MA, USA
[e] School of Computer Science, Carnegie Mellon University, 15213, Pittsburgh, PA, USA
[f] University of Science, VNU-HCM, Ho Chi Minh City, Viet Nam
[g] KT Research & Development Center, KT Corporation, 06763, Seoul, South Korea
[h] Department of Software and Computer Engineering, Ajou University, 16499, Suwon, South Korea
[i] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, 200240, Shanghai, China
[j] Wadhwani Institute for Artificial Intelligence, 400079, Mumbai, India
[k] Division of Radiology and Biomedical Engineering, Graduate School of Medicine, The University of Tokyo, 113-0033, Tokyo, Japan
[l] BioMedical AI Team, AIX Future R&D Center, SK Telecom, 04539, Seoul, South Korea
[m] Interdisciplinary Program in AI (IPAI), Seoul National University, 02504, Seoul, South Korea
[n] Department of Mathematical Sciences, Seoul National University, 02504, Seoul, South Korea
[o] Laboratory for Computational Physiology, Massachusetts Institute of Technology, 02139, Cambridge, MA, USA
[p] Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, 02215, Boston, MA, USA
[q] Department of Biostatistics, Harvard T.H. Chan School of Public Health, 02115, Boston, MA, USA
[r] National Center for Biotechnology Information, National Library of Medicine, 20894, Bethesda, MD, USA
[s] Clinical Center, National Institutes of Health, 20892, Bethesda, MD, USA
[t] Department of Radiology, Weill Cornell Medicine, 10065, New York, NY, USA

ARTICLE INFO

Dataset link: https://physionet.org/content/cxr-lt-iccv-workshop-cvamd/1.1.0/

Keywords:
Chest X-ray
Long-tailed learning
Computer-aided diagnosis

ABSTRACT

Many real-world image recognition problems, such as diagnostic medical imaging exams, are "long-tailed" – there are a few common findings followed by many more relatively rare conditions. In chest radiography, diagnosis is both a *long-tailed* and *multi-label* problem, as patients often present with multiple findings simultaneously. While researchers have begun to study the problem of long-tailed learning in medical image recognition, few have studied the interaction of label imbalance and label co-occurrence posed by long-tailed, multi-label disease classification. To engage with the research community on this emerging topic, we conducted an open challenge, **CXR-LT**, on long-tailed, multi-label thorax disease classification from chest X-rays (CXRs). We publicly release a large-scale benchmark dataset of over 350,000 CXRs, each labeled with at least one of 26 clinical findings following a long-tailed distribution. We synthesize common themes of top-performing solutions, providing practical recommendations for long-tailed, multi-label medical image classification. Finally, we use these insights to propose a path forward involving vision-language foundation models for few- and zero-shot disease classification.

* Corresponding author.
E-mail addresses: atlaswang@utexas.edu (Z. Wang), yip4002@med.cornell.edu (Y. Peng).

https://doi.org/10.1016/j.media.2024.103224
Received 24 October 2023; Received in revised form 1 April 2024; Accepted 27 May 2024
Available online 31 May 2024

1. Introduction

Like many diagnostic medical exams, chest X-rays (CXRs) yield a long-tailed distribution of clinical findings. This means that while a small subset of diseases are routinely observed, the majority are quite rare (Zhou et al., 2021). This long-tailed distribution challenges conventional deep learning methods, as they tend to favor common classes and often overlook the infrequent yet crucial classes. In response, several methods (Zhang et al., 2023b) have been proposed lately with a focus on addressing label imbalance in long-tailed medical image recognition tasks (Zhang et al., 2021a; Ju et al., 2021, 2022; Yang et al., 2022). Of note, diagnosing from CXRs is not only a long-tailed problem, but also *multi-label*, since patients often present with multiple disease findings simultaneously. Despite this, only a limited number of studies have incorporated knowledge of label co-occurrence into their learning process (Chen et al., 2020; Wang et al., 2023; Chen et al., 2019a).

Owing to the fact that most large-scale image classification benchmarks feature single-label images with a predominately balanced label distribution, we establish a new benchmark for long-tailed, multi-label medical image classification. Specifically, we expanded the MIMIC-CXR (Johnson et al., 2019a) dataset by increasing the set of target disease findings from 14 to 26. This is achieved by introducing 12 new disease findings by parsing the radiology reports associated with each CXR study.

In our effort to engage with the community on this emerging interdisciplinary topic, we have released the data and launched the **CXR-LT** challenge on long-tailed, multi-label thorax disease classification on CXRs. In this paper, we summarize the CXR-LT challenge, consolidate key insights from top-performing solutions, and offer practical perspective for advancing long-tailed, multi-label medical image classification. Finally, we use our findings to suggest a path forward toward few- and zero-shot disease classification in the long-tailed, multi-label setting by leveraging multimodal foundation models.

Our contributions can be summarized as follows:

1. We have publicly released a large multi-label, long-tailed CXR dataset containing 377,110 images. Each image is labeled with one or multiple labels from a set of 26 disease findings. In addition, we have provided a "gold standard" subset encompassing human-annotated consensus labels.
2. We conducted **CXR-LT**, a challenge for long-tailed, multi-label thorax disease classification on CXRs. We summarize insights from top-performing teams and offer practical recommendations for advancing long-tailed, multi-label medical image classification.
3. Based on the insights from CXR-LT, we propose a methodological path forward for few- and zero-shot generalization to unseen disease findings via multimodal foundation models.

2. Methods

2.1. Dataset curation

In this section, we detail the data curation process of two datasets: (i) the CXR-LT dataset used in the challenge, and (ii) a manually annotated "gold standard" test set used for additional evaluation of top-performing solutions after the conclusion of the challenge.

2.1.1. CXR-LT dataset

The CXR-LT challenge dataset[1] was created by extending the label set of the MIMIC-CXR dataset[2] (Johnson et al., 2019a), resulting in a more challenging, long-tailed label distribution. MIMIC-CXR is a large-scale, publicly available dataset of de-identified CXRs and radiology reports. The dataset contains a total of 377,110 frontal and lateral CXR images acquired from 227,835 studies conducted during routine clinical practice at the Beth Israel Deaconess Medical Center (Boston, Massachusetts, USA) emergency department from 2011–2016.

Following Holste et al. (2022), the radiology reports associated with each CXR study were parsed via RadText (Wang et al., 2022), a radiology text analysis tool, to extract the presence status of *12 new rare disease findings*: (1) Calcification of the Aorta, (2) Emphysema, (3) Fibrosis, (4) Hernia, (5) Infiltration, (6) Mass, (7) Nodule, (8) Pleural Thickening, (9) Pneumomediastinum, (10) Pneumoperitoneum, (11) Subcutaneous Emphysema, and (12) Tortuous Aorta. These particular findings were selected by (i) identifying diseases found in the NIH ChestXRay dataset (Wang et al., 2017) labels that were not present in the MIMIC-CXR labels and (ii) discussing with radiologists which findings might be important to include that were not captured by existing public CXR datasets. The latter point led to the inclusion of Calcification of the Aorta and Tortuous Aorta (cardiac findings) as well as Pneumomediastinum, Pneumoperitoneum, and Subcutaneous Emphysema (trapped air in undesirable cavities or under the skin).

The resulting dataset consisted of 377,110 CXRs, each labeled with at least one of 26 disease findings following a long-tailed distribution (Fig. 1). Though MIMIC-CXR contained the images and text reports needed for additional labeling, we used images from the MIMIC-CXR-JPG dataset (Johnson et al., 2019b) in this challenge since the preprocessed JPEG images (~600 GB) would be more accessible to participants than the raw DICOM data (~4.7 TB) provided in MIMIC-CXR.[3] Finally, the dataset was randomly split into training (70%), development (10%), and test sets (20%) at the patient level to avoid label leakage. Challenge participants would have access to all images, but only have access to labels for the training set.

2.1.2. Gold standard test set

While the CXR-LT dataset is large and challenging due to heavy label imbalance and label co-occurrence, it inevitably suffers from label noise much like other datasets with automatically text-mined labels (Abdalla and Fine, 2023). To remedy this, we aimed to construct a "gold standard" set, derived from the challenge test set, with labels that were manually annotated after analyzing the radiology reports. This smaller, but higher quality, dataset would then be used as an auxiliary test set to perform additional evaluation of the top-performing CXR-LT solutions after the conclusion of the challenge.

To build a gold standard set for evaluation, we first curated a sample of test set reports with at least two positive disease findings as determined by RadText. This was necessary to ensure a sufficient number of positive examples for tail classes; given the extreme rarity of certain findings (as low as 0.2% prevalence), a random sample of several hundred reports may not even yield a single instance of several rare findings, prohibiting proper performance evaluation. A subset of 451 such reports were manually annotated by six human readers, who marked the presence or absence of the 26 disease findings in each radiology report. Before manual labeling, all reports were preprocessed through RadText (Wang et al., 2022) to identify and highlight all relevant disease mentions in the text in order to ease the annotation process. Each annotator was then provided with the reports and a list of synonyms for each of the 26 findings. Annotators were asked to select all disease findings that were *conclusively* affirmed positive in the report. Following MIMIC-CXR, annotators could select "No Finding" if no other findings (except "Support Devices") were present.

Before annotation, a training session was held to align the standards among annotators where each annotator practiced by labeling 10 reports. Any disagreements in this phase were discussed until consensus was reached, leading to the formulation of a shared annotation

[1] https://physionet.org/content/cxr-lt-iccv-workshop-cvamd/1.1.0/.
[2] https://physionet.org/content/mimic-cxr/2.0.0/.
[3] https://physionet.org/content/mimic-cxr-jpg/2.0.0/.

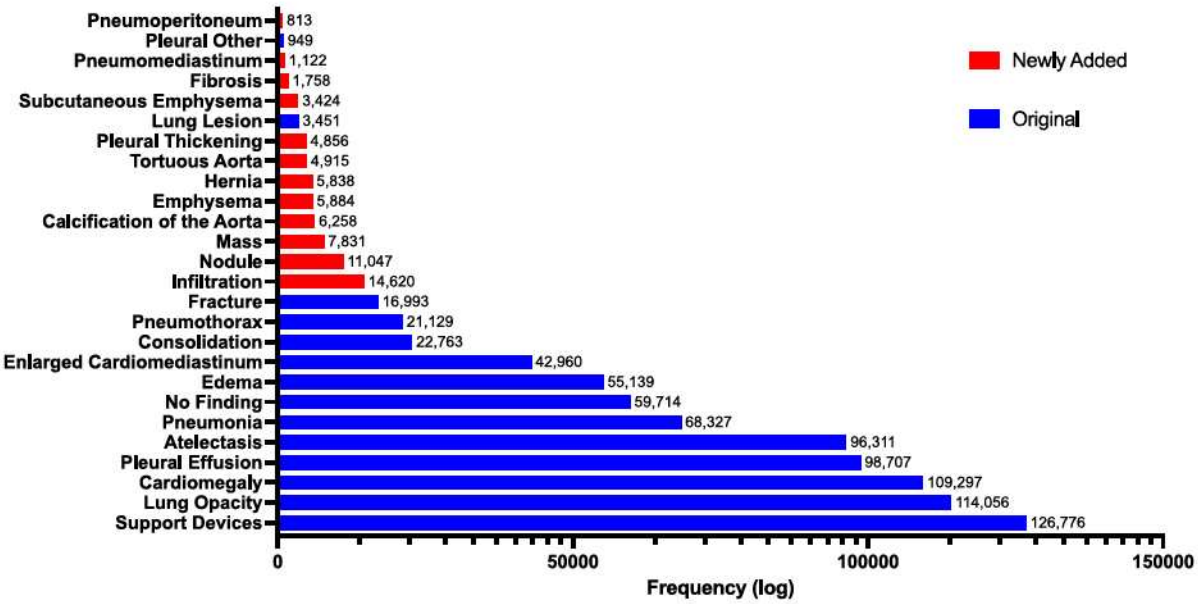

Fig. 1. Long-tailed distribution of the CXR-LT 2023 challenge dataset. The dataset was formed by extending the MIMIC-CXR (Johnson et al., 2019a) benchmark to include 12 new clinical findings (red) by parsing radiology reports.

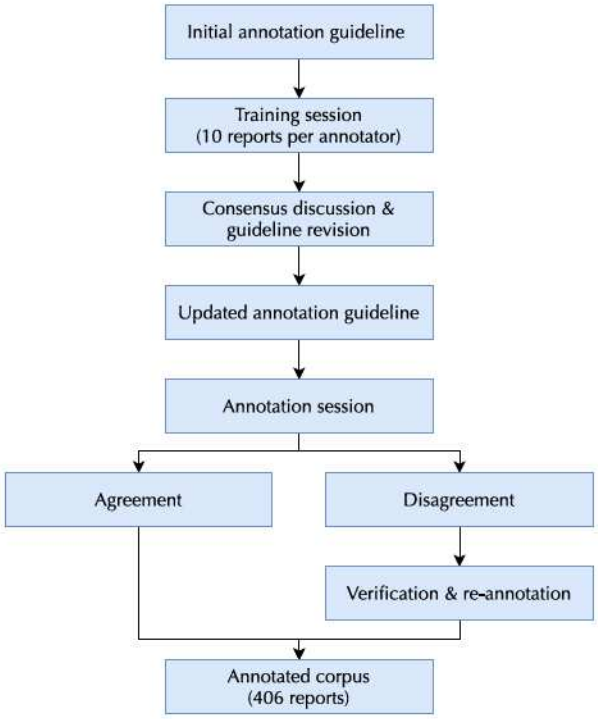

Fig. 2. Flowchart describing CXR-LT gold standard dataset annotation.

guideline (Fig. 2). Following the training session, the official annotation process consisted of two rounds: the first round covering 200 reports and the second round covering 251 reports. After each round, individual disease-level disagreements between annotators on a given report were compiled and adjudicated by a third annotator. For the first round, the overall agreement rate was 93.2% and the Cohen's Kappa coefficient was 0.795; for the second round, the agreement rate was 94.9% with a Cohen's kappa of 0.778. After removing reports that were not annotated by at least two readers, the CXR-LT gold standard set consisted of 406 cases. The resulting label distribution of the gold standard set can be found in Supplementary Fig. 1.

2.2. CXR-LT challenge task

The CXR-LT challenge was formulated as a 26-way multi-label classification problem. Given a CXR, participants were tasked with detecting all disease findings present. If no findings were present, participants could predict "No Finding", with the exception that "No Finding" can co-occur with "Support Devices" as this is not a clinically meaningful *diagnostic* finding. Since this is a multi-label classification problem with severe label imbalance, the primary evaluation metric was mean average precision (mAP), specifically, the "macro-averaged" AP across the 26 classes. While area under the receiver operating characteristic curve (AUROC) is a standard metric employed for related datasets (Wang et al., 2017; Seyyed-Kalantari et al., 2021), AUROC can be heavily inflated in the presence of class imbalance (Fernández et al., 2018; Davis and Goadrich, 2006). Instead, mAP is more appropriate for the long-tailed, multi-label setting since it measures performance across decision thresholds and does not degrade under class imbalance (Rethmeier and Augenstein, 2022). For thoroughness, mean AUROC (mAUROC) and mean F1 score – using a decision threshold of 0.5 for each class – were also calculated.

The challenge was conducted on CodaLab (Pavao et al., 2023). Any registered CodaLab user could apply to participate, but since this competition used MIMIC-CXR-JPG data (Johnson et al., 2019b), which requires credentialing and training through PhysioNet (Goldberger et al., 2000), participants were required to submit proof of PhysioNet credentials to enter. During the Development Phase, registered participants downloaded the labeled training set and (unlabeled) development set, for which they would generate a comma-separated values (CSV) file with predictions to upload. Submissions were then evaluated on the held-out development set and results were updated to a live, public leaderboard. During the Test Phase, test set images (without labels) were released. Participants were asked to submit CSV files with predictions on the much larger, held-out test set and were only given a maximum of 5 successful attempts. For this phase, the leaderboard was kept hidden and the single best-scoring submission (by mAP) by each team was retained. The final Test Phase leaderboard was used to rank participants, primarily by mAP, then by mAUROC in the event of ties.

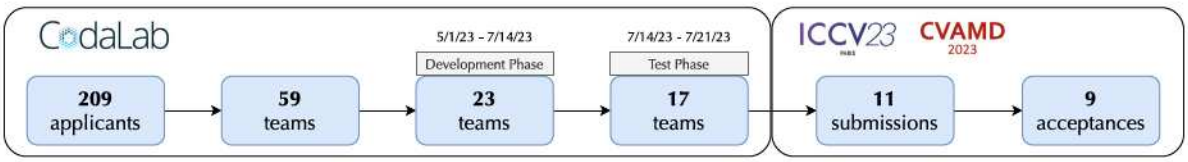

Fig. 3. Flowchart describing CXR-LT challenge participation. Over 200 teams applied to participate in the challenge on CodaLab, and 59 teams met registration requirements. Of the 17 teams that participated in the Test Phase, 11 submitted their written solutions for presentation at the ICCV CVAMD 2023 workshop. The top 9 of these submissions were accepted to the workshop and are described in this paper.

Table 1
Overview of top-performing CXR-LT challenge solutions. ENS = ensemble; RW = loss re-weighting.

Team	Rank	Image resolution	Backbone	ENS	RW	Pretraining	Notes
T1	1	1024	ConvNeXt-S		✓	ImageNet → CheXpert, NIH, VinDr	Two-stage training; cross-view Transformer; ML-Decoder classifier (label as text)
T2	2	512, 768	EfficientNetV2-S, ConvNeXt-S	✓	✓	ImageNet	Heavy mosaic augmentation
T3	3	448	ConvNeXt-B	✓	✓	ImageNet21K → NIH	Ensemble of "head" and "tail" experts
T4	4	384	ConvNeXt-B		✓	ImageNet	Custom robust asymmetric loss (RAL)
T5	5	512	ResNet50[a]	✓	✓	ImageNet[a]	Vision-language modeling (label as text); co-train on NIH, CheXpert
T6	6	448	ResNeXt101, DenseNet161	✓		ImageNet → CheXpert, NIH, PadChest	Used synthetic data to augment tail classes
T7	7	224–512	EfficientNetV2-S	✓		ImageNet, ImageNet21k	Three-stage training with increasing resolution
T8	8	448	TResNet50	✓		ImageNet	Heavy CutMix-like augmentation; feature pyramid with deep supervision
T9	11	1024	ResNet101	✓		ImageNet	RIDE mixture of experts; LSE pooling; label as text/graph with cross-modal attention

[a] T5 additionally used a Transformer text encoder pretrained on PubMedBERT (Gu et al., 2020) and Clinical-T5 (Lehman and Johnson, 2023).

3. Results

3.1. CXR-LT challenge participation

The CXR-LT challenge received 209 team applications on CodaLab,[4] of which 59 were approved after providing proof of credentialed access to MIMIC-CXR-JPG (Johnson et al., 2019b). During the Development Phase, 23 teams participated, contributing a total of 525 unique submissions to the public leaderboard. Ultimately, 17 teams participated in the final Test Phase, and 11 of these teams submitted papers describing their challenge solution to the ICCV CVAMD 2023 workshop.[5] The 9 accepted workshop papers, representing the top-performing teams in the CXR-LT challenge, were used for study in this paper (Fig. 3).

3.2. Methods of top-performing teams

A summary of top-performing solutions can be found in Table 1, including Test Phase rank, image resolution, backbone architecture used, and other methodological characteristics. Though each solution is described in the paragraphs below, please refer to the paper in each subsection title for full details.

3.2.1. T1 (Kim, 2023)

This team used a two-stage framework that aggregated features across views (e.g., frontal and lateral CXRs). In the first stage, a ConvNeXt-S model (Liu et al., 2022) model was pretrained with Noisy Student (Xie et al., 2020) self-training on the external CXR datasets NIH ChestXRay (Wang et al., 2017), CheXpert (Irvin et al., 2019), and VinDrCXR (Nguyen et al., 2022). Using this backbone as a frozen feature extractor, a Transformer then aggregated multi-view features in a given study. T1 used the ML-Decoder (Ridnik et al., 2023) classification head, which represents the labels as text and performs cross-attention over image (CXR) and text (label) features. Finally, this team utilized a weighted asymmetric loss (Ridnik et al., 2021a) to combat the inter-class imbalance caused by the long-tailed distribution and intra-class imbalance caused by the dominance of negative labels in multi-label classification.

3.2.2. T2 (Nguyen-Mau et al., 2023)

This team utilized augmentation, ensemble, and re-weighting methods for imbalanced multi-label classification. Specifically, they used an ensemble of EfficientNetV2-S (Tan and Le, 2021) and ConvNeXt-S (Liu et al., 2022) models. Of note, the team made use of heavy "mosaic" augmentation (Bochkovskiy et al., 2020), randomly tiling four CXRs into a single image and using the union of their label sets as ground truth. They used a weighted focal loss (Lin et al., 2017) to handle imbalance, then test-time augmentation and a multi-level ensemble across both model architectures and individual models obtained by stratified cross-validation to improve generalization.

3.2.3. T3 (Jeong et al., 2023)

This team proposed an ensemble method based on ConvNeXt-B (Liu et al., 2022) with the CSRA classifier (Zhu and Wu, 2021). After pretraining on the NIH ChestXRay14 (Wang et al., 2017) dataset, T3 trained three separate models, respectively, on CXR-LT data only from "head" classes, "tail" classes, and all classes; an average of these three models formed the final output. Each model utilized a weighted cross-entropy loss and the Lion optimizer (Chen et al., 2023).

[4] https://codalab.lisn.upsaclay.fr/competitions/12599.
[5] https://cvamd2023.github.io/.

Table 2
Final Test Phase results of the CXR-LT 2023 challenge. Presented is average precision (AP) of each team's final model on all 26 classes evaluated on the test set. The best AP for a given class is highlighted in bold.

	T1	T2	T3	T4	T5	T6	T7	T8	T9
Atelectasis	**0.622**	0.609	0.611	0.607	0.606	0.610	0.602	0.595	0.546
Calcification of the Aorta	**0.162**	0.140	0.145	0.143	0.135	0.109	0.130	0.116	0.111
Cardiomegaly	0.661	0.652	0.652	0.648	0.653	0.652	**0.668**	0.640	0.581
Consolidation	**0.240**	0.228	0.234	0.219	0.228	0.230	0.225	0.218	0.171
Edema	**0.563**	0.553	0.559	0.556	0.554	0.557	0.551	0.545	0.497
Emphysema	**0.210**	0.193	0.193	0.193	0.180	0.184	0.161	0.165	0.128
Enlarged Cardiomediastinum	0.186	0.184	0.184	0.184	**0.186**	0.185	0.183	0.177	0.140
Fibrosis	**0.167**	0.163	0.157	0.153	0.154	0.154	0.116	0.132	0.120
Fracture	**0.379**	0.262	0.269	0.289	0.243	0.262	0.171	0.219	0.171
Hernia	0.570	**0.585**	0.563	0.551	0.539	0.538	0.499	0.484	0.343
Infiltration	**0.063**	0.057	0.060	0.060	0.057	0.058	0.056	0.055	0.049
Lung Lesion	0.034	**0.042**	0.041	0.038	0.040	0.031	0.031	0.031	0.021
Lung Opacity	**0.617**	0.597	0.603	0.597	0.594	0.598	0.590	0.584	0.529
Mass	**0.250**	0.224	0.213	0.206	0.200	0.227	0.187	0.167	0.112
Nodule	**0.267**	0.192	0.204	0.200	0.180	0.196	0.137	0.166	0.117
Pleural Effusion	**0.843**	0.829	0.831	0.832	0.830	0.805	0.822	0.822	0.781
Pleural Other	**0.070**	0.037	0.043	0.040	0.039	0.016	0.059	0.042	0.007
Pleural Thickening	**0.137**	0.108	0.116	0.110	0.126	0.083	0.119	0.097	0.055
Pneumomediastinum	0.332	0.384	0.376	0.339	**0.387**	0.284	0.326	0.308	0.096
Pneumonia	**0.312**	0.305	0.311	0.309	0.304	0.308	0.292	0.294	0.258
Pneumoperitoneum	**0.324**	0.316	0.261	0.283	0.303	0.237	0.262	0.235	0.155
Pneumothorax	**0.602**	0.533	0.549	0.553	0.511	0.546	0.451	0.474	0.427
Subcutaneous Emphysema	**0.598**	0.556	0.564	0.560	0.570	0.520	0.507	0.538	0.492
Support Devices	**0.918**	0.906	0.916	0.913	0.910	0.910	0.894	0.903	0.887
Tortuous Aorta	**0.066**	0.061	0.060	0.060	0.058	0.056	0.063	0.053	0.045
No Finding	0.486	0.478	**0.488**	0.479	0.485	0.468	0.471	0.469	0.428
Mean	**0.372**	0.354	0.354	0.351	0.349	0.339	0.330	0.328	0.279

3.2.4. T4 (Park et al., 2023)

This team proposed a novel robust asymmetric loss (RAL) for multi-label long-tailed classification. RAL improves upon the popular focal loss (Lin et al., 2017) by including a Hill loss term (Zhang et al., 2021b), which mitigates sensitivity to the negative term of the original focal loss. The team used an ImageNet-pretrained ConvNeXt-B (Liu et al., 2022) with the proposed RAL loss and data augmentation following (Azizi et al., 2021; Chen et al., 2019b).

3.2.5. T5 (Hong et al., 2023)

This team used a vision-language modeling approach leveraging large pre-trained models. The authors utilized an ImageNet-pretrained ResNet50 (He et al., 2016) and text encoder pre-trained on PubMedBERT (Gu et al., 2020) and Clinical-T5 (Lehman and Johnson, 2023) to extract features from images and label text, respectively. For multi-label classification, they employed a multi-label Transformer query network to aggregate image and text features. To handle imbalance, the team used class-specific loss re-weighting informed by validation set performance. They also incorporated external training data (NIH ChestXRay 14 Wang et al., 2017 and CheXpert Irvin et al., 2019), used test-time augmentation, and performed "class-wise" ensembling to improve generalization.

3.2.6. T6 (Verma, 2023)

This team used domain-specific pretraining, ensembling, and synthetic data augmentation to improve performance. With an ImageNet-pretrained ResNeXt101 (Xie et al., 2017) and DenseNet101 (Huang et al., 2017), the team further pretrained on the CXR benchmarks NIH ChestXRay14 (Wang et al., 2017), CheXpert (Irvin et al., 2019), and PadChest (Bustos et al., 2020). This team also used RoentGen (Chambon et al., 2022), a multimodal generative model for synthesizing CXRs from natural language, to generate additional CXRs for tail classes in order to combat imbalance.

3.2.7. T7 (Yamagishi and Hanaoka, 2023)

This team used a multi-stage training scheme with ensembling and oversampling. For the first stage, an ImageNet21k-pretrained EfficientNetV2-S (Tan and Le, 2021) was trained on 224 × 224 resolution images. The weights from this model were then used to train on 320 × 320 and 384 × 384 images, then 512 × 512 images. An ensemble was then formed by averaging the predictions of four models trained on various resolutions, with some models leveraging oversampling of minority classes to mitigate class imbalance. Test-time augmentation and view-based post-processing were also used to boost performance.

3.2.8. T8 (Kim et al., 2023)

This team utilized an ImageNet-pretrained TResNet50 (Ridnik et al., 2021b) with heavy augmentation and ensembling. The authors made use of MixUp (Zhang et al., 2017), which linearly combines training images and their labels, and CutMix (Yun et al., 2019), which "cuts and pastes" regions from one training image onto another. They also used a "feature pyramid" approach, extracting pooled features from four layers throughout the network and aggregating these multi-scale features.

3.2.9. T9 (Seo et al., 2023)

This team built upon ML-GCN (Chen et al., 2019b), a framework for multi-label image classification, which uses GloVe (Pennington et al., 2014) to embed each label as a node within a graph capable of incorporating the co-occurrence patterns of labels. To combat the long-tailed distribution of classes, the authors trained a ResNet101 (He et al., 2016) with class-balanced sampling and the Routing DIverse Experts (RIDE) method (Wang et al., 2020) to diversify the members of their ensemble (Zhang et al., 2023b). They also used log-sum-exp (LSE) pooling (Pinheiro and Collobert, 2015) and a Transformer encoder to attend over image and text features.

3.3. CXR-LT challenge results

3.3.1. CXR-LT test phase results

Detailed Test Phase results of 9 top-performing CXR-LT teams can be found in Table 2. T1, the 1st-placed team, reached an mAP of 0.372, considerably outperforming the 2nd–5th-placed teams, who performed similarly with mAP ranging from 0.349 to 0.354; further, T1 achieved best performance on 20 out of 26 classes. Maximum per-class AP ranged

Table 3
Long-tailed classification performance on "head", "medium", and "tail" classes by average mAP within each category. These categories were determined by relative frequency of each class in the training set (denoted in parentheses). The rightmost column denotes the average of head, medium, and tail mAP. The best mAP in each column appears in bold.

	Overall	Head (>10%)	Medium (1–10%)	Tail (<1%)	Avg
T1	**0.372**	**0.499**	**0.246**	0.242	**0.329**
T2	0.354	0.482	0.226	0.227	0.311
T3	0.354	0.477	0.226	**0.246**	0.316
T4	0.351	0.480	0.221	0.220	0.307
T5	0.349	0.474	0.218	0.243	0.312
T6	0.339	0.476	0.210	0.179	0.288
T7	0.330	0.460	0.195	0.216	0.290
T8	0.328	0.461	0.195	0.195	0.284
T9	0.279	0.420	0.154	0.086	0.220

Table 4
Comparison of disease prevalence with automated text-mined labeling vs. manual human annotation of the 406 radiology reports in our gold standard test set.

	Automated	Human	Δ
Atelectasis	0.488	0.308	−37%
Calcification of the Aorta	0.062	0.116	+88%
Cardiomegaly	0.448	0.392	−13%
Consolidation	0.214	0.182	−15%
Edema	0.313	0.249	−20%
Emphysema	0.113	0.071	−37%
Enlarged Cardiomediastinum	0.313	0.291	−7%
Fibrosis	0.081	0.054	−33%
Fracture	0.133	0.118	−11%
Hernia	0.074	0.047	−37%
Infiltration	0.113	0.03	−74%
Lung Lesion	0.076	0.015	−81%
Lung Opacity	0.571	0.49	−14%
Mass	0.096	0.049	−49%
No Finding	0.091	0.091	+0%
Nodule	0.096	0.081	−15%
Pleural Effusion	0.562	0.451	−20%
Pleural Other	0.052	0.047	−10%
Pleural Thickening	0.052	0.054	+5%
Pneumomediastinum	0.099	0.086	−12%
Pneumonia	0.283	0.054	−81%
Pneumoperitoneum	0.084	0.059	−29%
Pneumothorax	0.214	0.121	−44%
Subcutaneous Emphysema	0.108	0.103	−4%
Support Devices	0.564	0.544	−4%
Tortuous Aorta	0.059	0.081	+38%

widely from 0.063 (Infiltration) to 0.918 (Support Devices), likely owing to the challenges posed by label imbalance and noise. Interestingly, T1 demonstrated outstanding performance particularly on the Fracture (0.379 mAP vs. next-best 0.289) and Nodule (0.267 mAP vs. next-best 0.204) classes. Since T1 was the only team to explicitly fuse multi-view information, it is conceivable that this learned aggregation of frontal- and lateral-view information aided performance particularly for findings like Fracture and Nodule, which may better be resolved by multiple views. Additional Test Phase results by AUROC can be found in Supplementary Table 1.

3.3.2. Long-tailed classification performance

To examine predictive performance by label frequency, we split the 26 target classes into "head" (>10%), "medium" (1%–10%), and "tail" (<1%) categories based on prevalence in the training set. Category-wise mAP is presented in Table 3, as well as a "category-wise average" of head, medium, and tail mAP. We first observe that T1 considerably outperformed all other teams particularly on the more common head and medium classes, though other teams performed comparably or slightly better (T3 and T5) on tail classes. We also find that T1–T5 on average outperformed the remaining four teams on tail classes much more so than head and medium classes. Critically, T1–T5 were the only five teams to use loss re-weighting, which appears to have provided clear benefits in modeling rare diseases.

3.4. Evaluation on a gold standard test set

As described in Section 2.1.2, a "gold standard" test set of 406 newly labeled CXRs was curated for additional evaluation on a small subset of the challenge test set with higher-quality labels. This section describes differences in human vs. automated annotation patterns and predictive performance with human vs. automated labels in this gold standard set.

3.4.1. Differences in human vs. text-mined annotation

To form a head-to-head comparison of disease annotation patterns of human readers vs. automatic text mining tools, we examined the distribution of manual and automated labels on the 406 gold standard reports (Table 4). Overall, we find that automatic labeling produced a higher volume of positive findings per image (median: 5, mean ± std: 5.4 ± 1.9) than manual annotation (median: 4, mean ± std: 4.2 ± 1.8). This is likely a consequence of the conservative labeling approach adopted by annotators after the initial training session. While we elected to only mark the presence of a finding if it was unambiguously affirmed positive in the report, text analysis tools like RadText do not necessarily adhere to the same rule. For example, potentially ambiguous phrases like "cannot rule out pneumonia" and "likely effusion" might be marked positive by a text analysis tool but not by human annotators.

While manual labeling produced fewer positive findings in general, we find that certain diseases were substantially more and less likely to be marked positive by a human reader than RadText. For example, as seen in Table 4, Pneumonia and Lung Lesion both saw an >80% drop in prevalence with human labeling. On the other hand, Calcification of the Aorta (88%) and Tortuous Aorta (38%) – both newly added cardiac findings – were considerably more prevalent upon human annotation, suggesting a low recall for these aortic findings with RadText. Interestingly, there was no difference in the number of reports marked No Finding, indicating that it is perhaps straightforward for both humans and text analysis tools to recognize the absence of findings in a normal report.

Since this is a multi-label classification problem, we also examine differences in co-occurrence behavior between findings in the text-mined vs. gold standard labels. We first computed a conditional co-occurrence probability matrix for each dataset, C^{test} and C^{gold}, where entry (i, j) represents the observed conditional probability that finding j was present given finding i was present: $C_{i,j} = P(j|i)$. We then computed $C^{\text{gold}} - C^{\text{test}}$, the difference in conditional co-occurrence probability between human vs. automated labels, visualized as a heatmap in Supplementary Fig. 2. We observe certain irregularities such as an >0.6 drop in P(Enlarged Cardiomediastinum | Pneumomediastinum) in the gold standard labels. Upon examination, we find that P(Enlarged Cardiomediastinum | Pneumomediastinum) = 1 in the text-mined labels, meaning every time Pneumomediastinum occurred, so did Enlarged Cardiomediastinum. This potentially reflects a shortcoming of text-mined labeling, where the presence of "mediastinum" may be falsely interpreted as a positive assertion of both Pneumomediastinum and Enlarged Cardiomediastinum; in reality, these are two very different clinical findings that are unlikely to occur together.

3.4.2. Gold standard test set results

Detailed results of top-performing teams on the gold standard test set can be found in Table 5 (additional results by AUROC in Supplementary Table 2). AP values were generally higher in the gold standard set than the original challenge test set, with certain classes experience large changes in performance — for example, the maximum AP jumps from 0.162 to 0.688 for Calcification of the Aorta and from 0.598 to 0.887 for Subcutaneous Emphysema. However, these dramatic differences are to be expected when considering that the gold standard test set is not a representative subset of the challenge test set; for reasons outlined in Section 2.1.2, the gold standard set consisted of studies that were

Table 5
Gold standard test set results from CXR-LT 2023 participants. Presented is average precision (AP) of each team's final model on all 26 classes evaluated on our human-annotated gold standard test set. The best AP for a given class is highlighted in bold.

	T1	T2	T3	T4	T5	T6	T7	T8	T9
Atelectasis	0.465	0.481	0.494	0.453	**0.500**	0.464	0.444	0.455	0.449
Calcification of the Aorta	0.658	0.609	**0.688**	0.662	0.621	0.578	0.544	0.613	0.541
Cardiomegaly	0.696	0.718	0.718	0.704	0.720	0.732	**0.754**	0.718	0.664
Consolidation	0.415	0.449	0.471	0.474	0.436	**0.476**	0.411	0.426	0.406
Edema	0.600	0.590	**0.601**	0.572	0.589	0.587	0.571	0.587	0.540
Emphysema	0.298	0.356	0.362	0.309	0.359	0.388	**0.394**	0.279	0.292
Enlarged Cardiomediastinum	0.351	**0.370**	0.349	0.349	0.337	0.341	0.328	0.314	0.337
Fibrosis	0.417	0.491	0.460	0.458	0.487	**0.497**	0.397	0.437	0.428
Fracture	**0.583**	0.455	0.535	0.494	0.524	0.501	0.385	0.366	0.389
Hernia	0.759	**0.808**	0.804	0.766	0.722	0.723	0.714	0.708	0.514
Infiltration	0.065	0.049	0.059	0.048	0.049	0.046	0.053	0.080	**0.095**
Lung Lesion	0.028	**0.071**	0.033	0.042	0.030	0.032	0.066	0.040	0.044
Lung Opacity	0.642	0.651	**0.678**	0.656	0.656	0.650	0.655	0.652	0.623
Mass	0.410	**0.477**	0.389	0.376	0.369	0.412	0.321	0.363	0.128
Nodule	**0.435**	0.325	0.296	0.300	0.272	0.359	0.362	0.257	0.212
Pleural Effusion	**0.856**	0.835	0.836	0.842	0.831	0.808	0.822	0.837	0.804
Pleural Other	**0.385**	0.212	0.224	0.259	0.230	0.121	0.249	0.245	0.138
Pleural Thickening	**0.386**	0.249	0.249	0.204	0.244	0.190	0.251	0.210	0.132
Pneumomediastinum	0.771	0.830	0.795	0.800	0.823	0.757	**0.837**	0.776	0.569
Pneumonia	0.114	**0.131**	0.127	0.123	0.111	0.111	0.114	0.103	0.096
Pneumoperitoneum	0.586	0.539	**0.632**	0.539	0.557	0.443	0.499	0.526	0.487
Pneumothorax	**0.675**	0.649	0.701	0.641	0.598	0.663	0.562	0.568	0.587
Subcutaneous Emphysema	0.845	0.825	0.852	**0.887**	0.874	0.777	0.774	0.813	0.837
Support Devices	0.952	0.950	0.957	**0.960**	0.956	0.951	0.932	0.946	0.944
Tortuous Aorta	0.362	0.329	0.322	0.309	0.336	**0.386**	0.265	0.299	0.280
No Finding	0.731	**0.841**	0.834	0.729	0.852	0.828	0.815	0.807	0.766
Mean	**0.519**	0.511	0.518	0.498	0.503	0.493	0.481	0.478	0.435

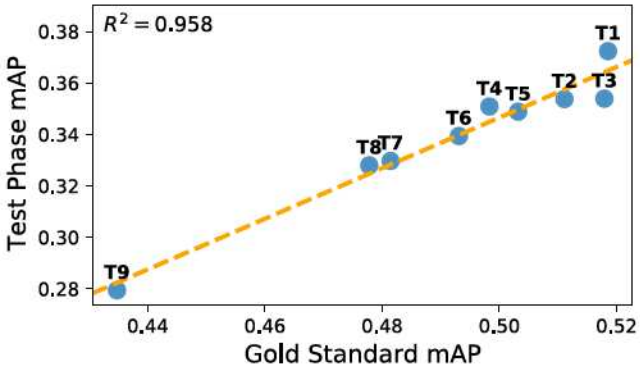

Fig. 4. Comparison of performance on CXR-LT Test Phase data (Section 2.1.1) and gold standard test data (Section 2.1.2).

far more likely to contain many disease findings than the overall test set (median: 2, mean ± std: 2.4 ± 1.5). A direct comparison of class-wise performance – specifically, the mean AP for each class across all 9 teams – using automated text-mined vs. human labels in the 406 gold standard test cases can be found in Supplementary Table 3. Here, we observe that performance on "gold standard" human labels vs. text-mined labels dropped for 19 out of 26 classes and experienced a 12.3% drop in AP on average. This can be attributed to the large label distribution shift outlined in Section 3.4.1, meaning models have been optimized on the particular noise patterns inherent in the text-mined training set labels.

Despite this large distribution shift, the overall correspondence of team results remained consistent between the official CXR-LT challenge test set and the gold standard set (Fig. 4; $R^2 = 0.958$, $r = 0.979$, $P = 4.7 \times 10^{-6}$). Even the team rankings remained consistent with the exception of T3 outperforming T2 and T5 outperforming T4 on the gold standard test set. This is to be expected when considering that teams T2–T5 were separated by, at most, 0.005 mAP in the CXR-LT Test Phase.

3.5. Ranking stability analysis

All results presented thus far are point estimates that ignore the variability in predictive performance metrics — for instance, how meaningful is it that T3 outperformed T4 by 0.001 mAP? We address this problem with a ranking stability analysis, whereby we perform 500 bootstrap samples of both the CXR-LT challenge test set and gold standard test set. Fig. 5 depicts the percentage of simulated bootstrap "trials" for which each team achieved a certain rank. On the CXR-LT challenge test set, ranking was generally stable, with the most common ("modal") simulated rank corresponding to the actual observed rank for all 9 teams. While T1 and T9 always, respectively, placed 1st and 9th, T2 and T3 were nearly interchangeable, with T2 placing 2nd in 52% of trials and T3 placing 2nd in the remaining 48%.

In contrast, simulated team rankings were far more volatile when evaluated on the gold standard test set. For example, T1 – which considerably outperformed other teams in the official challenge – placed 1st in under half of all simulated rankings, placing as low as 5th in some instances. Additionally, T2 placed 1st in 40% of trials, perhaps supported by the fact that the cosine similarity between the predicted probabilities of T1 and T2 was 0.95, considerably higher than between T1 and any other team (Supplementary Fig. 3). Overall, this variability is to be expected given the small sample size of the gold standard set; even when sampling "disease-heavy" studies to ensure that every finding was represented, certain tail classes were still only observed a handful of times (as few as 9), resulting in noisy estimates of performance. While improved label quality is valuable, this highlights the pervasive importance of sample size and justifies our choice for using the large-scale, text-mined labels for CXR-LT Test Phase evaluation. Despite this variability, however, the simulated modal rank actually corresponded to the original CXR-LT Test Phase rank in all instances but one: on the gold standard set, T4 placed 5th most often (43%) and T5 placed 4th most often (52%).

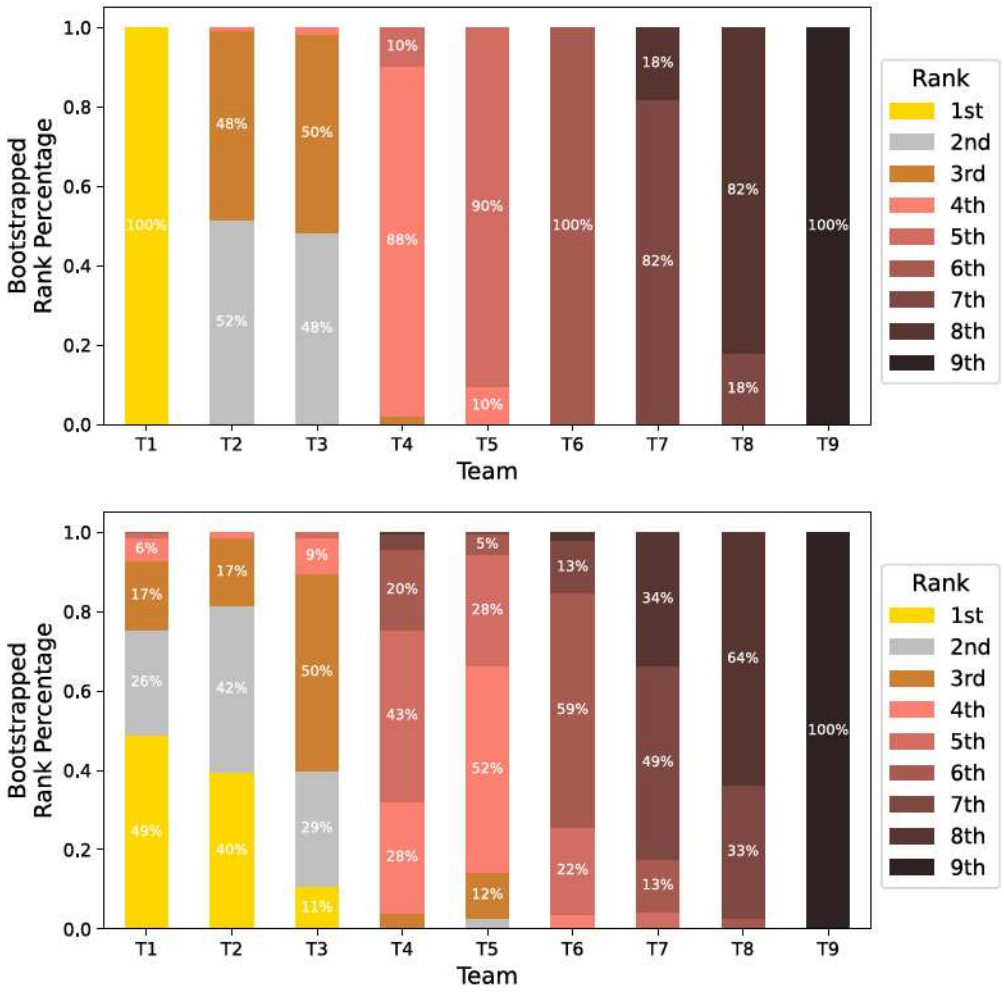

Fig. 5. Ranking stability analysis on CXR-LT Test Phase data (top) and gold standard test data (bottom). Simulated challenge rankings were computed via 500 bootstrap samples of each test set. Presented is the percentage of bootstrap trials for which each team achieved a given rank.

4. Discussion

4.1. Themes of successful solutions

As outlined in Table 1, several salient patterns emerge among top-performing challenge solutions. In summary, many successful CXR-LT solutions leveraged

- Relatively high image resolution (>300 × 300)
- Modern CNNs like ConvNeXt and EfficientNetV2
- Large-scale domain-specific pretraining on CXR data
- Strong data augmentation and ensemble learning
- Loss re-weighting to amplify tail classes
- Multimodal learning via text-based label representations.

High image resolution. Compared to the vast majority of research efforts in natural image recognition, top-performing CXR-LT solutions used relatively high image resolution — usually greater than the standard 224 × 224 and as high as 1024 × 1024. It is well-understood that increased image resolution improves visual recognition with deep neural networks (Touvron et al., 2019; Tan and Le, 2019, 2021), particularly in medical applications such as radiology (Thambawita et al., 2021; Sabottke and Spieler, 2020), where diagnosis can often rely on resolving faint or small abnormalities. In fact, Haque et al. (2023) specifically found 1024 × 1024 (used by two top-performing teams) to be an optimal spatial resolution for MIMIC-CXR-JPG disease classification, observing that 2048 × 2048 resolution actually degraded performance.

Modern convolutional neural network backbones. Also, despite the recent popularity of Vision Transformers (ViTs) (Khan et al., 2022), all top-performing solutions used a convolutional neural network (CNN) as the image encoder. The most popular choice was ConvNeXt (Liu et al., 2022), followed by the EfficientNet (Tan and Le, 2021) and ResNet (He et al., 2016) architecture families. This observation is echoed by a recent comprehensive benchmark of architectures on a wide array of visual recognition tasks (Goldblum et al., 2024): "Despite the recent attention paid to transformer-based architectures and self-supervised learning, high-performance convolutional networks pretrained via supervised learning outperform transformers on the majority of tasks..". The authors also found ConvNeXt, the most popular architecture among top CXR-LT performers, to outperform other backbones.

Large-scale pretraining. In this vein, all top-performing solutions utilized supervised pretraining or transfer learning of some kind. While many used standard ImageNet-pretrained models (some leveraging the larger ImageNet21k), several teams performed additional domain-specific pretraining on publicly available, external CXR datasets such as NIH ChestXRay (Wang et al., 2017), CheXpert (Irvin et al., 2019), and PadChest (Bustos et al., 2020). Specifically, T1, T3, and T5 all performed multi-stage pretraining, whereby they fine-tuned an ImageNet-pretrained backbone on large external CXR datasets for disease classification. Such a two-stage pretraining scheme with (1) "generalist" then (2) domain-specific pretraining has proven successful in prior works such as REMEDIS (Azizi et al., 2023) and Reed et al. (2022) in the context of self-supervised pretraining.

Ensemble learning and data augmentation. Many (7 out of 9) top solutions involved ensemble learning, aggregating the outputs of multiple models for improved generalization (Ganaie et al., 2022; Fort et al., 2019). Ensembles often benefit from diverse constituent models, and teams formed diverse ensembles in several unique ways: T2 and T5 formed ensembles across different model architectures, T2 and T7 formed ensembles across different image resolutions, and T3 formed an ensemble of head (common) class and tail (rare) class models. However, ensembling was not *strictly* necessary for high performance, evidenced by the fact that the 1st- and 4th-placed teams utilized a single well-trained model. Additionally, all teams used image augmentation, another standard technique to boost generalization (Xu et al., 2023). Notably, T2 and T8 employed "mosaic" and CutMix augmentation, respectively, effectively blending input images and labels as a form of regularization.

Loss re-weighting. Owing to the challenging long-tailed nature of this problem, the top five solutions all used loss re-weighting in order to adequately model rare classes. From Table 3, one can see that the performance gap between T1-T5 and the remaining four teams is most apparent on the uncommon medium and tail classes. Specifically, T1 used a weighted asymmetric loss (Ridnik et al., 2021a) specifically designed for imbalanced multi-label classification, T2 used a class-weighted focal loss (Lin et al., 2017), and T4 used a novel robust asymmetric loss that includes an additional regularization term to a weighted focal loss. It should also be noted that teams addressed the long-tailed distribution in ways other than loss re-weighting, such as mixture of experts (T3 and T9) and synthetic data generation of tail classes (T6).

Multimodal vision-language learning. Multimodal vision-language learning has recently become a popular approach in deep learning for radiology, typically as a means of pretraining on paired CXR imaging and free-text radiology reports (Chen et al., 2019a; Yan and Pei, 2022; Delbrouck et al., 2022; Moon et al., 2022; Li et al., 2024; Moor et al., 2023). While this challenge did not use radiology report data, three top-performing teams found success by directly representing the disease label information as *text*. For example, T1 used the ML-Decoder (Ridnik et al., 2023) classification head, which considers labels as text "queries" that participate in cross-attention with image features. T5, on the other hand, used a Transformer pretrained on large amounts of clinical text (Gu et al., 2020; Lehman and Johnson, 2023) to directly encode disease labels. For example, encoding the text "Pleural Effusion" rather than a one-hot label enables the model to embed this information in a semantically meaningful way in relation to the language it has already encountered during pretraining on medical text.

4.2. Limitations and future work

In addition to the common themes of successful solutions outlined above, it should be emphasized that the unique and often novel aspects of each team's solution also contributed to their success. For example, Kim (2023) leveraged a cross-view Transformer to aggregate information across radiographic views; Park et al. (2023) proposed a novel robust asymmetric loss (RAL), with additional experiments demonstrating improved performance on other long-tailed medical image classification tasks; Verma (2023) leveraged a vision-language foundation model, RoentGen (Chambon et al., 2022), to synthesize "tail" class examples to combat the long-tailed problem; Hong et al. (2023) took a vision-language approach leveraging Transformers pretrained on clinical text in order to learn rich representations of the multi-label disease information. One promising observation is that many teams reached very similar overall performance – measured by Test Phase mAP – with dramatically different methods. This suggests that the solutions from our participants may represent *orthogonal* contributions that, when combined, prove greater than the sum of their parts. Future work might unify the insights learned from top-performing solutions into a single long-tailed, multi-label medical image classification framework.

This study could also be strengthened by examining bias both in the data and the models put forth by top-performing teams. First, MIMIC-CXR was collected at a single institution – a major teaching hospital of Harvard Medical School – which may only represent the specific demographics of this setting. Second, prior work has shown that deep neural networks trained on single-institution CXR datasets often display disparities in predictive performance based on factors like race and sex (Seyyed-Kalantari et al., 2021). In fact, Seyyed-Kalantari et al. (2021) observed this effect on MIMIC-CXR-JPG specifically and noted that training on larger, multi-institutional datasets mitigated these disparities. It would be interesting to evaluate whether teams that pretrained on additional external CXR data not only achieved improved predictive performance but also improved group fairness. Future work may address subgroup fairness as an additional dimension of desired model behavior alongside predictive performance.

Regarding the data contributions of this work, we acknowledge that the CXR-LT dataset bears the same pitfalls as many other publicly available CXR benchmarks with automatically text-mined labels, namely label noise (Abdalla and Fine, 2023). We attempted to rectify this by manually annotating radiology reports to obtain a gold standard test set for additional evaluation. While this improved the label quality, this approach is limited in that it can only, at best, confirm the opinion of the individual radiologist writing the report. Even for highly trained experts, diagnosis from CXR is difficult and complex, leading to high inter-reader variability (Hopstaken et al., 2004; Sakurada et al., 2012). Ideally, a true "gold standard" dataset would consist of consensus labeling from multiple radiologists' interpretations. However, this is of course prohibitively expensive and time-consuming (Zhou et al., 2021) given the volume of labeled data required to train deep learning models and is the primary motivation for automatic disease labeling in the first place.

Though many diagnostic exams are long-tailed, most publicly available medical imaging datasets only include labels for a few common findings. The CXR-LT dataset thus represents a major contribution to due its large scale (>375,000 images), multi-label nature, and long-tailed label distribution of 26 clinical findings. However, a select few large-scale, long-tailed medical imaging datasets exist, such as HyperKvasir (Borgli et al., 2020) – containing >10,000 endoscopic images labeled with 23 findings – and PadChest (Bustos et al., 2020) – containing >160,000 CXRs labeled with 174 findings.

While CXR-LT and PadChest represent meaningful contributions to long-tailed learning from CXR, it should be noted that the "true" long tail of all clinical findings is at least an order of magnitude longer than any current publicly available dataset can offer. For example, Radiology Gamuts Ontology[6] (Budovec et al., 2014) documents 4691 unique radiological image findings. Thus, one way to enhance the CXR-LT dataset might be to include an even wider variety of automatically text-mined findings to mimic the extremely long tail of real-world CXR. However, this approach too has its own limitations. Even if we could construct a dataset with labels for up to 1000 clinical findings, ranging from common and well-studied to exceedingly rare, and train a model on this long-tailed data, what happens when a new finding is encountered? One might argue that the only way to tackle the *true* long-tailed distribution of imaging findings is to develop a model that can adaptively generalize to previously unseen diseases. Future iterations of the CXR-LT challenge will consider this problem through the lens of zero-shot classification: can participants train a model to accurately detect a clinical finding that the model has *not been trained on*?

[6] http://www.gamuts.net/about.php.

4.3. The future of long-tailed, multi-label learning

If zero-shot disease classification is the ultimate step toward clinically viable long-tailed medical image classification, then vision-language foundation models provide a very promising path forward. Several top-performing CXR-LT teams found success by encoding the label information as text, allowing for rich representation learning of the disease labels and their correlations. This allowed (Kim, 2023) to better handle the long-tailed, multi-label distribution via the text- and attention-based ML-Decoder classifier (Ridnik et al., 2023) and Seo et al. (2023) to exploit correlations between labels via graph learning of text-based label representations via ML-GCN (Chen et al., 2019b). While, for example, the approach of Hong et al. (2023) did not earn 1st place in this challenge, it would almost certainly prove the most useful when encountering a previously unseen disease finding, a practical scenario in real-world clinical deployment. Since (Hong et al., 2023) leverage Transformer encoders pretrained on large collections of clinical text including PubMedBERT (Gu et al., 2020) and Clinical-T5 (Lehman and Johnson, 2023), the model retains a rough *semantic* understanding of biomedical concepts via the textual representation of disease information. This in turn allows natural generalization to new findings by relating the text "prompt" of the potential new disease finding to relevant concepts encountered during pretraining, which is expressly not possible with standard unimodal deep learning methods.

Recent studies have shown that vision-language modeling with encoders pretrained on large collections of medical image and text data enable zero-shot disease classification, in some cases nearly reaching the performance of fully supervised approaches (Hayat et al., 2021; Tiu et al., 2022; Mishra et al., 2023; Zhang et al., 2023a; Li et al., 2024; Moor et al., 2023). For example, CheXzero (Tiu et al., 2022) performed Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) on paired CXR images and radiology reports, where the model was trained to properly match reports with their CXRs. At test time, this model could then encode text "prompts" of "<Pathology>" and "No <Pathology>", after which the model could determine which prompt best matched the given CXR, effectively performing zero-shot disease classification on par with board-certified radiologists. Critically, this ability to flexibly encode *any* new disease information would enable zero-shot learning of tail classes that might otherwise be difficult or impossible to obtain traditional labels for. Further, such a multimodal approach can be readily combined with many of the methods employed by top-performing CXR-LT teams such as loss re-weighting, heavy data augmentation, and ensemble learning. This would allow for a computer-aided diagnosis system capable of adaptively generalizing to unseen findings, thus capturing the true long tail of all imaging findings.

5. Conclusion

In summary, we curated and publicly released a large-scale dataset of over 375,000 CXR images for long-tailed, multi-label disease classification. We then hosted an open challenge, CXR-LT, to engage with the research community on this important task. We compiled and synthesized common threads through the most successful challenge solutions, providing practical recommendations for long-tailed, multi-label medical image classification. Lastly, we identify a path forward toward tackling the "true" long tail of all imaging findings via multimodal vision-language foundation models capable of zero-shot generalization to unseen diseases, which future iterations of the CXR-LT challenge will address.

CRediT authorship contribution statement

Gregory Holste: Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing, Visualization. **Yiliang Zhou:** Data curation, Software, Validation. **Song Wang:** Data curation, Methodology, Resources, Validation. **Ajay Jaiswal:** Data curation. **Mingquan Lin:** Data curation, Methodology, Writing – review & editing. **Sherry Zhuge:** Data curation. **Yuzhe Yang:** Data curation. **Dongkyun Kim:** Methodology, Software, Writing – review & editing. **Trong-Hieu Nguyen-Mau:** Methodology, Software, Writing – review & editing. **Minh-Triet Tran:** Methodology, Software, Writing – review & editing. **Jaehyup Jeong:** Methodology, Software, Writing – review & editing. **Wongi Park:** Methodology, Software, Writing – review & editing. **Jongbin Ryu:** Methodology, Software, Writing – review & editing. **Feng Hong:** Methodology, Software, Writing – review & editing. **Arsh Verma:** Methodology, Software, Writing – review & editing. **Yosuke Yamagishi:** Methodology, Software, Writing – review & editing. **Changhyun Kim:** Methodology, Software, Writing – review & editing. **Hyeryeong Seo:** Methodology, Resources, Writing – review & editing. **Myungjoo Kang:** Methodology, Resources, Writing – review & editing. **Leo Anthony Celi:** Supervision, Writing – review & editing. **Zhiyong Lu:** Supervision, Writing – review & editing. **Ronald M. Summers:** Supervision, Writing – review & editing. **George Shih:** Supervision, Writing – review & editing. **Zhangyang Wang:** Conceptualization, Funding acquisition, Supervision, Validation, Writing – original draft, Writing – review & editing. **Yifan Peng:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: R.M.S has received royalties for patent or software licenses from iCAD, Philips, PingAn, ScanMed, Translation Holdings, and MGB as well as research support form CRADA with PingAn. The remaining authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

CXR-LT challenge data is publicly available through https://physionet.org/content/cxr-lt-iccv-workshop-cvamd/1.1.0/.

Acknowledgments

This work was supported by the National Library of Medicine [grant number R01LM014306], the National Science Foundation [grant numbers 2145640, IIS-2212176], the Amazon Research Award, and the Artificial Intelligence Journal. It was also supported by the NIH Intramural Research Program, National Library of Medicine and Clinical Center. The authors would like to thank Alistair Johnson and the PhysioNet team for helping to publicize the challenge and host the data.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103224.

References

Abdalla, M., Fine, B., 2023. Hurdles to artificial intelligence deployment: Noise in schemas and "gold" labels. Radiol.: Artif. Intell. 5 (2), e220056.

Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., et al., 2023. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nat. Biomed. Eng. 1–24.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al., 2021. Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3478–3488.

Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al., 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data 7 (1), 283.

Budovec, J.J., Lam, C.A., Kahn Jr., C.E., 2014. Informatics in radiology: radiology gamuts ontology: Differential diagnosis for the Semantic Web. Radiographics 34 (1), 254–264.

Bustos, A., Pertusa, A., Salinas, J.-M., De La Iglesia-Vaya, M., 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Med. Image Anal. 66, 101797.

Chambon, P., Bluethgen, C., Delbrouck, J.-B., Van der Sluijs, R., Połacin, M., Chaves, J.M.Z., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A., 2022. RoentGen: Vision-language foundation model for chest x-ray generation. arXiv preprint arXiv:2211.12737.

Chen, B., Li, J., Lu, G., Yu, H., Zhang, D., 2020. Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. IEEE J. Biomed. Health Inform. 24 (8), 2292–2302.

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., Pham, H., Dong, X., Luong, T., Hsieh, C.-J., et al., 2023. Symbolic discovery of optimization algorithms. arXiv preprint arXiv:2302.06675.

Chen, H., Miao, S., Xu, D., Hager, G.D., Harrison, A.P., 2019a. Deep hierarchical multi-label classification of chest X-ray images. In: International Conference on Medical Imaging with Deep Learning. PMLR, pp. 109–120.

Chen, Z.-M., Wei, X.-S., Wang, P., Guo, Y., 2019b. Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5177–5186.

Davis, J., Goadrich, M., 2006. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 233–240.

Delbrouck, J.-b., Saab, K., Varma, M., Eyuboglu, S., Chambon, P., Dunnmon, J., Zambrano, J., Chaudhari, A., Langlotz, C., 2022. ViLMedic: A framework for research at the intersection of vision and language in medical AI. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 23–34.

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F., 2018. Learning from Imbalanced Data Sets. Vol. 10, Springer.

Fort, S., Hu, H., Lakshminarayanan, B., 2019. Deep ensembles: A loss landscape perspective. arXiv preprint arXiv:1912.02757.

Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M., Suganthan, P.N., 2022. Ensemble deep learning: A review. Eng. Appl. Artif. Intell. 115, 105151.

Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E., 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 101 (23), e215–e220.

Goldblum, M., Souri, H., Ni, R., Shu, M., Prabhu, V., Somepalli, G., Chattopadhyay, P., Ibrahim, M., Bardes, A., Hoffman, J., et al., 2024. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. Adv. Neural Inf. Process. Syst. 36.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2020. Domain-specific language model pretraining for biomedical natural language processing. arXiv:arXiv:2007.15779.

Haque, M.I.U., Dubey, A.K., Danciu, I., Justice, A.C., Ovchinnikova, O.S., Hinkle, J.D., 2023. Effect of image resolution on automated classification of chest X-rays. J. Med. Imaging 10 (4), 044503.

Hayat, N., Lashen, H., Shamout, F.E., 2021. Multi-label generalized zero shot learning for the classification of disease in chest radiographs. In: Machine Learning for Healthcare Conference. PMLR, pp. 461–477.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Holste, G., Wang, S., Jiang, Z., Shen, T.C., Shih, G., Summers, R.M., Peng, Y., Wang, Z., 2022. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In: MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. Springer, pp. 22–32.

Hong, F., Dai, T., Yao, J., Zhang, Y., Wang, Y., 2023. Bag of tricks for long-tailed multi-label classification on chest X-Rays. arXiv preprint arXiv:2308.08853.

Hopstaken, R., Witbraad, T., Van Engelshoven, J., Dinant, G., 2004. Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections. Clin. Radiol. 59 (8), 743–752.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, pp. 590–597.

Jeong, J., Jeoun, B., Park, Y., Han, B., 2023. An optimized ensemble framework for multi-label classification on long-tailed chest X-ray data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2739–2746.

Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Mark, R.G., Horng, S., 2019a. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data 6 (1), 317.

Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S., 2019b. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042.

Ju, L., Wang, X., Wang, L., Liu, T., Zhao, X., Drummond, T., Mahapatra, D., Ge, Z., 2021. 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. Springer, pp. 3–12.

Ju, L., Wu, Y., Wang, L., Yu, Z., Zhao, X., Wang, X., Bonnington, P., Ge, Z., 2022. Flexible sampling for long-tailed skin lesion classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 462–471.

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2022. Transformers in vision: A survey. ACM Comput. Surv. (CSUR) 54 (10s), 1–41.

Kim, D., 2023. CheXFusion: Effective fusion of multi-view features using transformers for long-tailed chest X-Ray classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2702–2710.

Kim, C., Kim, G., Yang, S., Kim, H., Lee, S., Cho, H., 2023. Chest X-Ray feature pyramid sum model with diseased area data augmentation method. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2757–2766.

Lehman, E., Johnson, A., 2023. Clinical-t5: Large language models built using mimic clinical text.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J., 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Adv. Neural Inf. Process. Syst. 36.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Mishra, A., Mittal, R., Jestin, C., Tingos, K., Rajpurkar, P., 2023. Improving zero-shot detection of low prevalence chest pathologies using domain pre-trained language models. arXiv preprint arXiv:2306.08000.

Moon, J.H., Lee, H., Shin, W., Kim, Y.-H., Choi, E., 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE J. Biomed. Health Inf. 26 (12), 6070–6080.

Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P., 2023. Med-flamingo: A multimodal medical few-shot learner. In: Machine Learning for Health. ML4H, PMLR, pp. 353–367.

Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al., 2022. Vindr-CXR: An open dataset of chest X-rays with radiologist's annotations. Sci. Data 9 (1), 429.

Nguyen-Mau, T.-H., Huynh, T.-L., Le, T.-D., Nguyen, H.-D., Tran, M.-T., 2023. Advanced augmentation and ensemble approaches for classifying long-tailed multi-label chest X-Rays. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2729–2738.

Park, W., Park, I., Kim, S., Ryu, J., 2023. Robust asymmetric loss for multi-label long-tailed learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2711–2720.

Pavao, A., Guyon, I., Letournel, A.-C., Tran, D.-T., Baro, X., Escalante, H.J., Escalera, S., Thomas, T., Xu, Z., 2023. CodaLab competitions: An open source platform to organize scientific challenges. J. Mach. Learn. Res. 24 (198), 1–6, URL http://jmlr.org/papers/v24/21-1436.html.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP, pp. 1532–1543.

Pinheiro, P.O., Collobert, R., 2015. From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1713–1721.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Reed, C.J., Yue, X., Nrusimha, A., Ebrahimi, S., Vijaykumar, V., Mao, R., Li, B., Zhang, S., Guillory, D., Metzger, S., et al., 2022. Self-supervised pretraining improves self-supervised pretraining. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2584–2594.

Rethmeier, N., Augenstein, I., 2022. Long-tail zero and few-shot learning via contrastive pretraining on and for small data. In: Computer Sciences & Mathematics Forum. Vol. 3, MDPI, p. 10.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L., 2021a. Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 82–91.

Ridnik, T., Lawen, H., Noy, A., Ben Baruch, E., Sharir, G., Friedman, I., 2021b. Tresnet: High performance gpu-dedicated architecture. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1400–1409.

Ridnik, T., Sharir, G., Ben-Cohen, A., Ben-Baruch, E., Noy, A., 2023. Ml-decoder: Scalable and versatile classification head. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 32–41.

Sabottke, C.F., Spieler, B.M., 2020. The effect of image resolution on deep learning in radiography. Radiol.: Artif. Intell. 2 (1), e190015.

Sakurada, S., Hang, N.T., Ishizuka, N., Toyota, E., Hung, L.D., Chuc, P.T., Lien, L.T., Thuong, P.H., Bich, P.T., Keicho, N., et al., 2012. Inter-rater agreement in the assessment of abnormal chest X-ray findings for tuberculosis between two Asian countries. BMC Infect. Dis. 12 (1), 1–8.

Seo, H., Lee, M., Cheong, W., Yoon, H., Kim, S., Kang, M., 2023. Enhancing multi-label long-tailed classification on chest X-Rays through ML-GCN augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2747–2756.

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I., Ghassemi, M., 2021. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. Vol. 26, pp. 232–243.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tan, M., Le, Q., 2021. Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning. PMLR, pp. 10096–10106.

Thambawita, V., Strümke, I., Hicks, S.A., Halvorsen, P., Parasa, S., Riegler, M.A., 2021. Impact of image resolution on deep learning performance in endoscopy image classification: An experimental study using a large dataset of endoscopic images. Diagnostics 11 (12), 2183.

Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P., 2022. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat. Biomed. Eng. 6 (12), 1399–1406.

Touvron, H., Vedaldi, A., Douze, M., Jégou, H., 2019. Fixing the train-test resolution discrepancy. Adv. Neural Inf. Process. Syst. 32.

Verma, A., 2023. How can we tame the long-tail of chest X-ray datasets? arXiv preprint arXiv:2309.04293.

Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S.X., 2020. Long-tailed recognition by routing diverse distribution-aware experts. arXiv preprint arXiv:2010.01809.

Wang, S., Lin, M., Ding, Y., Shih, G., Lu, Z., Peng, Y., 2022. Radiology text analysis system (RadText): Architecture and evaluation. In: 2022 IEEE 10th International Conference on Healthcare Informatics. ICHI, pp. 288–296. http://dx.doi.org/10.1109/ICHI54592.2022.00050.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106.

Wang, G., Wang, P., Cong, J., Liu, K., Wei, B., 2023. BB-GCN: A bi-modal bridged graph convolutional network for multi-label chest X-Ray recognition. arXiv preprint arXiv:2302.11082.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1492–1500.

Xie, Q., Luong, M.-T., Hovy, E., Le, Q.V., 2020. Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698.

Xu, M., Yoon, S., Fuentes, A., Park, D.S., 2023. A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognit. 137, 109347.

Yamagishi, Y., Hanaoka, S., 2023. Effect of stage training for long-tailed multi-label image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2721–2728.

Yan, B., Pei, M., 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 2982–2990.

Yang, Z., Pan, J., Yang, Y., Shi, X., Zhou, H.-Y., Zhang, Z., Bian, C., 2022. Proco: Prototype-aware contrastive learning for long-tailed medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 173–182.

Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032.

Zhang, Y., Cheng, Y., Huang, X., Wen, F., Feng, R., Li, Y., Guo, Y., 2021b. Simple and robust loss design for multi-label learning with missing labels. arXiv preprint arXiv:2112.07368.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. Mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

Zhang, R., Haihong, E., Yuan, L., He, J., Zhang, H., Zhang, S., Wang, Y., Song, M., Wang, L., 2021a. MBNM: Multi-branch network based on memory features for long-tailed medical image recognition. Comput. Methods Programs Biomed. 212, 106448.

Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J., 2023b. Deep long-tailed learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell..

Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y., 2023a. Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Commun. 14 (1), 4542.

Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M., 2021. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE 109 (5), 820–838.

Zhu, K., Wu, J., 2021. Residual attention: A simple but effective method for multi-label recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 184–193.