



CXR-LT 2024: A MICCAI challenge on long-tailed, multi-label, and zero-shot disease classification from chest X-ray

Mingquan Lin^{1,2,1b,a}, Gregory Holste^{3,a}, Song Wang³, Yiliang Zhou¹, Yishu Wei¹, Imon Banerjee⁴, Pengyi Chen⁵, Tianjie Dai^{5,6}, Yuexi Du⁷, Nicha C. Dvornek^{7,8}, Yuyan Ge⁹, Zuwei Guo¹⁰, Shouhei Hanaoka¹¹, Dongkyun Kim¹², Pablo Messina^{13,14}, Yang Lu¹⁵, Denis Parra^{13,14}, Donghyun Son¹⁶, Álvaro Soto¹³, Aisha Urooj⁴, René Vidal¹⁷, Yosuke Yamagishi¹¹, Pingkun Yan^{18,*}, Zefan Yang¹⁸, Ruichi Zhang¹⁵, Yang Zhou¹⁹, Leo Anthony Celi^{20,21,22}, Ronald M. Summers²³, Zhiyong Lu²⁴, Hao Chen²⁵, Adam Flanders²⁶, George Shih²⁷, Zhangyang Wang^{3,*}, Yifan Peng^{11b,*}

¹ Department of Population Health Sciences, Weill Cornell Medicine, NY, USA

² Department of Surgery, University of Minnesota, Minneapolis, USA

³ Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, USA

⁴ Department of Radiology, Mayo Clinic, Phoenix, USA

⁵ Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, China

⁶ Shanghai AI Laboratory, Shanghai, China

⁷ Department of Biomedical Engineering, Yale University, New Haven, USA

⁸ Department of Radiology & Biomedical Imaging, Yale University, New Haven, USA

⁹ Center for Innovation in Data Engineering and Science, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, USA

¹⁰ School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, USA

¹¹ Department of Radiology, The University of Tokyo Hospital, Tokyo, Japan

¹² School of Computer Science, Carnegie Mellon University, Pittsburgh, USA

¹³ Pontificia Universidad Católica de Chile, Santiago, Chile

¹⁴ Millennium Institute for Intelligent Healthcare Engineering (iHEALTH), National Center for Artificial Intelligence (CENIA), Santiago, Chile

¹⁵ School of Informatics, Xiamen University, Xiamen, China

¹⁶ Seoul National University, Seoul, South Korea

¹⁷ Center for Innovation in Data Engineering and Science, Departments of Electrical and Systems Engineering & Radiology, University of Pennsylvania, Philadelphia, USA

¹⁸ Department of Biomedical Engineering and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, USA

¹⁹ Institute of High Performance Computing, A*STAR, Singapore

²⁰ Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, USA

²¹ Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, USA

²² Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, USA

²³ Clinical Center, National Institutes of Health, Bethesda, USA

²⁴ Division of Intramural Research, National Library of Medicine, National Institutes of Health, Bethesda, USA, Bethesda, USA

²⁵ Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China

²⁶ Department of Radiology, Thomas Jefferson University Hospital, Philadelphia, USA

²⁷ Department of Radiology, Weill Cornell Medicine, NY, USA

ARTICLE INFO

Keywords:

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

ABSTRACT

The CXR-LT series is a community-driven initiative designed to enhance lung disease classification using chest X-rays (CXR). It tackles challenges in open long-tailed lung disease classification and enhances the measurability of state-of-the-art techniques. The first event, CXR-LT 2023, aimed to achieve these goals by providing high-quality benchmark CXR data for model development and conducting comprehensive evaluations to identify ongoing issues impacting lung disease classification performance. Building on the success of CXR-LT 2023, the **CXR-LT 2024** expands the dataset to 377,110 chest X-rays (CXRs) and 45 disease labels, including 19

* Corresponding authors.

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

^a Equal contribution.

new rare disease findings. It also introduces a new focus on zero-shot learning to address limitations identified in the previous event. Specifically, CXR-LT 2024 features three tasks: (i) long-tailed classification on a large, noisy test set, (ii) long-tailed classification on a manually annotated “gold standard” subset, and (iii) zero-shot generalization to five previously unseen disease findings. This paper provides an overview of CXR-LT 2024, detailing the data curation process and consolidating state-of-the-art solutions, including the use of multimodal models for rare disease detection, advanced generative approaches to handle noisy labels, and zero-shot learning strategies for unseen diseases. Additionally, the expanded dataset enhances disease coverage to better represent real-world clinical settings, offering a valuable resource for future research. By synthesizing the insights and innovations of participating teams, we aim to advance the development of clinically realistic and generalizable diagnostic models for chest radiography.

1. Introduction

The CXR-LT series marks a community-driven initiative to improve lung disease classification using chest X-rays (CXR) that addresses challenges in open long-tailed lung disease classification and advances the measurability of state-of-the-art techniques (Holste et al., 2022). These goals were pursued during the first event, CXR-LT 2023 (Holste et al., 2024), by offering high-quality benchmark CXR data for model development and conducting detailed evaluations to identify persistent issues affecting lung disease classification performance. CXR-LT 2023 attracted significant attention, with 59 teams yielding over 500 unique submissions. Since then, the task setup and data have provided a foundation for numerous studies (Hong et al., 2024; Huijben et al., 2024; Park and Ryu, 2024; Li et al., 2024a).

As the second event in the series, CXR-LT 2024 maintains the general design and goals of its predecessor while introducing a new emphasis on zero-shot learning. This addition addresses a limitation identified in CXR-LT 2023. The vast number of unique radiological findings, estimated to exceed 4500 (Budovec et al., 2014),¹ suggests that the actual distribution of clinical findings on CXR is at least two orders of magnitude greater than what current benchmarks can offer. Therefore, effectively addressing the “long-tail” of radiological abnormal findings necessitates the development of a model that can generalize to new classes in a “zero-shot” manner.

This paper provides an overview of the CXR-LT 2024 challenge, including two long-tailed tasks that attracted extensive participation and one newly introduced zero-shot task. Task 1 and Task 2 focus on long-tailed classification, with Task 1 using a large, noisy test set and Task 2 using a small, manually annotated test set. Task 3 concerns zero-shot generalization to previously unseen diseases. Each task adheres to the general framework established by CXR-LT 2023, providing participants with a large, automatically labeled training set consisting of over 250,000 CXR images with 40 binary disease labels. The final submissions from participants are evaluated against a separate held-out test set prepared in a similar manner.

In the following sections, we introduce each task setting and outline the evaluation criteria. Next, we detail the data curation process before presenting the results for each task. We then consolidate key insights from top-performing solutions and provide practical perspectives. Finally, we use our findings to suggest a path forward for few- and zero-shot disease classification, emphasizing the potential of leveraging multimodal foundation models.

2. Methods

2.1. Main tasks

The CXR-LT 2024 challenge includes three tasks: (1) long-tailed classification on a large, noisy test set, (2) long-tailed classification on a small, manually annotated test set, and (3) zero-shot generalization

to previously unseen diseases. All can be formulated as multi-label classification problems.

Given the severe label imbalance in these tasks, the primary evaluation metric was mean average precision (mAP), specifically the “macro-averaged” AP across classes. While the area under the receiver operating characteristic curve (AUROC) is often used for similar datasets (Wang et al., 2017; Seyyed-Kalantari et al., 2020), it can be heavily inflated in the presence of class imbalance (Fernández et al., 2018; Davis and Goadrich, 2006). In contrast, mAP is more suitable for long-tailed, multi-label settings as it measures the performance across decision thresholds without degrading under-class imbalance (Rethmeier and Augenstein, 2022). For thoroughness, mean AUROC (mAUC) and mean F1 score (mF1) – with a threshold of 0.5 – were computed as auxiliary classification metrics. We also calculated the mean expected calibration error (ECE) (Naeini et al., 2015) to quantify bias. To further enhance clinical interpretability, we also report per-class F1 scores, as well as macro- and micro-averaged F1 scores and false-negative rates for critical findings, in addition to the challenge’s primary evaluation metric. We believe these additions provide a more granular understanding of model performance in practical settings.

2.2. Dataset curation

Table 1 lists the characteristics of the datasets used in the three tasks. The same training dataset is utilized for all three tasks. Similarly, the same development dataset is shared across all three tasks, with Task 3 focusing on five unseen classes. Task 1 and Task 3 share the same test set; however, Task 3 explicitly evaluates performance on the five unseen classes. Task 2 is a subset of Task 1, containing 26 manually annotated classes. Fig. 1 lists the 45 classes, with the five unseen classes being Bulla, Cardiomyopathy, Hilum, Osteopenia, and Scoliosis. The remaining 40 classes exclude these five unseen classes, while the 14 classes are derived from the original MIMIC-CXR dataset and the 12 additional classes introduced in CXR-LT 2023. Fig. 2 shows example chest X-rays from the challenge dataset, where each image contains multiple annotated abnormalities.

In this section, we detail the data curation process of two datasets: (i) the automatically labeled CXR-LT dataset used in Tasks 1 and 3, and (ii) a manually annotated “gold standard” test set used in Task 2.

2.2.1. CXR-LT dataset

The CXR-LT challenge dataset² was developed by expanding the label set of the MIMIC-CXR dataset,³ (Johnson et al., 2019a) resulting in a more complex, long-tailed label distribution. This year, newly added clinical findings were selected from sources including the disease list of the PadChest dataset (Bustos et al., 2020) and the Fleischner glossary of thoracic imaging terms (Hansell et al., 2008). After ensuring that a sufficient number of occurrences were observed in the dataset for reliable evaluation, these 19 new disease findings are:

(1) Adenopathy, (2) Azygos Lobe, (3) Clavicle Fracture, (4) Fissure, (5) Hydropneumothorax, (6) Infarction, (7) Kyphosis, (8) Lobar Atelectasis, (9) Pleural Other, (10) Pulmonary Embolism, (11) Pulmonary

¹ <http://www.gamuts.net/about.php>

² <https://physionet.org/content/cxr-lt-iccv-workshop-cvamd>

³ <https://physionet.org/content/mimic-cxr/2.0.0/>

Table 1
Characteristics of the datasets used in the three tasks.

Dataset	Task 1		Task 2		Task 3	
	Samples	Labels	Samples	Labels	Samples	Labels
Train	258,871	40	258,871	40	258,871	40
Development	39,293	40	39,293	40	39,293	5
Test	78,946	40	406	26	78,946	5

Hypertension, (12) Rib Fracture, (13) Round Atelectasis, (14) Tuberculosis, (15) Bulla, (16) Cardiomyopathy, (17) Hilum, (18) Osteopenia, (19) Scoliosis.

The last five abnormal findings – Bulla, Cardiomyopathy, Hilum, Osteopenia, and Scoliosis – were not included in the challenge training set and were held out for zero-shot evaluation in Task 3.

In addition, we replaced the “No Finding” class with a more intuitive “Normal” class. “No Finding” indicated that none of the abnormal findings in the label set were present. For instance, with the original 14 MIMIC-CXR classes, “No Finding” meant that none of these 14 findings were present; however, when expanding the label set to 26 classes, this “No Finding” label may, in fact, include one of the 12 added abnormalities. To avoid unclear interpretation of this label across tasks, we curated new labels for a simplified “Normal” class, signifying that no cardiopulmonary disease or abnormality was found in the report. Like in 2023, the radiology reports for each CXR study were parsed using RadText (Wang et al., 2022), a radiology text analysis tool, to extract the presence status of new diseases.

The final dataset included 377,110 CXR images, each labeled with one or more of the 45 diseases, following a long-tailed distribution (Fig. 1). Like CXR-LT 2023, we opted to use image data from the MIMIC-CXR-JPG dataset (Johnson et al., 2019b)⁴ to increase accessibility and reduce the burden of storing this dataset (~600 GB vs. ~4.7 TB using the raw DICOM data provided by MIMIC-CXR). The dataset was randomly partitioned into training (70%), development (10%), and test (20%) sets at the patient level; critically, this split was unique to CXR-LT 2024, meaning participants could not re-use models from the previous year’s challenge. Participants had access to all images, but were provided with labels only for the training set.

2.2.2. Gold standard test set

In our overview of CXR-LT 2023 (Holste et al., 2024), we used a manually annotated “gold standard” set derived from the challenge test set to evaluate the differences in manual vs. automated annotation as well as how top-performing solutions fared on this test set with reduced label noise. Specifically, six annotators reviewed 406 MIMIC-CXR radiology reports for the presence or absence of 26 disease findings considered in CXR-LT 2023. For complete data curation details of this gold standard set, please see Holste et al. (2024). This dataset provides a high-quality benchmark for evaluating model performance on a smaller, manually vetted test set. This year, in CXR-LT 2024, we used this gold standard dataset as the test set in Task 2.

2.3. Schedule

Table 2 shows the task schedule. The challenge was conducted on the CodaLab platform⁵, with a separate CodaLab page for each of the three tasks (Pavao et al., 2023). Any registered CodaLab user could apply, but would only be accepted after submitting proof of the necessary PhysioNet credentials required to access MIMIC-CXR-JPG.

Table 2
Schedule of CXR-LT 2024.

Event	Date	Teams
Registration	May 1, 2024	61
Development phase	May 1, 2024	29
Training data		
Development data		
Leaderboard		
Test phase		17
Test data	Aug 26, 2024	
Submission	Sept 6, 2024	
Workshop	Oct 10, 2024	9

During the Development Phase (May 1, 2024 - August 26, 2024), registered participants downloaded the labeled training set and the unlabeled development set, from which they generated a comma-separated values (CSV) file with predictions to upload. Submissions were evaluated on the held-out development set, and results were updated on a live, public leaderboard. During the Test Phase (August 26, 2024 - September 6, 2024), test set images (without labels) were released. Participants were asked to submit CSV files with test set predictions for final evaluation and ranking in each task. The leaderboard was hidden in this phase, and each team’s best-scoring submission was retained. The final Test Phase leaderboard ranked participants primarily by mAP, then by mAUC in the event of ties.

3. Results

3.1. Participation

The CXR-LT challenge received 96 team applications on CodaLab, of which 61 were approved after providing proof of credentialed access to MIMIC-CXR-JPG (Johnson et al., 2019b). During the Development Phase, 29 teams participated, submitting a total of 661, 349, and 364 unique submissions to the public leaderboard for Tasks 1, 2, and 3, respectively. In the final Test Phase, a total of 17 teams participated. We selected the top 9 teams for the invitation to present at the CXR-LT 2024 challenge event at MICCAI 2024⁶ and for inclusion in this study. Since two teams excelled in both Tasks 1 and 2, this comprised the top 4 solutions in Tasks 1 and 2 as well as the top 3 solutions for the zero-shot Task 3. Table 3 summarizes the top-performing groups participating in one or more of these tasks and system descriptions. Additional details, including all presentation slides, are available on GitHub,⁷ allowing readers to explore the specifics of all methods in greater depth.

3.2. System descriptions

Team A: zguo. This team proposed the ChexFusion+ model for CXR-LT classification, participating in Task 1 and Task 2. Their approach leveraged ensembles of 12 multi-resolution ConvNeXt (Liu et al., 2022) models trained with synthetically generated long-tail data. Specifically, they used input prompts and random Gaussian noise to generate two images through a conditional denoising U-Net and a variational autoencoder decoder (VAE) decoder. The two generated images, together with the input prompts and the corresponding MIMIC-CXR image, were used as input for the ConvNeXt for training. To address the data imbalance problem in the tail cases, they used pre-trained large generative models to generate 100 new images per rare disease class. These synthetic images were generated using carefully constructed prompts that specified multiple co-occurring thoracic abnormalities, such as “Round(ed) Atelectasis, Pneumothorax, Pleural Effusion, Lung Opacity, and Atelectasis”, to reflect realistic radiographic comorbidities.

⁴ <https://physionet.org/content/mimic-cxr-jpg/2.0.0/>

⁵ <https://codalab.lisn.upsaclay.fr/competitions/18601>, <https://codalab.lisn.upsaclay.fr/competitions/18603>, <https://codalab.lisn.upsaclay.fr/competitions/18604>

⁶ <https://cxr-lt.github.io/CXR-LT-2024/>

⁷ <https://github.com/CXR-LT/CXR-LT-2024>

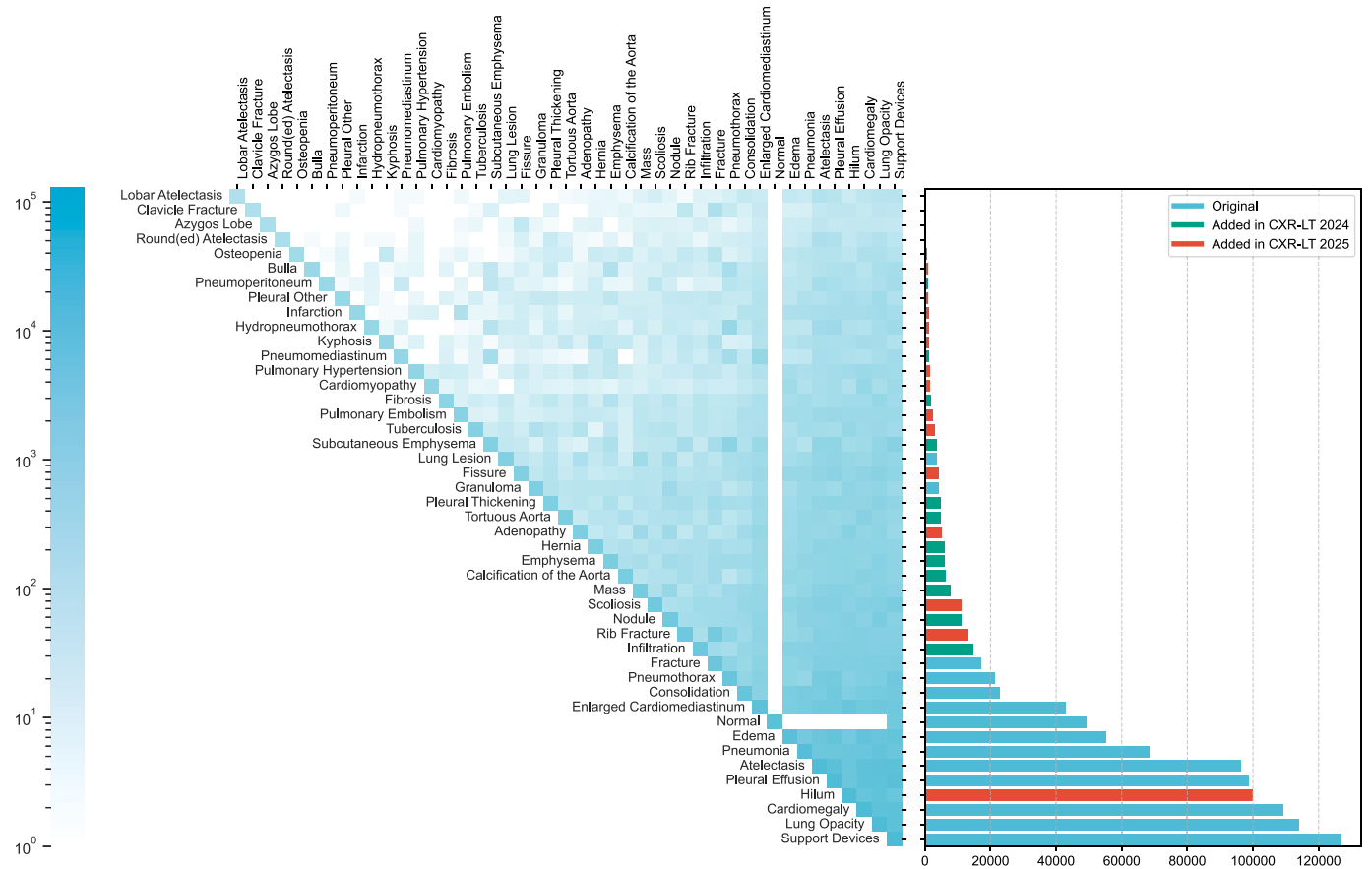
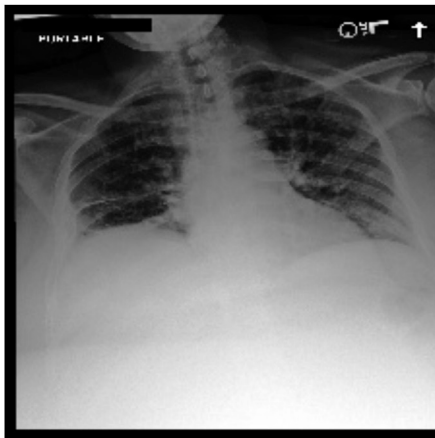
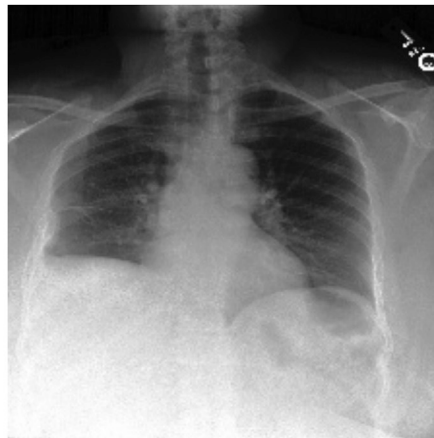


Fig. 1. Long-tailed distribution of the CXR-LT 2024 challenge dataset. The dataset was formed by extending the MIMIC-CXR benchmark to include 12 new clinical findings (red) by parsing radiology reports.



(a) Hilum, Cardiomeidiastinum, Fissure, Nodule, Pleural Effusion



(b) Fracture, Pleural Effusion



(c) Cardiomegaly, Edema, Lung Opacity

Fig. 2. Representative chest X-rays from the challenge dataset, each demonstrating multiple findings. (a) Includes the Hilum label (new in CXR-LT 2024); (b) shows Fracture (introduced in CXR-LT 2023); and (c) displays original MIMIC-CXR labels (Cardiomegaly, Edema, Lung Opacity).

Team B: tianjie_dai. This team leveraged a multimodal ensemble approach to address the imbalanced, multi-label classification challenge in the CXR-LT tasks. Specifically, they employed an ensemble of EfficientNetV2-Large (Tan and Le, 2021) and PubMedBERT (Gu et al., 2020) models, fine-tuned on a Unified Language Medical System (UMLS) knowledge graph (Dai et al., 2024), to integrate both image and text features. To address class imbalance, they used an asymmetric loss function (Kim, 2023), assigning higher weights to rare classes.

Test-time augmentation techniques were applied to improve model robustness and generalization, including resizing, cropping, and flipping. Additionally, they incorporated external datasets from multiple sources, such as ChestXRay-14 (Wang et al., 2017), CheXpert (Irvin et al., 2019), VinDr-CXR (Nguyen et al., 2022), and BRAX (Reis et al., 2022), to enhance representation learning of rare disease labels.

Team C: XYPB. This team addressed the CXR-LT challenge with a multimodal ensemble approach, leveraging multi-view and multi-scale

Table 3

Overview of top-performing CXR-LT 2024 challenge solutions. ENS - ensemble; LRW - loss reweighting; VL - vision-language.

Team	Institution	Image Resolution	Backbone	ENS	LRW	VL	Pre-training
A	Arizona State University	224, 384	ConvNeXt-S ConvNeXt-B ConvNeXt-T ConvNeXt V2-B	✓		✓	ImageNet
B	Shanghai Jiaotong University	512	EfficientNetV2-L	✓	✓	✓	ImageNet → NIH, CheXpert, VinDr-CXR, BRAX
C	Yale University	1024	ConvNeXt-S EfficientNetV2-S	✓	✓	✓	ImageNet → MIMIC-CXR
D	Carnegie Mellon University	1024	ConvNeXt-S		✓	✓	ImageNet → CheXpert, NIH, VinDr-CXR
E	Rensselaer Polytechnic Institute	336, 448, 512	ViT-L	✓			MIMIC-CXR, CheXpert, PadChest, NIH, BRAX
F	The University of Tokyo	384, 512	ConvNeXt V2-S MaxViT-T	✓	✓		ImageNet → NIH
G	University of Pennsylvania	224	ViT-L	✓	✓	✓	ImageNet → MIMIC-CXR, CXR-Concepts, Chest ImaGenome, CXR-LT
H	Xiamen University	224	ResNet50	✓		✓	None
I	Pontifical Catholic University of Chile	384, 416	DenseNet121 SigLIP Base ConvNeXt-S Uniformer	✓	✓	✓	ImageNet → MIMIC-CXR, IU X-ray, Chest ImaGenome, CheXpert, CheXlocalize, VinDr-CXR

image alignment to enhance classification in the long-tailed, multi-label setting. Their method builds on the CLEFT (Du et al., 2024a) and MaMA (Du et al., 2024b) frameworks, incorporating contrastive language-image pretraining (CLIP) with additional image-to-image and image-to-text contrastive learning across multi-view image pairs from chest X-ray studies. This approach allowed their model to capture information from varied perspectives, such as postero-anterior and lateral views. To tackle the multi-scale nature of medical imaging, they introduced a Symmetric Local Cross-Attention (SLA) alignment module that models region-specific visual-text correlations through cross-attention, aligning local image patches with descriptive text segments. To counteract class imbalance, they used a weighted asymmetric loss (Ridnik et al., 2021). For the image encoder, they utilized ConvNeXt-S (Liu et al., 2022) and EfficientNet V2-S (Tan and Le, 2021) backbones, pre-trained on ImageNet for classification and MIMIC-CXR using a CLIP-based approach, for Task 1 and Task 2, and use parameter efficient fine-tuned medical LLM, *i.e.*, BioMedLM (Bolton et al., 2024), as their language encoder.

Team D: dongkyunk. This team implemented a two-stage framework designed to effectively leverage the multiple views available for each patient. In the first stage, a single model was trained using the ML-Decoder (Ridnik et al., 2023) classification head alongside Noisy Student (Xie et al., 2020) self-training. In the second stage, a Transformer-based model called CheXFusion was introduced to aggregate multi-view features (Kim, 2023). The feature aggregation in CheXFusion is analogous to encoding multiple sentences in natural language processing, where each sentence represents a single chest X-ray image. Additionally, a weighted version of the asymmetric loss (Ridnik et al., 2021) was employed to address inter-class imbalance from the long-tailed distribution of diseases and intra-class imbalance due to the predominance of negative labels in multi-label classification.

Team E: yangz16. This team tackled the CXR-LT challenge with a foundation model-based approach. Their methodology incorporated three key components: DINOv2 foundation models (Oquab et al., 2023) utilizing the ViT-Large network architecture (Neil and Dirk, 2020) as the backbone, the ML-Decoder (Neil and Dirk, 2020) classification head, and multi-view/multi-resolution ensembling. The DINOv2 models were pre-trained on over 710,000 chest X-rays from diverse datasets, including MIMIC-CXR (Johnson et al., 2019a), CheXpert (Irvin et al.,

2019), PadChest (Bustos et al., 2020), NIH Chest X-ray14 (Wang et al., 2017), and BRAX (Reis et al., 2022), using self-supervised learning that combined a self-distillation loss and masked image modeling to learn robust representations. The ML-Decoder mapped local features from the foundation model to disease-specific predictions for classification, employing attention mechanisms to achieve finer localization of disease findings.

Team F: yyama. This team utilized an ensemble of ConvNeXt V2 (Liu et al., 2022) and MaxViT (Tu et al., 2022) models with domain-specific pretraining and view-based aggregation. The ConvNeXt V2 models were pre-trained on ImageNet, while the MaxViT model was further pretrained on the NIH Chest X-ray dataset. To address class imbalance, they applied an asymmetric loss function (Ridnik et al., 2021) combined with class weights, assigning higher importance to rare classes. Additionally, they implemented a view-based prediction aggregation method, combining predictions from frontal and lateral views, with a weighted average favoring the frontal view.

Team G: yyge. This team employed a dual-model strategy combining Vision-Language Models (VLM) and Multi-View Vision Models (MVM) for zero-shot and multi-label disease classification in chest X-rays. The VLM integrated DINOv2 (Oquab et al., 2023) as the image encoder and BERT (Boecking et al., 2022) for text encoding, utilizing fine-grained disease descriptions from ChatGPT (Achiam et al., 2023). The model was first pretrained on domain-specific datasets, and then was fine-tuned on the CXR-LT training set with the weighted binary cross-entropy loss for class imbalance. Meanwhile, the MVM converted zero-shot tasks into few-shot problems by mining disease-specific examples from radiology reports and aggregated multi-view features using DINOv2 and lightweight Transformers. Using a weighted asymmetric loss (Kim, 2023) and multi-view learning further addressed long-tail distributions. This combined framework effectively captured domain-specific knowledge and balanced performance across seen and unseen diseases.

Team H: ZhangRuichi. This team utilized a Visual-Language Model (VLM) inspired by MedKLIP (Wu et al., 2023), incorporating anatomical and textual information to enhance generalization and performance. The authors used a ResNet50 (He et al., 2016) architecture without pretraining for image encoding, as pretraining on ImageNet may introduce biases due to domain and distribution differences between natural

Table 4

mAP of top-4 team's final model on all 40 classes evaluated on the test set in Task 1. mAUROC, mF1, and mECE are also presented, with the numbers in parentheses indicating the rankings based on the corresponding evaluation metric.

Ranking	Team	mAP	mAUROC		mF1		mECE	
				🏆		🏆		🏆
1	A	0.281	0.847	(3)	0.289	(3)	0.589	(4)
2	B	0.279	0.843	(5)	0.286	(4)	0.592	(6)
3	C	0.277	0.849	(1)	0.299	(1)	0.603	(8)
4	D	0.277	0.842	(6)	0.285	(5)	0.602	(7)

images and chest X-ray datasets. By training the model from scratch, the authors aim to mitigate these domain and distribution gaps and better tailor the model to the characteristics of chest X-ray images. To enable zero-shot capability, categorical labels were augmented with GPT-4-generated descriptions, and BioClinical BERT (Alsentzer et al., 2019) was employed for text encoding, capturing rich semantic information. Anatomical location data from CheXpert was integrated, mapping MIMIC-CXR diseases to corresponding regions to address the lack of fine-grained labels. For training, they applied a cross-entropy loss to improve accuracy and a contrastive loss to link diseases with anatomical regions. During testing, a class-wise ensemble strategy and test-time fusion of predictions from different patient views were implemented to improve overall accuracy.

Team I: pameSSina. This team developed a multimodal model for the zero-shot classification task. The text encoder is the CXR Fact Encoder (CXRFE) (Messina et al., 2024), which computes fact embeddings from short factual sentences. The text encoder remains frozen during training. The image encoder is trained end-to-end, and the team experimented with several architectures, including DenseNet121 (Huang et al., 2017), SigLIP Base (Zhai et al., 2023), ConvNext-Small (Liu et al., 2022), and Uniformer (Li et al., 2022). The model uses the fact embeddings to modulate local and global features from the image encoder via FiLM (Perez et al., 2018) to predict both a binary coarse segmentation mask (representing the visual grounding of the fact) and a global binary classification of the fact for the entire image. The best results were achieved using an ensemble of 21 models, each with a different configuration of image encoder and training data. This work also used GPT-4 (Achiam et al., 2023) as an automatic labeler for MIMIC-CXR reports and included additional datasets, some of which contained bounding boxes to provide visual grounding supervision.

3.3. Task 1 primary evaluation results

CXR-LT test phase results. The results of the top 4 teams in Task 1 are listed in Table 4. Team A took first place with an mAP of 0.281, Team B came in second with an mAP of 0.279, while Teams C and D both achieved an mAP of 0.277. However, Team C ranked third due to a higher mAUC. In CXR-LT 2023, the top 4 teams achieved mAP scores ranging from 0.349 to 0.372, much higher than this year's scores due to the inclusion of 19 new rare classes in CXR-LT 2024. When evaluating this year's top solutions on the set of 26 CXR-LT 2023 labels, we observe mAP scores of 0.371, 0.373, 0.371, and 0.370, respectively. Compared to CXR-LT 2023, top performers in Task 1 improved their overall performance in these classes (e.g., the second-place finisher in CXR-LT 2023 reached 0.354 mAP). Supplementary Table 1 details the performance of each class for the top four teams. Additionally, Supplementary Tables 2 and 3 present per-class F1 scores, macro- and micro-averaged F1 scores, and false-negative rates for critical findings.

Long-tailed classification performance. To examine predictive performance by label frequency, we split the 40 target classes into “head” (>10%), “medium” (1%–10%), and “tail” (<1%) categories based on their prevalence in the training set, consisting of 9, 14, and 16 categories, respectively. Category-wise mAP is presented in Table 5, as well as a “category-wise average” of head, medium, and tail mAP. Team

Table 5

Long-tailed classification performance on “head”, “medium”, and “tail” classes by average mAP within each category. These categories were determined by the 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.

Team	Overall	Head (>10%)	Medium (1%–10%)	Tail (<1%)	Avg
A	0.281	0.567	0.263	0.136	0.322
B	0.279	0.569	0.260	0.133	0.321
C	0.277	0.570	0.253	0.136	0.320
D	0.277	0.568	0.264	0.125	0.320

A not only achieved the highest overall performance but also excelled in the “tail” group. However, the top three performances in the “tail” group were very close. Team A used pretrained large generative models to generate new images for these tail cases, while Teams B and C applied loss reweighting, indicating that both approaches can improve performance in the “tail” group.

3.4. Task 2 primary evaluation results

The results of the top 4 teams in Task 2 are listed in Table 6. The first place went to Team C with an mAP of 0.526. Team E placed second with an mAP of 0.511, Team A secured third with an mAP of 0.511, and Team F placed fourth with an mAP of 0.509. All four teams used ensembling methods to improve model performance, achieving results similar to last year; for instance, top CXR-LT 2023 performers achieved mAP scores of 0.519, 0.518, and 0.519 on the same gold standard test set. Supplementary Table 4 provides the detailed class-specific performance for these top teams. Additionally, Supplementary Tables 5 and 6 present per-class F1 scores, macro- and micro-averaged F1 scores, and false-negative rates for critical findings.

As mentioned in Section 2.2.2, the test set in Task 2 is the subset of the test set in Task 1, with only 26 manually annotated labels. Table 4 and 6 show that the first-place team in Task 2 ranked third in Task 1, while the second-place team in Task 2 was first in Task 1. Additionally, from the challenge leaderboards,⁸⁹ we can see that the third- and fourth-place teams in Task 2 ranked sixth and fifth in Task 1, with mAP scores of 0.269 and 0.273, respectively. We selected ten teams that submitted their results for both tasks, named them T1 to T10, and analyzed their performance based on the results. Despite a large distribution shift between these datasets, the overall performance consistency remained stable between Tasks 1 and 2 (Fig. 3; $R^2 = 0.946$, $r = 0.972$).

3.5. Task 3 primary evaluation results

Table 7 presents results for the top 3 performing teams in Task 3. Team G secured the first place with an mAP of 0.129. Following closely, Team H earned second place with an mAP of 0.116, while

⁸ <https://codalab.lisn.upsaclay.fr/competitions/18603#results>

⁹ <https://codalab.lisn.upsaclay.fr/competitions/18601#results>

Table 6

mAP of top-4 team's final model on all 26 classes evaluated on the Gold standard test set in Task 2. mAUROC, mF1, and mECE are also presented, with the numbers in parentheses indicating the rankings based on the corresponding evaluation metric.







Ranking	Team	mAP	mAUROC		mF1		mECE	
								
1	C	0.526	0.833	(3)	0.499	(1)	0.464	(6)
2	A	0.519	0.834	(2)	0.471	(4)	0.457	(30)
3	E	0.511	0.836	(1)	0.265	(9)	0.744	(10)
4	F	0.509	0.829	(5)	0.474	(3)	0.462	(5)

Table 7

Performance evaluation of the final models from the top 3 teams on the test set for all five unseen classes in Task 3. mAUROC, mF1, and mECE are also presented, with the numbers in parentheses indicating the rankings based on the corresponding evaluation metric.

Ranking	Team	mAP	mAUROC		mF1		mECE	
								
1	G	0.129	0.741	(2)	0.075	(6)	0.817	(8)
2	H	0.116	0.673	(8)	0.035	(7)	0.907	(9)
3	I	0.110	0.744	(1)	0.094	(4)	0.711	(6)

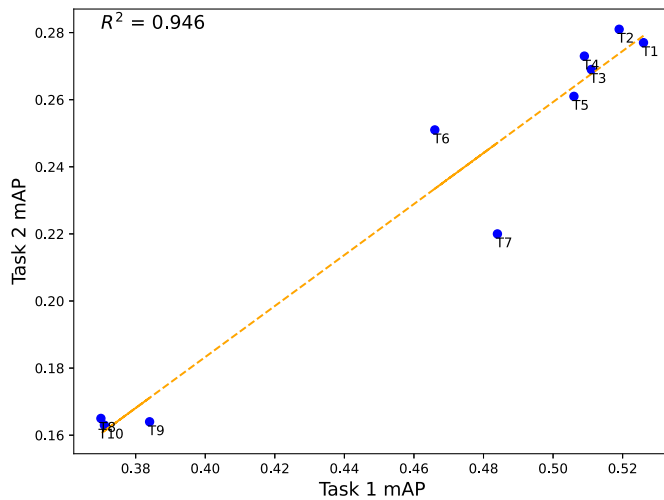


Fig. 3. Comparison of performance on CXR-LT Task 1 data (Section 2.2.1) and gold standard Task 2 data (Section 2.2.2).

Team I came in third with an mAP of 0.110. Compared to the other tasks, the relatively low performance in Task 3 can be attributed to the challenging zero-shot nature of detecting findings that were never seen during training. Supplementary Table 7 details the class-specific performance of these top teams. Additionally, Supplementary Table 8 and 9 report per-class F1 scores, macro- and micro-averaged F1 scores, and false-negative rates for critical findings.

3.6. Comparison of rule-based and GPT-4o labeling

In the CXR-LT dataset, labels were initially generated using a rule-based method. Approaches like this have proven successful in automatically labeling existing CXR datasets (Irvin et al., 2019; Wang et al., 2017; Bustos et al., 2020), but recent developments in large language models (LLMs) suggest they may be useful candidates for this task. Team I opted to use GPT-4 to label the data from MIMIC-CXR reports, instead of relying on the rule-based labels. With a dataset of 406 manually annotated samples, we calculated the precision to assess whether large language models could generate more accurate labels. Table 8 lists the performance comparison between the rule-based method and GPT-4o, using prompts suggested by Wei et al. (2024). The rule-based method achieved a precision of 0.711, whereas GPT-4o reached a precision of 0.786.

Table 8

Performance of models on long-tailed, multi-label disease classification evaluated using micro-precision on our gold standard test set.

	Rule-based	GPT-4o
Atelectasis	0.611	0.590
Calcification of the Aorta	1.000	0.857
Cardiomegaly	0.769	0.938
Consolidation	0.816	0.849
Edema	0.638	0.804
Emphysema	0.609	0.639
Enlarged Cardiomediastinum	0.583	1.000
Fibrosis	0.667	0.682
Fracture	0.870	0.937
Hernia	0.633	0.810
Infiltration	0.261	0.889
Lung Lesion	0.161	0.000
Lung Opacity	0.853	0.984
Mass	0.513	0.810
Normal	0.917	0.972
Nodule	0.821	0.844
Pleural Effusion	0.798	0.812
Pleural Other	0.810	0.500
Pleural Thickening	1.000	0.815
Pneumomediastinum	0.875	0.889
Pneumonia	0.191	0.435
Pneumoperitoneum	0.676	0.840
Pneumothorax	0.563	0.865
Subcutaneous Emphysema	0.955	0.889
Support Devices	0.948	0.933
Tortuous Aorta	0.958	0.861
Mean	0.711	0.786

4. Discussion

4.1. Themes of top CXR-LT 2024 solutions

As outlined in Table 3 the system descriptions, we observe several common themes among top-performing solutions across the three tasks, as well as some unique perspectives.

Modern convolutional neural network architectures. The top-performing solutions commonly used convolutional neural networks (CNNs) as image encoders, continuing the trend from CXR-LT 2023. ConvNeXt emerged as the most popular choice (Liu et al., 2022), followed by EfficientNet (Tan and Le, 2021), ResNet (He et al., 2016), and DenseNet (Huang et al., 2017). ConvNeXt consistently outperformed other architectures, both in 2023 and 2024. We attribute ConvNeXt's popularity and strong performance to two main factors: (1) its adoption by last year's top-performing solutions, which sets a standard, and (2) its

design for scalable performance across different image resolutions, making it well-suited to capture multi-scale information.

Vision transformers. While none of the top-performing solutions used Vision Transformers (ViTs) (Khan et al., 2022) as image encoders in 2023, there has been a noticeable shift in 2024, with four teams adopting ViT-based models. Specifically, Teams E and G dedicated their image encoding entirely to ViTs. In contrast, Teams F and I opted for a hybrid approach, combining ViT-based Transformers with CNNs. We attribute the increased adoption of ViT-based transformers to two main factors: (1) the complementary strengths offered by integrating both CNNs and ViTs (Pantelaios et al., 2024), which can enhance feature extraction and representation capabilities, and (2) the strategic use of additional datasets for pretraining the image encoders, which is particularly beneficial for effectively training ViT-based transformers, improving their robustness and generalization.

Large-scale pretraining. Eight of the nine top-performing solutions relied on supervised pretraining or transfer learning. While some teams used standard ImageNet-pretrained models, several others performed additional pretraining on publicly available “in-domain” CXR datasets such as ChestXRay-14 (Wang et al., 2017), CheXpert (Irvin et al., 2019), VinDr-CXR (Nguyen et al., 2022), and BRAX (Reis et al., 2022). Notably, Teams B, C, D, F, and G employed a multi-stage pretraining strategy, starting with general pretraining on natural images, followed by domain-specific pretraining on CXR data, similar to approaches used by several teams in CXR-LT 2023. In contrast, Team H reported that their proposed model achieved superior performance without pretraining.

Ensemble learning and data augmentation. As with CXR-LT 2023, many top solutions (eight out of nine) employed a variety of ensemble learning strategies to improve generalization (Ganaie et al., 2022; Fort et al., 2019). Teams B, C, F, G, and I created ensembles across different model architectures; Teams A and E formed ensembles by using different image resolutions; and Team H constructed an ensemble strategy based on multiple views of the same image, but using the same model architecture across these views. In addition to ensemble learning, all teams incorporated image augmentation, a well-established technique for enhancing generalization (Xu et al., 2023). Notably, Team A leveraged a diffusion model to generate synthetic images to augment rare tail classes.

Loss re-weighting. To address the long-tailed distribution of labels, five out of the nine top-performing solutions adopted loss re-weighting techniques to boost the importance of rare tail classes. These five teams (B, C, D, F, and G) all utilized a weighted asymmetric loss (Ridnik et al., 2021), which is specifically designed for handling imbalanced multi-label classification scenarios. Additionally, Team G implemented a weighted binary cross-entropy loss alongside this approach. The widespread adoption of the weighted asymmetric loss function can be attributed to its success in CXR-LT 2023, where it was employed by top-ranking teams. It is worth noting that two of the top solutions in Task 3 opted not to use weighted losses, primarily due to a lack of information about the distribution of the five unseen classes. However, one of the top solutions in Task 3 adopted a weighted loss approach by converting the zero-shot problem into a few-shot problem, leveraging prior knowledge about the unseen classes extracted from text-based descriptions.

Multimodal vision-language learning. Multimodal vision-language learning has recently gained popularity in deep learning for radiology, particularly as a pretraining approach using paired CXR images and free-text radiology reports (Chen et al., 2019; Yan and Pei, 2022; Delbrouck et al., 2022; Moon et al., 2022; Li et al., 2024b; Moor et al., 2023). This year, eight of the nine teams successfully leveraged both image and text data in some form. For example, Team C employed a combination of image-to-image and text-to-image contrastive learning to enhance feature representation. Meanwhile, Team D employed the

ML-Decoder (Ridnik et al., 2023) classification head, which treats labels as text “queries” that interact with image features via cross-attention. Notably, all three teams in Task 3 utilized multimodal vision-language models to enable zero-shot generalization to the five novel classes in Task 3.

Chatgpt/GPT-4. With LLMs’ increasing popularity in general and medical domains, three teams utilized ChatGPT or GPT-4 for Task 3. Team G used ChatGPT (Achiam et al., 2023) to create fine-grained disease descriptions, potentially enhancing the performance of the text encoder. Team H leveraged GPT-4 to generate descriptive text augmentations for categorical labels, thereby facilitating zero-shot learning. In contrast, Team I opted not to use the provided labels and instead employed GPT-4 as an automatic labeler for MIMIC-CXR reports. They further incorporated additional CXR datasets, including some with bounding boxes, to support visual grounding supervision.

Implications of synthetic data for long-tailed classification. Team A’s use of generative models to create synthetic data for rare classes appears to be a promising approach, as reflected in their performance. Synthetic data has the potential to mitigate extreme class imbalance by supplementing underrepresented classes, especially in domains where real data collection is costly or slow. However, it also raises questions about data fidelity, domain shift, and overfitting. As synthetic data generation techniques (e.g., diffusion models, GANs) continue to evolve, their role in addressing long-tailed medical image classification warrants further investigation, particularly regarding trustworthiness, generalizability, and clinical utility.

4.2. Limitations and future work

A key limitation of this study is the reliance on the MIMIC-CXR dataset, which was collected at a single academic medical center in the United States. As a result, the data may reflect institution-specific patient demographics, disease prevalence, imaging protocols, and equipment characteristics. These factors could limit the generalizability of the models to other geographic regions, healthcare systems, or clinical workflows. Although several top-performing teams incorporated additional publicly available CXR datasets (e.g., CheXpert, PadChest, VinDr-CXR) during training to enhance robustness, the final evaluation and leaderboard rankings were based solely on MIMIC-CXR test data. To more rigorously assess model generalization and ensure broader clinical applicability, future editions of the challenge should incorporate external test sets drawn from diverse institutions and populations. This would facilitate a more comprehensive evaluation of cross-site transferability and reveal potential sources of dataset shift or subgroup-specific bias.

Additionally, addressing bias in the models is critical. Existing studies have shown that deep neural networks trained on single-institution CXR datasets often exhibit disparities in predictive performance linked to factors like race and sex (Seyyed-Kalantari et al., 2020; Lin et al., 2023). While (Seyyed-Kalantari et al., 2020) observed that training on larger, multi-institutional datasets could help mitigate these disparities, their work focused on binary classification tasks. To date, no research has specifically examined bias in long-tailed, multi-label, and zero-shot classification tasks. Future investigations could explore methods to tackle these challenges, ensuring that models are both fair and generalizable across diverse populations and settings through rigorous subgroup analysis and multi-site validation.

Similar to most publicly available CXR benchmark datasets, the CXR-LT dataset is constrained by inherent label noise resulting from automatically extracted text-mined labels (Abdalla and Fine, 2023). However, with the advancement of LLMs, recent studies (Wei et al., 2024) have demonstrated that GPT-4 can potentially generate more accurate labels for CXR datasets than traditional methods such as rule-based approaches. In our study, Table 8 supports this observation,

showing that GPT-4 produces higher-quality labels as evidenced by improved mAP. Other LLMs may also surpass traditional methods in label generation, presenting a promising avenue to reduce label noise in the future. Future iterations of CXR-LT may leverage LLM-based labeling pipelines to generate structured labels for arbitrarily large, long-tailed CXR datasets, which have proven successful in recent efforts (Zheng et al., 2024). Additionally, as more classes are included, the prompts for LLMs become longer, which may cause performance degradation by overwhelming the model and potentially leading to forgetting some classes. To mitigate this issue, reframing the labeling task as a natural language inference (NLI) problem and focusing the prompt on one class at a time can be an effective strategy. Moreover, incorporating techniques like chain-of-thought (CoT) prompting can further enhance performance by improving reasoning and response generation. Alternatively, a knowledge graph can be employed to separate the classes into different subgroups before applying LLM-based labeling, providing a structured and systematic approach to addressing this challenge.

Moreover, while it is challenging to obtain sufficient samples for deep learning training through manual annotation by radiologists due to the prohibitively high costs and time requirements (Zhou et al., 2021), providing a “gold standard” dataset for testing purposes remains feasible. In this work, we leveraged and publicly released such a dataset with more reliable labels. However, this dataset was annotated by graduate students reviewing the clinical report text. In the future, this dataset could benefit from consensus re-annotation by radiology residents or attendings to enhance its quality. Additionally, manually annotating an external dataset for validation purposes could further enhance the evaluation of proposed methods, providing more reliable and accurate performance benchmarks.

As outlined in the overview of CXR-LT 2023 (Holste et al., 2024), zero-shot classification can be the ideal approach for clinically viable long-tailed medical image recognition, enabling adaptation to any novel finding. This year, the top three teams in Task 3 all utilized vision-language models to tackle this challenge, highlighting their potential. However, there remains significant room for improvement in several areas: aligning image and text representations more effectively, extracting information about unseen classes from textual data, and accurately detecting abnormal regions in images. Furthermore, efficient fine-tuning of vision-language models or instruction tuning will be crucial in addressing the challenges associated with the zero-shot disease classification problem.

Despite recent methodological advances, mean Average Precision (mAP) and F1 scores for chest X-ray (CXR) disease classification remain relatively modest. This raises a critical question: Are these performance metrics clinically acceptable? For example, an mAP in the range of ~0.28–0.52 indicates performance well above random chance, yet likely falls short of the reliability required for autonomous clinical use. Several factors contribute to these limitations. First, extreme class imbalance – especially with rare findings – can skew performance and reduce sensitivity for less-represented diseases. Second, both training and evaluation datasets often contain label noise due to weak supervision or annotation inconsistencies. Third, chest X-ray as an imaging modality inherently lacks the resolution or contrast to clearly distinguish certain pathologies, particularly those with subtle or overlapping visual cues. Additionally, most models rely on thresholded probabilistic outputs, which can suffer from calibration issues, further affecting the reliability of decisions. To mitigate these challenges and improve real-world utility, future research may explore approaches such as multimodal decision fusion, calibrated confidence estimation, clinician-in-the-loop validation, and active learning techniques for better rare class sampling. Moreover, reporting per-class metrics and conducting failure mode analysis can help contextualize model performance, guiding more informed deployment strategies in clinical settings.

5. Conclusion

In summary, we organized CXR-LT 2024 to address the challenges of long-tailed, multi-label disease classification and zero-shot learning from chest X-rays. For this purpose, we have curated and released a large, long-tailed, multi-label CXR dataset containing 377,110 images, each labeled with one or more findings from a set of 45 disease categories. Additionally, we have provided a publicly available “gold standard” subset with human-annotated consensus labels to facilitate further evaluation. Finally, we outline a pathway to enhance the reliability, generalizability, and practicality of methods, with the ultimate goal of making them applicable in real-world clinical settings.

CRediT authorship contribution statement

Mingquan Lin: Visualization, Formal analysis, Writing – review & editing, Validation, Data curation, Writing – original draft, Methodology. **Gregory Holste:** Writing – original draft, Formal analysis, Validation, Writing – review & editing, Methodology. **Song Wang:** Writing – review & editing, Methodology, Validation, Data curation, Resources. **Yiliang Zhou:** Resources, Methodology, Validation, Data curation. **Yishu Wei:** Writing – review & editing, Validation. **Imon Banerjee:** Writing – review & editing, Software, Methodology. **Pengyi Chen:** Writing – review & editing, Software, Methodology. **Tianjie Dai:** Writing – review & editing, Software, Methodology. **Yuexi Du:** Writing – review & editing, Software, Methodology. **Nicha C. Dvornek:** Writing – review & editing, Software, Methodology. **Yuyan Ge:** Writing – review & editing, Software, Methodology. **Zuwei Guo:** Methodology, Writing – review & editing, Software. **Shouhei Hanaoka:** Methodology, Writing – review & editing, Software. **Dongkyun Kim:** Methodology, Writing – review & editing, Software. **Pablo Messina:** Methodology, Writing – review & editing, Software. **Yang Lu:** Methodology, Writing – review & editing, Software. **Denis Parra:** Methodology, Writing – review & editing, Software. **Donghyun Son:** Methodology, Writing – review & editing, Software. **Álvaro Soto:** Methodology, Writing – review & editing, Software. **Aisha Urooj:** Methodology, Writing – review & editing, Software. **René Vidal:** Methodology, Writing – review & editing, Software. **Yosuke Yamagishi:** Methodology, Writing – review & editing, Software. **Pingkun Yan:** Methodology, Writing – review & editing, Software. **Zefan Yang:** Methodology, Writing – review & editing, Software. **Ruichi Zhang:** Methodology, Writing – review & editing, Software. **Yang Zhou:** Methodology, Writing – review & editing, Software. **Leo Anthony Celi:** Supervision, Writing – review & editing. **Ronald M. Summers:** Writing – review & editing, Supervision. **Zhiyong Lu:** Writing – review & editing, Supervision. **Hao Chen:** Supervision, Writing – review & editing. **Adam Flanders:** Writing – review & editing, Supervision. **George Shih:** Writing – review & editing, Supervision. **Zhangyang Wang:** Validation, Conceptualization, Writing – review & editing, Supervision, Writing – original draft, Funding acquisition. **Yifan Peng:** Supervision, Conceptualization, Writing – review & editing, Project administration, Writing – original draft, Funding acquisition.

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 from a 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.

Acknowledgments

This work was supported by the National Library of Medicine [grant number R01LM014306], the National Science Foundation (NSF) [grant numbers 2145640, IIS-2212176], the Amazon Research Award, Cornell–HKUST Global Strategic Collaboration Award, the Artificial Intelligence Journal, the National Institute of Health [grant number R01EB017205], DS-I Africa [grant number U54TW012043-01], Bridge2AI [grant number OT2OD032701], the National Science Foundation, United States through ITEST #2148451, the National Center for Artificial Intelligence CENIA [grant number FB210017], and the ANID - Millennium Science Initiative Program [grant number ICN2021_004]. It was also supported by the NIH Intramural Research Program, National Library of Medicine and Clinical Center.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.media.2025.103739>.

References

- Abdalla, M., Fine, B., 2023. Hurdles to artificial intelligence deployment: Noise in schemas and “gold” labels. *Radiol. Artif. Intell.* 5 (2), e220056.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., McDermott, M., 2019. Publicly available clinical BERT embeddings. arXiv preprint [arXiv:1904.03323](https://arxiv.org/abs/1904.03323).
- Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al., 2022. Making the most of text semantics to improve biomedical vision–language processing. In: *European Conference on Computer Vision*. Springer, pp. 1–21.
- Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., Daneshjoui, R., Frankle, J., Liang, P., Carbin, M., et al., 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. arXiv preprint [arXiv:2403.18421](https://arxiv.org/abs/2403.18421).
- 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.
- Chen, H., Miao, S., Xu, D., Hager, G.D., Harrison, A.P., 2019. Deep hierarchical multi-label classification of chest X-ray images. In: *International Conference on Medical Imaging with Deep Learning*. PMLR, pp. 109–120.
- Dai, T., Zhang, R., Hong, F., Yao, J., Zhang, Y., Wang, Y., 2024. UniChest: Conquer-and-divide pre-training for multi-source chest X-Ray classification. *IEEE Trans. Med. Imaging*.
- 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. ViLMed: 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.
- Du, Y., Chang, B., Dvornek, N.C., 2024a. CLEFT: Language-image contrastive learning with efficient large language model and prompt fine-tuning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 465–475.
- Du, Y., Onofrey, J., Dvornek, N.C., 2024b. Multi-view and multi-scale alignment for contrastive language-image pre-training in mammography. arXiv preprint [arXiv:2409.18119](https://arxiv.org/abs/2409.18119).
- 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:1901.02757](https://arxiv.org/abs/1901.02757).
- 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.
- 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.
- Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Muller, N.L., Remy, J., 2008. Fleischner society: glossary of terms for thoracic imaging. *Radiology* 246 (3), 697–722.
- 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.
- Holste, G., Zhou, Y., Wang, S., Jaiswal, A., Lin, M., Zhuge, S., Yang, Y., Kim, D., Nguyen-Mau, T.-H., Tran, M.-T., et al., 2024. Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge. *Med. Image Anal.* 103224.
- Hong, Y., Zhang, X., Zhang, X., Zhou, J.T., 2024. Evolution-aware Variance (EVA) coreset selection for medical image classification. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, New York, NY, USA, pp. 301–310. [http://dx.doi.org/10.1145/3664647.3681592](https://doi.org/10.1145/3664647.3681592).
- 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.
- Huijben, E.M.C., Pluim, J.P.W., van Eijnatten, M.A.J.M., 2024. Denoising diffusion probabilistic models for addressing data limitations in chest X-ray classification. *Inf. Med. Unlocked* 50 (101575), 101575. [http://dx.doi.org/10.1016/j.imu.2024.101575](https://doi.org/10.1016/j.imu.2024.101575).
- Irvine, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Illcus, S., Chute, C., Marklund, H., Haghighi, 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, (01), pp. 590–597.
- 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](https://arxiv.org/abs/1901.07042).
- Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2022. Transformers in vision: A survey. *ACM Comput. Surv.* 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*. pp. 2702–2710.
- Li, Y., Liu, T., Shen, W., Cui, Y., Lu, W., 2024a. Improving generalization and personalization in long-tailed federated learning via classifier retraining. In: *European Conference on Parallel Processing*. Springer, pp. 408–423.
- Li, K., Wang, Y., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y., 2022. Uniformer: Unified transformer for efficient spatiotemporal representation learning. arXiv preprint [arXiv:2201.04676](https://arxiv.org/abs/2201.04676).
- Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J., 2024b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* 36.
- Lin, M., Li, T., Yang, Y., Holste, G., Ding, Y., Van Tassel, S.H., Kovacs, K., Shih, G., Wang, Z., Lu, Z., et al., 2023. Improving model fairness in image-based computer-aided diagnosis. *Nat. Commun.* 14 (1), 6261.
- 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.
- Messina, P., Vidal, R., Parra, D., Soto, Á., Araujo, V., 2024. Extracting and encoding: Leveraging large language models and medical knowledge to enhance radiological text representation. arXiv preprint [arXiv:2407.01948](https://arxiv.org/abs/2407.01948).
- 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. Heal. Inform.* 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.
- Naeini, M.P., Cooper, G., Hauskrecht, M., 2015. Obtaining well calibrated probabilities using bayesian binning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, (1).
- Neil, H., Dirk, W., 2020. Transformers for image recognition at scale. Online: <https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at-scale.html>.
- 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.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint [arXiv:2304.07193](https://arxiv.org/abs/2304.07193).
- Pantelatos, D., Theofilou, P.-A., Tzouveli, P., Kollias, S., 2024. Hybrid CNN-vit models for medical image classification. In: *2024 IEEE International Symposium on Biomedical Imaging. ISBI, IEEE*, pp. 1–4.
- Park, W., Ryu, J., 2024. Fine-grained self-supervised learning with Jigsaw puzzles for medical image classification. *Comput. Biol. Med.* 174 (108460), 108460. [http://dx.doi.org/10.1016/j.combiomed.2024.108460](https://doi.org/10.1016/j.combiomed.2024.108460).

- 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>.
- Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A., 2018. Film: Visual reasoning with a general conditioning layer. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, (1).
- Reis, E.P., De Paiva, J.P., Da Silva, M.C., Ribeiro, G.A., Paiva, V.F., Bulgarelli, L., Lee, H.M., Santos, P.V., Brito, V.M., Amaral, L.T., et al., 2022. BRAX, Brazilian labeled chest x-ray dataset. *Sci. Data* 9 (1), 487.
- 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, (1), MDPI, p. 10.
- Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L., 2021. Asymmetric loss for multi-label classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 82–91.
- 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.
- Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M., 2020. CheX-clusion: Fairness gaps in deep chest X-ray classifiers. In: *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*. World Scientific, pp. 232–243.
- Tan, M., Le, Q., 2021. Efficientnetv2: Smaller models and faster training. In: *International Conference on Machine Learning*. PMLR, pp. 10096–10106.
- Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y., 2022. Maxvit: Multi-axis vision transformer. In: *European Conference on Computer Vision*. Springer, pp. 459–479.
- 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.
- Wei, Y., Wang, X., Ong, H., Zhou, Y., Flanders, A., Shih, G., Peng, Y., 2024. Enhancing disease detection in radiology reports through fine-tuning lightweight LLM on weak labels. *arXiv preprint arXiv:2409.16563*.
- Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W., 2023. Medklip: Medical knowledge enhanced language-image pre-training in radiology. *arXiv preprint arXiv:2301.02228*.
- 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.
- 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, (3), pp. 2982–2990.
- Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L., 2023. Sigmoid loss for language image pre-training. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11975–11986.
- Zheng, Q., Zhao, W., Wu, C., Zhang, X., Dai, L., Guan, H., Li, Y., Zhang, Y., Wang, Y., Xie, W., 2024. Large-scale long-tailed disease diagnosis on radiology images. *Nat. Commun.* 15 (1), 10147.
- 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.