CXR-LT 2024：一种关于从胸部 X 光片进行长尾、多标签和零样本疾病分类的 MICCAI 挑战

Mingquan Lin[1,†], Gregory Holste[2,†], Song Wang[2], Yiliang Zhou[3], Yishu Wei[3], Imon Banerjee[4], Pengyi Chen[5], Tianjie Dai[5,6], Yuexi Du[7], Nicha C. Dvornek[7,8], Yuyan Ge[9], Zuwei Guo[10], Shouhei Hanaoka[11], Dongkyun Kim[12], Pablo Messina[13,14], Yang Lu[15], Denis Parra[13,14], Donghyun Son[16], Álvaro Soto[13], Aisha Urooj[4], René Vidal[26], Yosuke Yamagishi[11], Zefan Yang[17], Ruichi Zhang[15], Yang Zhou[18], Leo Anthony Celi[20,21,22], Ronald M. Summers[23], Zhiyong Lu[24], Hao Chen[25], Adam Flanders[19], George Shih[3], Zhangyang Wang[2,*], Yifan Peng[3,*]

[*]Corresponding author(s). Email(s): atlaswang@utexas.edu, yip4002@med.cornell.edu
[†]These authors contributed equally to this work.

1. Department of Surgery, University of Minnesota, Minneapolis, USA
2. Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, USA
3. Department of Population Health Sciences, Weill Cornell Medicine, New York, 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. Department of Biomedical Engineering and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, USA
18. Institute of High Performance Computing, A*STAR, Singapore
19. Department of Radiology, Thomas Jefferson University Hospital, Philadelphia, USA
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. National Center for Biotechnology Information, National Library of Medicine, Bethesda, USA
25. Department of Computer Science and Engineering, Hong Kong University of Science and

Technology, Hong Kong, China
26. Center for Innovation in Data Engineering and Science, Departments of Electrical and Systems Engineering & Radiology, University of Pennsylvania, Philadelphia, USA

Abstract

CXR-LT 系列是一个社区驱动的倡议，旨在通过胸部 X 光（CXR）增强肺病分类。该系列解决了开放的长尾肺病分类中的挑战，并提升了尖端技术的可测量性。首次活动 CXR-LT 2023 旨在通过提供高质量的基准 CXR 数据用于模型开发，并进行全面评估，以识别影响肺病分类性能的持续问题，来实现这些目标。基于 CXR-LT 2023 的成功，CXR-LT 2024 扩展了数据集至 377,110 张胸部 X 光片（CXR）和 45 个疾病标签，包括 19 项新的罕见疾病发现。它还引入了零样本学习的新重点，以解决在先前活动中识别出的限制。具体来说，CXR-LT 2024 设有三个任务：（i）在大规模噪声测试集上的长尾分类，（ii）在手动注释的“黄金标准”子集上的长尾分类，（iii）对五项以前未见的疾病发现的零样本泛化。本文概述了 CXR-LT 2024，详细介绍了数据整理过程并整合了尖端解决方案，包括使用多模态模型进行罕见疾病检测，处理噪声标签的先进生成方法以及针对未见疾病的零样本学习策略。此外，扩展的数据集提升了疾病涵盖范围，以更好地代表现实临床环境，为未来研究提供了宝贵的资源。通过综合参与团队的洞察力和创新，我们旨在推进临床上真实且可推广的胸部影像诊断模型的发展。

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

1. 介绍

CXR-LT 系列标志着一个社区驱动的倡议，旨在改善使用胸部 X 光片（CXR）进行肺部疾病分类，该倡议解决了开放长尾肺部疾病分类中的挑战，并推进了先进技术可测量性的进步 [1] 。这些目标是在首次活动 CXR-LT 2023 [2] 期间追求的，通过提供高质量的基准 CXR 数据用于模型开发，并进行详细评估以识别影响肺部疾病分类性能的持续问题。CXR-LT 2023 吸引了广泛关注，共有 59 个团队提交了 500 多份独特的参赛作品。从那时起，任务设置和数据为大量研究提供了基础 [3–6] 。

作为该系列的第二个事件，CXR-LT 2024 保持了其前身的总体设计和目标，同时引入了对零射学习的新重视。这个新增的内容解决了在 CXR-LT 2023 中识别出的一个限制。估计超过 4,500 种独特的放射学发现表明，CXR 上临床发现的实际分布至少是当前基准能够提供的两个数量级。因此，有效地解决放射学异常发现的“长尾”问题，需要开发一个能够以“零射”方式泛化到新类别的模型。

本文概述了 CXR-LT 2024 挑战赛，包含了吸引广泛参与的两项长尾任务以及一个新引入的零样本任务。任务 1 和任务 2 专注于长尾分类，任务 1 使用一个大规模且噪声较大的测试集，而任务 2 使用一个小规模且人工标注的测试集。任务 3 涉及对之前未见过的疾病进行零样本泛化。每个任务都遵循由 CXR-LT 2023 建立的一般框架，为参与者提供一个由超过 250,000 个胸部 X 光图像和 40 个二元疾病标签组成的大型自动标注训练集。参与者最终的提交结果将与以类似方式准备的单独保留的测试集进行评估。

在接下来的部分中，我们介绍每个任务的设置并概述评估标准。接下来，我们详细描述数据整理过程，然后展示每个任务的结果。然后，我们整合表现优异的解决方案中的关键信息，并提供实用的观点。最后，我们利用我们的研究结果建议未来少样本和零样本疾病分类的路径，强调利用多模态基础模型的潜力。

2. 方法

2.1 主要任务

CXR-LT 2024 挑战包括三个任务：1）在一个大型且噪声多的测试集上进行长尾分类，（2）在一个小型且手动标注的测试集上进行长尾分类，以及（3）对以前未见过的疾病进行零样本泛化。所有这些都可以表述为多标签分类问题。

Table 1: 三项任务中使用的数据集的 特征。

	Task 1		Task 2		Task 3	
Dataset	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

考虑到这些任务中严重的标签不平衡，主要的评估指标是平均精度（mAP），特别是跨类的“宏平均”精度。虽然接收器操作特征曲线下面积（AUROC）通常用于类似的数据集 [7, 8] ，但在存在类别不平衡的情况下，它可能会被严重夸大 [9, 10] 。相比之下，mAP 更适合长尾、多标签的情境，因为它在不削弱类别不平衡的情况下衡量跨决策阈值的性能 [11] 。为了详尽，我们计算了平均 AUROC（mAUROC）和平均 F1 分数（mF1）——阈值为 0.5——作为辅助分类指标。我们还计算了平均预期校准误差（ECE）[12] 以量化偏差。为了进一步增强临床可解释性，我们也报告了每类的 F1 分数，以及宏平均和微平均 F1 分数及关键发现的假阴性率，除了挑战的主要评估指标之外。我们相信这些补充能够在实际环境中提供对模型性能的更细致的了解。

2.2 数据集整理

表 1 列出了用于这三个任务的数据集的特征。所有三个任务使用相同的训练数据集。同样，所有三个任务共享相同的开发数据集，其中任务 3 专注于五个未见过的类别。任务 1 和任务 3 共享相同的测试集；但是，任务 3 明确评估五个未见过类别的表现。任务 2 是任务 1 的一个子集，包含 26 个手动注释的类别。图 1 列出了 45 个类别，其中五个未见过的类别为 Bulla、Cardiomyopathy、Hilum、Osteopenia 和 Scoliosis。其余 40 个类别排除了这五个未见过的类别，而 14 个类别是源自原始 MIMIC-CXR 数据集的和在 CXR-LT 2023 引入的 12 个附加类别。图 2 展示了挑战数据集中的胸部 X 光片示例，其中每个图像包含多个注释不正常现象。

在本节中，我们详细介绍了两个数据集的数据整理过程：(i) 在任务 1 和任务 3 中使用的自动标记的 CXR-LT 数据集，(ii) 在任务 2 中使用的人工注释的“黄金标准”测试集。

2.2.1 CXR-LT 数据集

CXR-LT 挑战数据集是通过扩展 MIMIC-CXR 数据集的标签集 [13] ，[1] 而开发的，结果是一个更加复杂的、长尾的标签分布。今年，新添加的临床发现是从包括 PadChest 数据集的疾病列表 [14] 和胸部成像术语的 Fleischner 词汇表 [15] 等来源中选择的。在确保在数据集中观察到足够的出现次数以进行可靠评估之后，这 19 个新疾病发现是：(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 Hypertension, (12) Rib Fracture, (13) Round Atelectasis, (14) Tuberculosis, (15) Bulla, (16) Cardiomyopathy, (17) Hilum, (18) Osteopenia, (19) Scoliosis. 最后五个异常发现 – 肺大泡、心肌病、肺门、骨质减少和脊柱侧弯 – 未包含在挑战训练集中，并被保留用于任务 3 中的零样本评估。

此外，我们将“无发现”类别替换为更直观的“正常”类别。“无发现”表示标签集中没有出现异常发现。例如，使用原始的 14 个 MIMIC-CXR 类别时，“无发现”意味着这 14 个发现中没有任何发现；然而，当标签集扩展到 26 个类别时，这个“无发现”标签实际上可能包括 12 个新增异常中的一个。为避免在不同任务中对这一标签的解释不清，我们为简化的“正常”类别创建了新标签，表示报告中未发现心肺疾病或异常。如在 2023 年一样，每次 CXR 研究的放射报告都使用 RadText [16] ，一个放射学文本分析工具，进行解析以提取新疾病的存在状态。

[1] https://physionet.org/content/mimic-cxr/2.0.0/

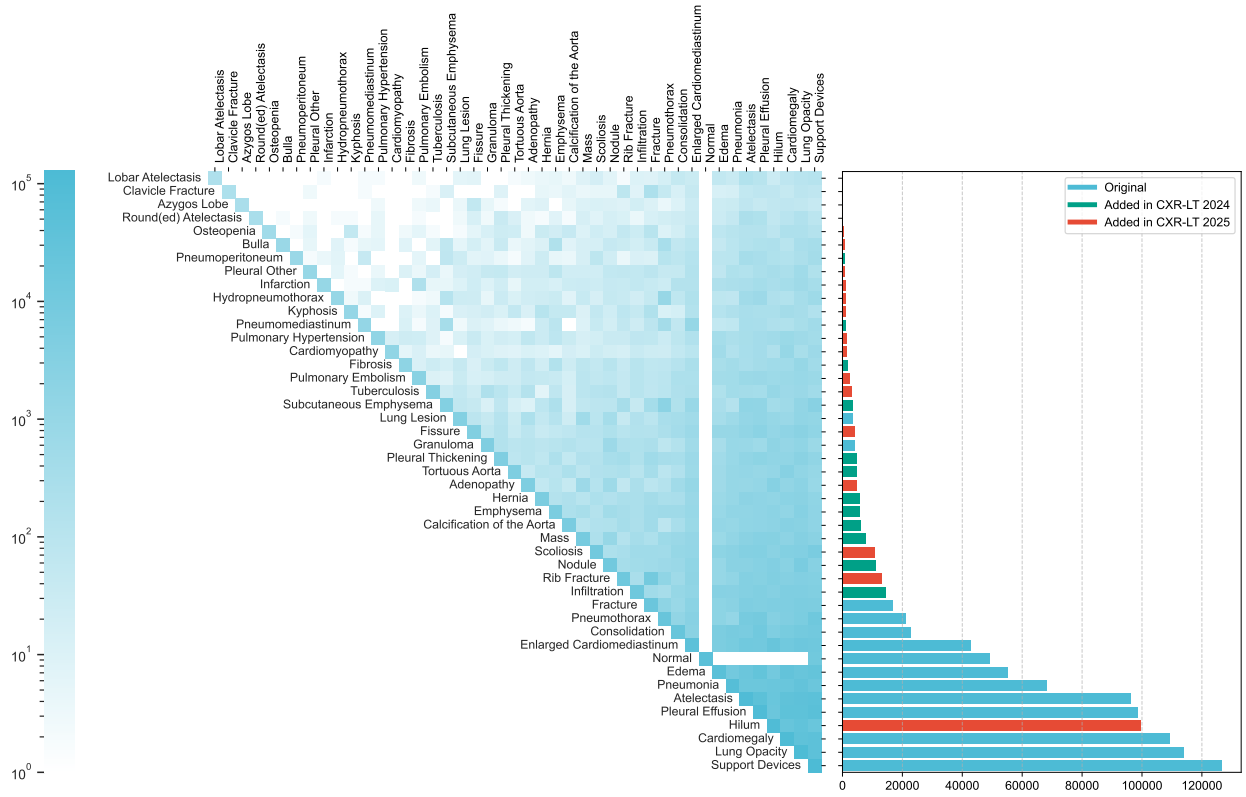

Figure 1: CXR-LT 2024 挑战数据集的长尾分布。该数据集通过解析放射学报告，将 MIMIC-CXR 基准扩展为包含 12 个新的临床发现（红色）而形成。

最终数据集包含 377,110 张 CXR 图像，每张图像都标有一种或多种 45 种疾病，呈现长尾分布（图 1 ）。与 CXR-LT 2023 类似，我们选择使用 MIMIC-CXR-JPG 数据集 [17] 中的图像数据，以增加可访问性并减少存储此数据集的负担（使用 MIMIC-CXR 提供的原始 DICOM 数据为 ∼ 600 GB 对比 ∼ 4.7 TB）。[2] 数据集在患者级别上随机划分为训练集（70 %）、开发集（10 %）和测试集（20 %）；重要的是，这一划分是 CXR-LT 2024 独有的，这意味着参与者不能重复使用前一年挑战的模型。参与者可以访问所有图像，但仅提供训练集的标签。

2.2.2 黄金标准测试集

在我们的 CXR-LT 2023 [2] 综述中，我们使用了一个从挑战测试集衍生的手动标注“金标准”集来评估手动和自动标注之间的差异，以及顶级解决方案在标签噪声减少的情况下如何在该测试集中表现。具体来说，六名标注员审查了 406 份 MIMIC-CXR 放射报告，以确定 CXR-LT 2023 中考虑的 26 种疾病发现的存在与否。有关此金标准集的完整数据整理细节，请参见 Holste et al. [2] 。该数据集为评估模型在较小、经过手动审核的测试集上的表现提供了高质量的基准。今年，在 CXR-LT 2024 中，我们将此金标准数据集用作任务 2 的测试集。

2.3 日程

表 2 显示了任务时间表。该挑战在 CodaLab 平台上进行，[3] 每个三个任务都有一个单独的 CodaLab 页面 [18] 。任何注册的 CodaLab 用户都可以申请，但只有在提交能够访问 MIMIC-CXR-JPG 所

[2] https://physionet.org/content/mimic-cxr-jpg/2.0.0/

[3] https://codalab.lisn.upsaclay.fr/competitions/18601 , https://codalab.lisn.upsaclay.fr/competitions/18603 , https://codalab.lisn.upsaclay.fr/competitions/18604

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

(b) Fracture, Pleural Effusion

(c) Cardiomegaly, Edema, Lung Opacity

Figure 2: 来自挑战数据集的代表性胸片，每张都展示了多个发现。(a) 包括门脉标签（CXR-LT 2024 中新引入）；(b) 显示骨折（在 CXR-LT 2023 中引入）；(c) 显示原始 MIMIC-CXR 标签（心脏肥大、水肿、肺不透明）。

Table 2: 2024 年 CXR-LT 的时间表。

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

需的 PhysioNet 凭证证明后才会被接受。

在开发阶段（2024 年 5 月 1 日至 2024 年 8 月 26 日），已注册的参与者下载了带标签的训练集和未标签的开发集，并由此生成一个带有预测结果的逗号分隔值（CSV）文件以进行上传。提交内容在保留的开发集上进行评估，结果被实时更新在公开的排行榜上。在测试阶段（2024 年 8 月 26 日至 2024 年 9 月 6 日），发布了测试集图像（无标签）。要求参与者提交带有测试集预测结果的 CSV 文件，以进行最终评估和各任务排名。在这一阶段排行榜是隐藏的，每个团队得分最高的提交被保留。最终测试阶段排行榜主要根据 mAP 进行排名，若出现并列则按 mAUROC 排序。

3. 结果

3.1 参与

在 CXR-LT 挑战中，共有 96 个团队在 CodaLab 上提交申请，其中有 61 个在提供 MIMIC-CXR-JPG [17] 访问凭据证明后获得批准。在开发阶段，有 29 个团队参与，共提交了 661 次、349 次和 364 次独特的提交，分别用于任务 1、2 和 3 的公共排行榜。在最终测试阶段，共有 17 个团队参与。我们选择了排名前 9 的团队，邀请其在 MICCAI 2024 [4] 的 CXR-LT 2024 挑战活动中进行展示，并将其纳入本研究。由于有两个团队在任务 1 和任务 2 中表现优异，最终包含了任务 1 和 2

[4] https://cxr-lt.github.io/CXR-LT-2024/

Table 3: 顶尖的 CXR-LT 2024 挑战解决方案概述。ENS - 集成；LRW - 损失重加权；VL - 视觉语言。

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

的前四名解决方案以及零样例任务 3 的前三名解决方案。表 3 总结了参与一个或多个任务并提交系统说明的表现最佳的团队。包括所有演示幻灯片在内的更多详细信息可在 GitHub [5] 上获得，使读者能够更深入地探索所有方法的细节。

3.2 系统描述

团队 A: `zguo` 该团队提出了用于 CXR-LT 分类的 ChexFusion+ 模型，参与了任务 1 和任务 2。他们的方法利用了通过合成生成的长尾数据训练的 12 个多分辨率 ConvNeXt [19] 模型。具体而言，他们使用输入提示和随机高斯噪声，通过条件去噪 U-Net 和变分自编码器解码器（VAE）解码器生成两个图像。这两个生成的图像，以及输入提示和对应的 MIMIC-CXR 图像，被用作 ConvNeXt 的训练输入。为了解决长尾案例中的数据不平衡问题，他们使用预训练的大型生成模型为每个罕见疾病类别生成 100 个新图像。这些合成图像使用精心构建的提示生成，提示中指定了多种共现的胸腔异常，例如“圆形（或）肺不张、气胸、胸腔积液、肺浑浊和肺不张”，以反映真实的放射合并症。

小组 B: `tianjie_dai` 这个团队利用多模态集成方法解决 CXR-LT 任务中的不平衡多标签分类挑战。具体而言，他们采用 EfficientNetV2-Large [20] 和 PubMedBERT [21] 模型的集成，并在统一医学语言系统（UMLS）知识图谱 [22] 上进行微调，以整合图像和文本特征。为了解决类别不平衡

[5] `https://github.com/CXR-LT/CXR-LT-2024`

问题，他们使用非对称损失函数 [23] ，为稀有类别赋予更高的权重。为了提高模型的稳健性和泛化能力，采用了测试时增强技术，包括调整大小、裁剪和翻转。此外，他们从多个来源引入外部数据集，如 ChestXRay-14 [7] 、CheXpert [24] 、VinDr-CXR [25] 和 BRAX [26] ，以增强对罕见疾病标签的表示学习。

团队 C： `XYPB` 该团队通过多模态集成方法解决了 CXR-LT 挑战，利用多视角和多尺度图像对齐来增强长尾多标签环境下的分类。他们的方法建立在 CLEFT [27] 和 MaMA [28] 框架的基础上，结合了对比语言-图像预训练（CLIP），并在胸部 X 光研究的多视图图像对中加入了额外的图像-图像和图像-文本对比学习。这种方法使他们的模型能够从不同的视角获取信息，如后前位和侧位视图。为了解决医学影像的多尺度特性，他们引入了一个对称的局部交叉注意力（SLA）对齐模块，通过交叉注意建模特定区域的视觉-文本相关性，将局部图像区域与描述性文本段对齐。为了解决类别不平衡问题，他们使用了加权不对称损失 [29] 。对于图像编码器，他们采用了 ConvNeXt-S [19] 和 EfficientNet V2-S [20] 主干，这些主干在 ImageNet 上进行了分类训练，并在 MIMIC-CXR 上使用基于 CLIP 的方法进行了训练，用于任务 1 和任务 2，并使用参数高效微调的医学大型语言模型，即 BioMedLM [30] ，作为他们的语言编码器。

团队 D： `dongkyunk` 该团队实施了一个两阶段框架，旨在有效利用每位患者可用的多视图。在第一阶段，使用 ML-Decoder [31] 分类头与 Noisy Student [32] 自学习一起训练了一种单一模型。在第二阶段，引入了一个基于 Transformer 的模型，称为 CheXFusion，用于聚合多视图特征 [23] 。CheXFusion 中的特征聚合类似于自然语言处理中的多句编码，其中每个句子代表一个单独的胸部 X 光片。此外，采用了一种加权版本的非对称损失 [29] 来解决由于疾病长尾分布产生的类间不平衡以及多标签分类中负标签占多数导致的类内不平衡问题。

团队 E: `yangz16` 该团队采用了一种基于基础模型的方法来应对 CXR-LT 挑战。他们的方法包含三个关键组件：以 ViT-Large 网络架构 [33] 为骨干的 DINOv2 基础模型 [34] ，ML-Decoder [33] 分类头，以及多视图/多分辨率集成。DINOv2 模型在超过 71 万张来自不同数据集的胸部 X 光片上进行了预训练，这些数据集包括 MIMIC-CXR [13] 、CheXpert [24] 、PadChest [14] 、NIH Chest X-ray14 [7] 和 BRAX [26] ，通过自监督学习，结合自蒸馏损失和掩码图像建模来学习稳健的表示。ML-Decoder 将基础模型中的局部特征映射到特定疾病的预测结果用于分类，采用注意力机制以实现疾病发现的更精细定位。

队伍 F: `YYama` 这个团队利用了 ConvNeXt V2 [19] 和 MaxViT [35] 模型的集合，并进行了领域特定的预训练和基于视图的聚合。ConvNeXt V2 模型在 ImageNet 上进行了预训练，而 MaxViT 模型则进一步在 NIH Chest X-ray 数据集上进行了预训练。为了解决类别不平衡的问题，他们应用了一个不对称损失函数 [29] ，并结合类别权重，给予稀有类别更高的重要性。此外，他们实施了一种基于视图的预测聚合方法，将来自正面和侧面的视图预测相结合，并通过加权平均来偏向正面视图。

团队 G： `yyge` 该团队采用了一种双模型策略，将视觉语言模型（VLM）和多视图视觉模型（MVM）结合，用于胸部 X 光片的零样本和多标签疾病分类。VLM 集成了 DINOv2 [34] 作为图像编码器和 BERT [36] 用于文本编码，利用来自 ChatGPT [37] 的精细化疾病描述。该模型首先在特定领域的数据集上进行预训练，然后在 CXR-LT 训练集上进行了微调，使用加权二元交叉熵损失来处理类别不平衡问题。同时，MVM 通过从放射报告中挖掘特定疾病示例，将零样本任务转化为少样本问题，并利用 DINOv2 和轻量级 Transformer 聚合多视图特征。使用加权不对称损失 [23] 和多视图学习进一步解决了长尾分布的问题。这一结合框架有效捕获了特定领域的知识，并在已知和未知疾病之间平衡了性能。

团队 H： `ZhangRuichi` 该团队利用了一种受 MedKLIP [38] 启发的视觉语言模型（VLM），结合了解剖和文本信息，以增强泛化能力和性能。作者使用了一个未经预训练的 ResNet50 [39] 架构进行图像编码，因为在 ImageNet 上进行预训练可能会由于自然图像和胸部 X 光片数据集之间的领域

Table 4: 在任务 1 的测试集中，对所有 40 个类别进行评估时，前四名团队最终模型的 mAP。还展示了 mAUROC、mF1 和 mECE，括号中的数字表示根据相应评估指标的排名。

			mAUROC	mF1	mECE
Ranking	Team	mAP	🏆	🏆	🏆
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)

和分布差异而引入偏差。通过从头训练模型，作者旨在减小这些领域和分布的差距，更好地根据胸部 X 光片图像的特征进行模型定制。为了实现零样本能力，类别标签被补充了 GPT-4 生成的描述，并使用了 BioClinical BERT [40] 进行文本编码，以捕捉丰富的语义信息。整合了来自 CheXpert 的解剖位置数据，将 MIMIC-CXR 疾病映射到相应的区域，以解决缺乏细粒度标签的问题。在训练过程中，他们应用交叉熵损失以提高准确性，并使用对比损失将疾病与解剖区域关联。在测试期间，实施了按类别的集成策略和来自不同患者视图的预测的测试时融合以提高整体准确性。

队伍 I： `pamessina` 这个团队为零样本分类任务开发了一个多模态模型。文本编码器是 CXR 事实编码器（CXRFE）[41]，它从简短的事实性句子中计算事实嵌入。训练过程中，文本编码器保持冻结状态。图像编码器则进行端到端训练，团队尝试了多种架构，包括 DenseNet121 [42]、SigLIP Base [43]、ConvNext-Small [19] 和 Uniformer [44]。该模型使用事实嵌入通过 FiLM [45] 调节来自图像编码器的局部和全局特征，以预测二元粗略分割掩膜（表示事实的视觉定位）和整个图像的事实全局二元分类。使用 21 个模型的集成体达到了最佳效果，每个模型使用不同的图像编码器和训练数据配置。此外，本研究还使用 GPT-4 [37] 作为 MIMIC-CXR 报告的自动标注器，并包括了一些额外的数据集，其中一些数据集包含边界框以提供视觉定位监督。

3.3 任务 1 主要评估结果

CXR-LT 测试阶段结果 任务 1 中前四名团队的结果列在表 4 中。A 队以 mAP 0.281 获得第一名，B 队以 mAP 0.279 名列第二，而 C 队和 D 队都获得了 mAP 0.277。然而，由于 C 队的 mAUC 更高，排名第三。在 CXR-LT 2023 中，前四名团队获得的 mAP 分数从 0.349 到 0.372，比今年的分数高得多，这是因为 CXR-LT 2024 增加了 19 个新的罕见类别。当用 CXR-LT 2023 的 26 个标签集合评估今年的顶尖解决方案时，我们观察到的 mAP 分数分别为 0.371、0.373、0.371 和 0.370。与 CXR-LT 2023 相比，任务 1 的顶尖选手在这些类别上的整体表现有所提高（例如，CXR-LT 2023 中获得第二名的选手达到了 0.354 mAP）。补充表 1 详细列出了前四名团队每个类别的表现。此外，补充表 2 和表 3 展示了每个类别的 F1 分数、宏平均和微平均 F1 分数，以及关键发现的假阴性率。

长尾分类性能 为了根据标签频率检查预测性能，我们将 40 个目标类别分为“头部”（> 10 %）、“中等”（1 % –10 %）和“尾部”（< 1 %）类别，基于它们在训练集中的流行度，分别由 9、14 和 16 个类别组成。类别的 mAP 在表 5 中展示，同时有头部、中等和尾部 mAP 的“类别平均值”。团队 `A` 不仅在整体性能上取得了最高成绩，而且在“尾部”组也取得了优异表现。然而，“尾部”组的前三名表现非常接近。团队 `A` 使用预训练的大型生成模型为这些尾部案例生成新图像，而团队 `B` 和 `C` 应用了损失重加权，表明这两种方法都可以提高“尾部”组的性能。

任务 2 中前四名团队的结果列在表 6 中。第一名由团队 `C` 获得，mAP 为 0.526。团队 `E` 获得第二名，mAP 为 0.511，团队 `A` 获得第三名，mAP 为 0.511，团队 `F` 获得第四名，mAP 为 0.509。所有四个团队都使用集成方法来提高模型性能，取得了与去年相似的结果；例如，CXR-LT 2023 的顶级表现者在同一金标准测试集上取得了 0.519、0.518 和 0.519 的 mAP 分数。补充表 4 提供了这

Table 5: 通过计算每个类别内的平均 mAP，来评估长尾分类在“头部”、“中部”和“尾部”类别上的表现。这些类别是根据训练集中每个类别的相对频率确定的（在括号中标出）。最右边的列表示头部、中部和尾部 mAP 的平均值。每列中最佳的 mAP 以粗体显示。

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

Table 6: 在任务 2 中，顶级 4 支团队最终模型在黄金标准测试集上所有 26 个类别的 mAP。同时还展示了 mAUROC、mF1 和 mECE，括号中的数字表示基于相应评估指标的排名。

			mAUROC	mF1	mECE
Ranking	Team	mAP	🏆	🏆	🏆
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: 对于任务 3 中所有五个未见类别，测试集中前三名队伍的最终模型性能评估。还展示了 mAUROC、mF1 和 mECE，其括号中的数字表示基于相应评估指标的排名。

			mAUROC	mF1	mECE
Ranking	Team	mAP	🏆	🏆	🏆
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)

些顶级团队的详细类别特定表现。此外，补充表 5 和 6 提供了每个类别的 F1 分数、宏平均和微平均 F1 分数，以及关键发现的漏报率。

如第 2.2.2 节所述，任务 2 的测试集是任务 1 测试集的一个子集，仅有 26 个手动标注的标签。表 4 和 6 显示，任务 2 中的第一支队伍在任务 1 中排名第三，而任务 2 中的第二支队伍在任务 1 中名列第一。此外，从挑战排行榜 [6] [7] 可以看出，任务 2 中排名第三和第四的队伍在任务 1 中分别排名第六和第五，其 mAP 分数分别为 0.269 和 0.273。我们选择了十支同时提交任务结果的队伍，命名为 T1 到 T10，并根据结果分析了它们的表现。尽管这些数据集之间存在较大的分布差异，但任务 1 和任务 2 之间的整体性能一致性仍然保持稳定（图 3；$R^2 = 0.946$，$r = 0.972$）。

3.4 任务 3 主要评估结果

表 7 展示了任务 3 中表现最好的三个团队的结果。团队 `G` 以 0.129 的 mAP 获得第一名。紧随其后，团队 `H` 以 0.116 的 mAP 获得第二名，而团队 `I` 以 0.110 的 mAP 获得第三名。与其他任务相比，任务 3 中表现相对较低可以归因于检测训练期间从未见过的发现的零样本性质的挑战。附录表 7 详述了这些顶部团队的特定类别表现。此外，附录表 8 和表 9 报告了每类的 F1 分数、宏平均和微平均 F1 分数，以及关键发现的假阴性率。

[6] `https://codalab.lisn.upsaclay.fr/competitions/18603#results`
[7] `https://codalab.lisn.upsaclay.fr/competitions/18601#results`

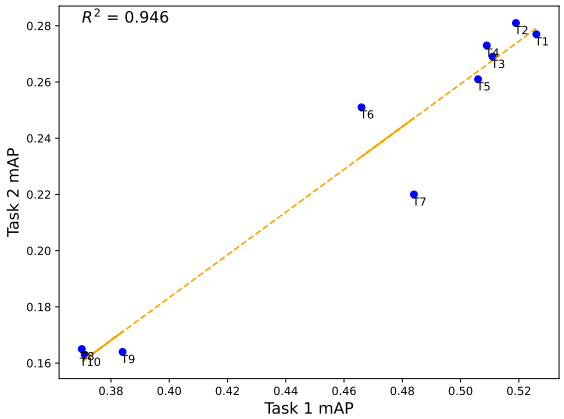

Figure 3: 对 CXR-LT 任务 1 数据（第 2.2.1 节）和黄金标准任务 2 数据（第 2.2.2 节）的性能比较。

3.5 基于规则和 GPT-4o 标签的比较

在 CXR-LT 数据集中，标签最初是使用基于规则的方法生成的。此类方法已被证明在自动标注现有 CXR 数据集 [7, 14, 24] 中取得成功，但近期大型语言模型（LLMs）的发展表明它们可能是完成此任务的有用选择。`I` 团队选择使用 GPT-4 对 MIMIC-CXR 报告中的数据进行标注，而不是依赖于基于规则的标签。我们使用 406 个手动注释的样本集计算了精度，以评估大型语言模型是否可以生成更准确的标签。表 **??** 列出了基于规则的方法与 GPT-4o 之间的性能比较，使用的是 Wei et al. [46] 建议的提示。基于规则的方法获得了 0.711 的精度，而 GPT-4o 达到了 0.786 的精度。

4. 讨论

4.1 顶级 CXR-LT 2024 解决方案的主题

如表格 3 中的系统描述所示，我们观察到在这三个任务中表现突出的解决方案中有几个共同的主题，以及一些独特的视角。

现代卷积神经网络架构 表现最佳的解决方案通常使用卷积神经网络（CNNs）作为图像编码器，延续了 CXR-LT 2023 的趋势。ConvNeXt 成为最受欢迎的选择 [19] ，其次是 EfficientNet [20] 、ResNet [39] 和 DenseNet [42] 。无论是在 2023 年还是 2024 年，ConvNeXt 稳定地超越其他架构。我们将 ConvNeXt 的受欢迎程度和强劲表现归因于两个主要因素：（1）去年的顶尖解决方案的使用树立了一个标准，（2）其设计可在不同图像分辨率中实现可扩展的性能，使其非常适合捕捉多尺度信息。

视觉变压器 虽然在 2023 年，没有任何表现最好的解决方案使用 Vision Transformers (ViTs) [47] 作为图像编码器，但到了 2024 年，情况发生了显著变化，有四个团队采用了基于 ViT 的模型。具体来说，团队 `E` 和 `G` 完全依赖 ViTs 进行图像编码。相反，团队 `F` 和 `I` 选择了混合方法，结合了基于 ViT 的 Transformers 和 CNNs。我们将 ViT-based transformers 更多的采用归因于两个主要因素：（1）结合 CNNs 和 ViTs [48] 所提供的互补优势，可以增强特征提取和表示能力，以及（2）策略性地使用额外的数据集来对图像编码器进行预训练，这对有效训练基于 ViT 的 transformers，提高其鲁棒性和泛化能力尤其有利。

大规模预训练 在表现最佳的九个解决方案中，有八个依赖于有监督的预训练或迁移学习。尽管有些团队使用了标准的 ImageNet 预训练模型，然而其他一些团队在公开可用的“领域内”CXR 数

Table 8: 使用微精度评估模型在长尾多标签疾病分类中的表现，我们的金标准测试集上进行评估。

	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

据集，如 ChestXRay-14 [7] 、CheXpert [24] 、VinDr-CXR [25] 和 BRAX [26] 上进行了额外预训练。值得注意的是，团队 B 、C 、D 、F 和 G 采用了多阶段预训练策略，首先在自然图像上进行一般预训练，然后在 CXR 数据上进行特定领域预训练，类似于 CXR-LT 2023 中一些团队使用的方法。相比之下，团队 H 报告他们提出的模型无需预训练就取得了更好的性能。

集成学习和数据增强 与 CXR-LT 2023 一样，许多顶级解决方案（九个中有八个）使用了多种集成学习策略来提高泛化能力 [49, 50] 。团队 B 、C 、F 、G 和 I 在不同的模型架构之间创建了集成；团队 A 和 E 通过使用不同的图像分辨率形成集成；而团队 H 则基于同一图像的多个视角构建了集成策略，但在这些视角之间使用相同的模型架构。除了集成学习，所有团队还采用了图像增强这一已被验证的技术来提升泛化能力 [51] 。值得注意的是，团队 A 利用扩散模型生成合成图像来加强稀有尾类的增强。

损失重加权 为了应对标签的长尾分布，九个最佳解决方案中有五个采用了损失重加权技术，以提高稀有尾部类别的重要性。这五个团队（B 、C 、D 、F 和 G ）都采用了加权非对称损失 [29] ，这种损失专门用于处理不平衡的多标签分类场景。此外，团队 G 在这种方法的基础上实施了加权二元交叉熵损失。加权非对称损失函数的广泛采用可以归因于它在 CXR-LT 2023 中的成功，当时排名靠前的团队采用了这种方法。值得注意的是，在任务 3 中有两个顶尖解决方案选择不使用加权损失，主要是由于缺乏关于五个未见类别分布的信息。然而，任务 3 中有一个顶尖解决方案通过将零样本问题转换为少样本问题，采用了加权损失方法，利用从基于文本的描述中提取的关于未见类别的先验知识。

多模态视觉语言学习 多模态视觉语言学习最近在放射学的深度学习领域变得流行，特别是作为一种使用成对的胸片（CXR）图像和自由文本放射报告的预训练方法 [52–57] 。今年，九个团队中有八个成功地利用了图像和文本数据。例如，团队 C 使用了图像到图像和文本到图像的对比学习组合来增强特征表示。同时，团队 D 使用了 ML-Decoder [31] 分类头，利用跨注意力机制将标签视为与图像特征交互的文本“查询”。值得注意的是，在任务 3 中，所有三个团队都使用了多模态视觉语言模型，以实现对任务 3 中新出现的五个类别的零样本泛化。

ChatGPT/GPT-4 随着大型语言模型在通用和医学领域的日益普及，三个团队在第三项任务中使用了 ChatGPT 或 GPT-4。团队 G 使用 ChatGPT [37] 创建了细粒度的疾病描述，可能提高文本编码器的性能。团队 H 利用 GPT-4 为类别标签生成描述性文本增强，从而促进零样本学习。相比之下，团队 I 选择不使用提供的标签，而是使用 GPT-4 作为 MIMIC-CXR 报告的自动标注器。他们进一步结合了额外的 CXR 数据集，其中包括一些带有边界框的数据集，以支持视觉固定监督。

合成数据对长尾分类的影响 团队 A 使用生成模型为稀有类别创建合成数据是一种有前景的方法，这在他们的表现中得以体现。合成数据有潜力通过补充代表性不足的类别来缓解极端的类别不平衡，特别是在真实数据收集成本高昂或速度缓慢的领域。然而，这也引发了关于数据保真度、领域转移和过拟合的问题。随着合成数据生成技术（如扩散模型、GANs）的不断发展，它们在解决长尾医学图像分类中的作用值得进一步研究，特别是在可信度、可推广性和临床实用性方面。

本研究的一个关键局限是依赖于 MIMIC-CXR 数据集，该数据集是在美国的一所学术医疗中心收集的。因此，这些数据可能反映特定机构的患者人口统计学特征、疾病流行率、成像协议和设备特性。这些因素可能限制模型在其他地理区域、医疗系统或临床工作流程中的泛化性。尽管有几个表现优异的团队在训练时整合了其他公开的 CXR 数据集（如 CheXpert、PadChest、VinDr-CXR）以增强稳健性，但最终的评估和排行榜排名仅基于 MIMIC-CXR 测试数据。为了更严格地评估模型的泛化能力并确保更广泛的临床适用性，未来的挑战版应结合来自不同机构和人群的外部测试集。这将有助于更全面地评估跨站点的可迁移性，并揭示数据集转移或亚组特定偏差的潜在来源。

此外，解决模型中的偏差问题是至关重要的。现有研究表明，基于单一机构的 CXR 数据集训练的深度神经网络在预测性能上常常表现出与种族和性别等因素相关的差异。虽然 [8] 观察到在较大、多个机构的数据集上训练可以帮助减轻这些差异，但他们的工作主要集中在二分类任务上。迄今为止，没有研究专门检查长尾、多标签和零样本分类任务中的偏差。未来的研究可以探索应对这些挑战的方法，通过严格的亚组分析和多地点验证确保模型在不同人群和环境中既公平又具有广泛适用性。

与大多数公开的 CXR 基准数据集类似，CXR-LT 数据集受限于由自动提取的文本挖掘标签 [58] 造成的固有标签噪声。然而，随着大型语言模型（LLMs）的发展，最近的研究 [46] 表明，相较于传统的方法（如基于规则的方法），GPT-4 在 CXR 数据集中可能生成更准确的标签。在我们的研究中，表格 **??** 支持这一观察，显示 GPT-4 通过改进的 mAP 生成了更高质量的标签。其他 LLMs 在生成标签上也可能超越传统方法，呈现出一个有前景的途径来减少未来的标签噪声。CXR-LT 的未来迭代可能利用基于 LLM 的标签生成管道为任意大规模、长尾的 CXR 数据集生成结构化标签，这在最近的努力中已被证明成功 [59] 。此外，随着更多类别的加入，LLMs 的提示变得更长，这可能导致模型性能下降，并可能遗忘某些类别。为缓解这一问题，将标签任务重新框定为自然语言推理（NLI）问题并一次专注于一个类别的提示这可以作为一个有效的策略。此外，采用链式思考（CoT）提示等技术可以通过改善推理和响应生成来进一步提高性能。或者，可以利用知识图谱在应用基于 LLM 的标签生成之前将类别分成不同的子群体，从而提供一种结构化和系统的方法来解决这一挑战。

而且，由于放射科医生手动标注的成本和时间要求极高，获得足够样本进行深度学习训练具有挑战性 [60] ，但提供一个用于测试目的的“黄金标准”数据集仍是可行的。在这项工作中，我们利用并公开发布了这样一个具有更可靠标签的数据集。然而，该数据集是由研究生通过审阅临床报告文本

进行标注的。未来，该数据集可以通过放射科住院医生或主治医生达成共识重新标注来提高其质量。此外，手动标注一个外部数据集以进行验证也可以进一步增强所提出方法的评估，为其提供更可靠和准确的性能基准。如在 CXR-LT 2023 [2] 概述中指出的，零样本分类可以成为临床上可行的长尾医学图像识别的理想方法，使其能够适应任何新的发现。今年，任务 3 的前三名团队都使用了视觉-语言模型来应对这一挑战，凸显了其潜力。然而，在几个领域仍有显著的改进空间：更有效地对齐图像和文本表示，从文本数据中提取未见类别的信息，以及准确检测图像中的异常区域。此外，视觉-语言模型的高效微调或指令调优将是解决零样本疾病分类问题相关挑战的关键。

尽管最近在方法上有所进展，胸部 X 光片 (CXR) 疾病分类的平均精度 (mAP) 和 F1 分数仍然相对有限。这提出了一个关键问题：这些性能指标在临床上是否可以接受？例如，mAP 在 0.28 至 0.52 之间的范围内表明性能远高于随机机会，但可能无法满足自动临床使用所需的可靠性。有几个因素造成了这些局限性。首先，极端的类别不平衡——尤其是罕见发现的情况——会使性能偏颇并降低对较少代表性疾病的灵敏度。其次，训练和评估数据集由于弱监督或标注不一致，常常含有标签噪声。第三，胸部 X 光作为一种成像模式本质上缺乏分辨率或对比度，以清晰区分某些病理，特别是在视觉线索微妙或重叠的情况下。此外，大多数模型依赖于带有阈值的概率输出，这可能由于校准问题而影响决策的可靠性。为减轻这些挑战并提高实际应用的效用，未来的研究可以探讨多模态决策融合、校准置信度估计、临床医生参与验证以及主动学习技术以改善罕见类别的采样。此外，报告每个类别的指标并进行失效模式分析有助于将模型性能置于背景中，并指导在临床设置中更为明智的部署策略。

5. 结论

总而言之，我们组织了 CXR-LT 2024，以应对胸部 X 光片中长尾、多标签疾病分类和零样本学习的挑战。为此，我们策划并发布了一个大型的、长尾的、多标签 CXR 数据集，其中包含 377,110 张图像，每张图像标注了一组 45 种疾病类别中的一个或多个发现。此外，我们还提供了一个公开可用的“黄金标准”子集，其中包含人工标注的共识标签，以促进进一步评估。最后，我们概述了一条提高方法的可靠性、泛化能力和实用性的途径，最终目标是使它们能够应用于真实临床环境中。

6.

声明竞争利益

作者声明以下财务利益/个人关系可能被视为潜在的利益冲突：R.M.S 已收到来自 iCAD、Philips、PingAn、ScanMed、Translation Holdings 和 MGB 的专利或软件许可证的版税，以及来自与 PingAn 合作的 CRADA 的研究支持。其余作者声明，他们没有已知的可能影响本文所报道工作的竞财利益或个人关系。

7.

致谢

本工作得到了以下机构和项目的支持：美国国家医学图书馆 [资助号 R01LM014306]，国家科学基金会 [资助号 2145640, IIS-2212176]，亚马逊研究奖，康奈尔–香港科技大学全球战略合作奖，《人工智能杂志》，国家卫生研究院 [资助号 R01EB017205]，DS-I 非洲 [资助号 U54TW012043-01]，Bridge2AI [资助号 OT2OD032701]，以及通过 ITEST # 2148451 的国家科学基金会。此外，还得到了 NIH 内部研究项目、国家医学图书馆和临床中心的支持。

References

[1] Gregory Holste, Song Wang, Ziyu Jiang, Thomas C Shen, George Shih, Ronald M Summers, Yifan Peng, and Zhangyang Wang. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 22–32. Springer, 2022.

[2] Gregory Holste, Yiliang Zhou, Song Wang, Ajay Jaiswal, Mingquan Lin, Sherry Zhuge, Yuzhe Yang, Dongkyun Kim, Trong-Hieu Nguyen-Mau, Minh-Triet Tran, et al. Towards long-tailed, multi-label disease classification from chest x-ray: Overview of the cxr-lt challenge. *Medical Image Analysis*, page 103224, 2024.

[3] Yuxin Hong, Xiao Zhang, Xin Zhang, and Joey Tianyi Zhou. Evolution-aware VAriance (EVA) coreset selection for medical image classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 301–310, New York, NY, USA, 28 October 2024. ACM. doi: 10.1145/3664647.3681592.

[4] Evi M C Huijben, Josien P W Pluim, and Maureen A J M van Eijnatten. Denoising diffusion probabilistic models for addressing data limitations in chest X-ray classification. *Inform. Med. Unlocked*, 50(101575):101575, 1 January 2024. ISSN 2352-9148. doi: 10.1016/j.imu.2024.101575.

[5] Wongi Park and Jongbin Ryu. Fine-grained self-supervised learning with jigsaw puzzles for medical image classification. *Comput. Biol. Med.*, 174(108460):108460, May 2024. ISSN 1879-0534,0010-4825. doi: 10.1016/j.compbiomed.2024.108460.

[6] Yuhang Li, Tong Liu, Wenfeng Shen, Yangguang Cui, and Weijia Lu. Improving generalization and personalization in long-tailed federated learning via classifier retraining. In *European Conference on Parallel Processing*, pages 408–423. Springer, 2024.

[7] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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*, pages 2097–2106, 2017.

[8] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[9] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.

[10] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[11] Nils Rethmeier and Isabelle Augenstein. Long-tail zero and few-shot learning via contrastive pretraining on and for small data. In *Computer Sciences & Mathematics Forum*, volume 3, page 10. MDPI, 2022.

[12] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[13] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317, 2019.

[14] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

[15] David M Hansell, Alexander A Bankier, Heber MacMahon, Theresa C McLoud, Nestor L Muller, and Jacques Remy. Fleischner society: glossary of terms for thoracic imaging. *Radiology*, 246(3):697–722, 2008.

[16] Song Wang, Mingquan Lin, Ying Ding, George Shih, Zhiyong Lu, and Yifan Peng. Radiology text analysis system (radtext): Architecture and evaluation. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 288–296, 2022. doi: 10.1109/ICHI54592.2022.00050.

[17] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[18] Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6, 2023. URL `http://jmlr.org/papers/v24/21-1436.html`.

[19] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[20] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[21] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.

[22] Tianjie Dai, Ruipeng Zhang, Feng Hong, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Unichest: Conquer-and-divide pre-training for multi-source chest x-ray classification. *IEEE Transactions on Medical Imaging*, 2024.

[23] Dongkyun Kim. 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*, pages 2702–2710, 2023.

[24] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[25] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist' s annotations. *Scientific Data*, 9(1):429, 2022.

[26] Eduardo P Reis, Joselisa PQ De Paiva, Maria CB Da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, et al. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487, 2022.

[27] Yuexi Du, Brian Chang, and Nicha C Dvornek. 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*, pages 465–475. Springer, 2024.

[28] Yuexi Du, John Onofrey, and Nicha C Dvornek. Multi-view and multi-scale alignment for contrastive language-image pre-training in mammography. *arXiv preprint arXiv:2409.18119*, 2024.

[29] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021.

[30] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*, 2024.

[31] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–41, 2023.

[32] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[33] Houlsby Neil and Weissenborn Dirk. Transformers for image recognition at scale. *Online: https://ai. googleblog. com/2020/12/transformers-for-image-recognitionat. html*, 2020.

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[35] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022.

[36] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[37] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[38] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training in radiology. *arXiv preprint arXiv:2301.02228*, 2023.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[40] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[41] Pablo Messina, René Vidal, Denis Parra, Álvaro Soto, and Vladimir Araujo. Extracting and encoding: Leveraging large language models and medical knowledge to enhance radiological text representation. *arXiv preprint arXiv:2407.01948*, 2024.

[42] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[43] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[44] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022.

[45] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[46] Yishu Wei, Xindi Wang, Hanley Ong, Yiliang Zhou, Adam Flanders, George Shih, and Yifan Peng. Enhancing disease detection in radiology reports through fine-tuning lightweight llm on weak labels. *arXiv preprint arXiv:2409.16563*, 2024.

[47] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54 (10s):1–41, 2022.

[48] Dimitrios Pantelaios, Paraskevi-Antonia Theofilou, Paraskevi Tzouveli, and Stefanos Kollias. Hybrid cnn-vit models for medical image classification. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2024.

[49] Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022.

[50] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[51] Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, 137:109347, 2023.

[52] Haomin Chen, Shun Miao, Daguang Xu, Gregory D Hager, and Adam P Harrison. Deep hierarchical multi-label classification of chest x-ray images. In *International conference on medical imaging with deep learning*, pages 109–120. PMLR, 2019.

[53] Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2982–2990, 2022.

[54] Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 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*, pages 23–34, 2022.

[55] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multimodal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022.

[56] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[57] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[58] Mohamed Abdalla and Benjamin Fine. Hurdles to artificial intelligence deployment: Noise in schemas and "gold" labels. *Radiology: Artificial Intelligence*, 5(2):e220056, 2023.

[59] Qiaoyu Zheng, Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Lisong Dai, Hengyu Guan, Yuehua Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. Large-scale long-tailed disease diagnosis on radiology images. *Nature Communications*, 15(1):10147, 2024.

[60] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

Supplementary Material

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

Supplementary Table 1. Final test phase results of the CXR-LT 2024 competition for Task 1. The table presents the precision of the top-4 teams' final models for each of the 40 classes evaluated on the test set.

Disease	zguo	Tianjie_dai	XYPB	dongkyunk
Atelectasis	0.626	0.630	0.627	0.628
Cardiomegaly	0.669	0.683	0.678	0.675
Consolidation	0.250	0.259	0.253	0.251
Edema	0.551	0.551	0.553	0.547
Enlarged Cardiomediastinum	0.189	0.196	0.192	0.188
Fracture	0.380	0.364	0.328	0.367
Lung Lesion	0.068	0.069	0.067	0.062
Lung Opacity	0.628	0.627	0.629	0.633
Pleural Effusion	0.845	0.845	0.843	0.842
Pneumonia	0.335	0.336	0.342	0.339
Pneumothorax	0.608	0.597	0.597	0.618
Support Devices	0.927	0.925	0.927	0.926
Granuloma	0.317	0.269	0.296	0.342
Normal	0.331	0.333	0.335	0.334
Calcification of the Aorta	0.172	0.163	0.164	0.173
Emphysema	0.263	0.262	0.258	0.274
Fibrosis	0.169	0.167	0.169	0.138
Hernia	0.572	0.583	0.580	0.550
Infiltration	0.064	0.064	0.067	0.064
Mass	0.251	0.287	0.250	0.258
Nodule	0.246	0.233	0.232	0.254
Pleural Thickening	0.153	0.146	0.138	0.137
Pneumomediastinum	0.307	0.340	0.338	0.300
Pneumoperitoneum	0.299	0.299	0.351	0.286
Subcutaneous Emphysema	0.613	0.604	0.580	0.609
Tortuous Aorta	0.067	0.068	0.070	0.072
Adenopathy	0.103	0.136	0.129	0.112
Azygos Lobe	0.156	0.133	0.117	0.146
Clavicle Fracture	0.040	0.013	0.006	0.003
Fissure	0.223	0.198	0.174	0.197
Hydropneumothorax	0.179	0.185	0.173	0.142
Infarction	0.011	0.008	0.011	0.011
Kyphosis	0.086	0.086	0.107	0.073
Lobar Atelectasis	0.008	0.004	0.005	0.004
Pleural Other	0.077	0.066	0.067	0.071
Pulmonary Embolism	0.012	0.013	0.013	0.012
Pulmonary Hypertension	0.032	0.047	0.049	0.035
Rib Fracture	0.284	0.269	0.256	0.289
Round(ed) Atelectasis	0.053	0.028	0.057	0.041
Tuberculosis	0.062	0.060	0.073	0.064
Mean	0.281	0.279	0.277	0.277

Supplementary Table 2. Final test phase results of the CXR-LT 2024 competition for Task 1. The table presents the F1 of the top-4 teams' final models for each of the 40 classes evaluated on the test set.

Disease	zguo	Tianjie dai	XYPB	dongkyunk
Atelectasis	0.566	0.549	0.557	0.556
Cardiomegaly	0.541	0.517	0.543	0.543
Consolidation	0.259	0.275	0.275	0.275
Edema	0.453	0.440	0.468	0.455
Enlarged Cardiomediastinum	0.232	0.220	0.229	0.229
Fracture	0.347	0.320	0.359	0.324
Lung Lesion	0.136	0.156	0.139	0.117
Lung Opacity	0.560	0.550	0.565	0.574
Pleural Effusion	0.707	0.682	0.704	0.708
Pneumonia	0.346	0.338	0.347	0.347
Pneumothorax	0.491	0.500	0.493	0.477
Support Devices	0.841	0.813	0.841	0.845
Granuloma	0.346	0.347	0.393	0.358
Normal	0.360	0.359	0.379	0.373
Calcification of the Aorta	0.258	0.214	0.254	0.224
Emphysema	0.293	0.313	0.314	0.283
Fibrosis	0.245	0.255	0.244	0.225
Hernia	0.522	0.498	0.566	0.438
Infiltration	0.114	0.109	0.119	0.109
Mass	0.263	0.299	0.300	0.263
Nodule	0.260	0.228	0.259	0.242
Pleural Thickening	0.202	0.208	0.209	0.177
Pneumomediastinum	0.370	0.392	0.407	0.391
Pneumoperitoneum	0.444	0.443	0.481	0.386
Subcutaneous Emphysema	0.593	0.572	0.605	0.564
Tortuous Aorta	0.113	0.133	0.125	0.126
Adenopathy	0.151	0.156	0.172	0.146
Azygos Lobe	0.163	0.326	0.221	0.289
Clavicle Fracture	0.065	0.000	0.000	0.000
Fissure	0.257	0.230	0.247	0.217
Hydropneumothorax	0.313	0.270	0.258	0.231
Infarction	0.000	0.000	0.025	0.017
Kyphosis	0.167	0.123	0.125	0.129
Lobar Atelectasis	0.000	0.000	0.000	0.000
Pleural Other	0.117	0.121	0.125	0.152
Pulmonary Embolism	0.000	0.000	0.006	0.015
Pulmonary Hypertension	0.011	0.062	0.099	0.079
Rib Fracture	0.326	0.287	0.313	0.293
Round(ed) Atelectasis	0.000	0.000	0.043	0.081
Tuberculosis	0.143	0.136	0.162	0.151
Macro F1	0.289	0.286	0.299	0.285
Micro F1	0.471	0.455	0.474	0.462

Supplementary Table 3. Final test phase results of the CXR-LT 2024 competition for Task 1. The table presents the False Negative Rate (FNR) of the top-4 teams' final models for each of the 40 classes evaluated on the test set.

Disease	zguo	Tianjie dai	XYPB	dongkyunk
Atelectasis	0.069	0.052	0.059	0.556
Cardiomegaly	0.064	0.036	0.063	0.543
Consolidation	0.271	0.305	0.307	0.275
Edema	0.121	0.108	0.131	0.455
Enlarged Cardiomediastinum	0.168	0.083	0.139	0.229
Fracture	0.362	0.394	0.438	0.324
Lung Lesion	0.851	0.814	0.761	0.117
Lung Opacity	0.042	0.033	0.045	0.574
Pleural Effusion	0.061	0.048	0.059	0.708
Pneumonia	0.050	0.024	0.049	0.347
Pneumothorax	0.264	0.277	0.263	0.477
Support Devices	0.055	0.049	0.056	0.845
Granuloma	0.529	0.614	0.596	0.358
Normal	0.101	0.101	0.130	0.373
Calcification of the Aorta	0.468	0.331	0.420	0.224
Emphysema	0.440	0.467	0.472	0.283
Fibrosis	0.509	0.512	0.545	0.225
Hernia	0.379	0.329	0.378	0.438
Infiltration	0.802	0.854	0.768	0.109
Mass	0.560	0.530	0.621	0.263
Nodule	0.568	0.474	0.571	0.242
Pleural Thickening	0.554	0.557	0.679	0.177
Pneumomediastinum	0.497	0.527	0.510	0.391
Pneumoperitoneum	0.597	0.551	0.473	0.386
Subcutaneous Emphysema	0.109	0.087	0.127	0.564
Tortuous Aorta	0.836	0.713	0.784	0.126
Adenopathy	0.819	0.675	0.748	0.146
Azygos Lobe	0.849	0.472	0.660	0.289
Clavicle Fracture	0.958	1.000	1.000	0.000
Fissure	0.541	0.519	0.645	0.217
Hydropneumothorax	0.636	0.594	0.693	0.231
Infarction	1.000	1.000	0.986	0.017
Kyphosis	0.703	0.533	0.713	0.129
Lobar Atelectasis	1.000	1.000	1.000	0.000
Pleural Other	0.913	0.889	0.858	0.152
Pulmonary Embolism	1.000	1.000	0.996	0.015
Pulmonary Hypertension	0.994	0.958	0.891	0.079
Rib Fracture	0.499	0.517	0.526	0.293
Round(ed) Atelectasis	1.000	1.000	0.971	0.081
Tuberculosis	0.759	0.756	0.665	0.151
Macro FNR	0.525	0.495	0.520	0.466
Micro FNR	0.151	0.135	0.153	0.142

Supplementary Table 4. Final test phase results of the CXR-LT 2024 competition for Task 2. The table presents the precision of the top-4 teams' final models for each of the 26 classes evaluated on the golden standard test set.

Disease	XYPB	zguo	yangz16	YYama
Atelectasis	0.464	0.457	0.454	0.465
Calcification of the Aorta	0.680	0.643	0.566	0.633
Cardiomegaly	0.747	0.711	0.717	0.739
Consolidation	0.466	0.437	0.450	0.466
Edema	0.537	0.582	0.564	0.585
Emphysema	0.364	0.386	0.379	0.335
Enlarged Cardiomediastinum	0.347	0.346	0.348	0.339
Fibrosis	0.549	0.456	0.543	0.536
Fracture	0.538	0.622	0.619	0.492
Hernia	0.789	0.748	0.840	0.741
Infiltration	0.052	0.049	0.056	0.045
Lung Lesion	0.037	0.048	0.055	0.042
Lung Opacity	0.657	0.658	0.642	0.662
Mass	0.430	0.388	0.384	0.401
Normal	0.726	0.795	0.741	0.727
Nodule	0.326	0.430	0.335	0.376
Pleural Effusion	0.837	0.850	0.842	0.835
Pleural Other	0.339	0.262	0.259	0.307
Pleural Thickening	0.333	0.321	0.372	0.258
Pneumomediastinum	0.832	0.791	0.768	0.823
Pneumonia	0.127	0.165	0.125	0.149
Pneumoperitoneum	0.732	0.605	0.641	0.581
Pneumothorax	0.665	0.657	0.630	0.639
Subcutaneous Emphysema	0.849	0.802	0.760	0.783
Support Devices: 0.9643	0.964	0.961	0.944	0.958
Tortuous Aorta	0.302	0.319	0.245	0.317
Mean	0.527	0.519	0.511	0.509

Supplementary Table 5. Final test phase results of the CXR-LT 2024 competition for Task 2. The table presents the F1 of the top-4 teams' final models for each of the 26 classes evaluated on the golden standard test set.

Disease	XYPB	zguo	yangz16	YYama
Atelectasis	0.517	0.522	0.486	0.523
Calcification of the Aorta	0.628	0.543	0.000	0.548
Cardiomegaly	0.618	0.625	0.603	0.619
Consolidation	0.439	0.432	0.053	0.432
Edema	0.547	0.530	0.413	0.545
Emphysema	0.321	0.414	0.000	0.274
Enlarged Cardiomediastinum	0.437	0.415	0.032	0.433
Fibrosis	0.546	0.455	0.000	0.526
Fracture	0.552	0.583	0.154	0.544
Hernia	0.737	0.684	0.690	0.743
Infiltration	0.094	0.042	0.000	0.000
Lung Lesion	0.048	0.000	0.000	0.000
Lung Opacity	0.684	0.685	0.625	0.685
Mass	0.426	0.358	0.261	0.415
Normal	0.737	0.673	0.053	0.646
Nodule	0.364	0.375	0.000	0.310
Pleural Effusion	0.714	0.721	0.745	0.714
Pleural Other	0.240	0.000	0.000	0.095
Pleural Thickening	0.383	0.339	0.000	0.391
Pneumomediastinum	0.767	0.789	0.326	0.776
Pneumonia	0.114	0.112	0.000	0.118
Pneumoperitoneum	0.681	0.579	0.222	0.585
Pneumothorax	0.513	0.545	0.619	0.539
Subcutaneous Emphysema	0.808	0.808	0.720	0.842
Support Devices: 0.9643	0.877	0.875	0.884	0.873
Tortuous Aorta	0.174	0.150	0.000	0.143
Macro F1	0.499	0.471	0.260	0.474
Micro F1	0.564	0.562	0.530	0.562

Supplementary Table 6. Final test phase results of the CXR-LT 2024 competition for Task 2. The table presents the False Negative Rate (FNR) of the top-4 teams' final models for each of the 26 classes evaluated on the golden standard test set.

Disease	XYPB	zguo	yangz16	YYama
Atelectasis	0.016	0.032	0.456	0.016
Calcification of the Aorta	0.426	0.532	1.000	0.511
Cardiomegaly	0.019	0.025	0.484	0.025
Consolidation	0.176	0.189	0.973	0.203
Edema	0.040	0.040	0.683	0.040
Emphysema	0.552	0.379	1.000	0.655
Enlarged Cardiomediastinum	0.246	0.297	0.983	0.229
Fibrosis	0.455	0.545	1.000	0.545
Fracture	0.229	0.229	0.917	0.292
Hernia	0.263	0.316	0.474	0.316
Infiltration	0.750	0.917	1.000	1.000
Lung Lesion	0.833	1.000	1.000	1.000
Lung Opacity	0.010	0.010	0.367	0.015
Mass	0.350	0.400	0.850	0.450
Normal	0.054	0.081	0.973	0.135
Nodule	0.636	0.636	1.000	0.667
Pleural Effusion	0.011	0.005	0.240	0.005
Pleural Other	0.842	1.000	1.000	0.947
Pleural Thickening	0.591	0.545	1.000	0.591
Pneumomediastinum	0.200	0.200	0.800	0.257
Pneumonia	0.182	0.182	1.000	0.136
Pneumoperitoneum	0.333	0.542	0.875	0.500
Pneumothorax	0.020	0.082	0.388	0.020
Subcutaneous Emphysema	0.048	0.048	0.357	0.048
Support Devices: 0.9643	0.045	0.050	0.122	0.050
Tortuous Aorta	0.879	0.909	1.000	0.909
Macro FNR	0.316	0.365	0.767	0.368
Micro FNR	0.148	0.165	0.570	0.167

Supplementary Table 7. Final test phase results of the CXR-LT 2024 competition for Task 3. The table presents the precision of the top-3 teams' final models for each of the 5 classes evaluated on the test set.

Disease	yyge	zhangRuichi	pamessina
Bulla	0.013	0.011	0.038
Cardiomyopathy	0.036	0.007	0.065
Hilum	0.359	0.389	0.252
Osteopenia	0.034	0.037	0.036
Scoliosis	0.205	0.158	0.036
Mean	0.129	0.116	0.110

Supplementary Table 8. Final test phase results of the CXR-LT 2024 competition for Task 3. The table presents the F1 of the top-3 teams' final models for each of the 5 classes evaluated on the test set.

Disease	yyge	zhangRuichi	pamessina
Bulla	0.000	0.000	0.028
Cardiomyopathy	0.062	0.000	0.026
Hilum	0.000	0.000	0.289
Osteopenia	0.061	0.000	0.019
Scoliosis	0.252	0.176	0.107
Macro F1	0.075	0.035	0.094
Micro F1	0.047	0.022	0.182

Supplementary Table 9. Final test phase results of the CXR-LT 2024 competition for Task 3. The table presents the False Negative Rate (FNR) of the top-3 teams' final models for each of the 5 classes evaluated on the test set.

Disease	yyge	zhangRuichi	pamessina
Bulla	1.000	1.000	0.260
Cardiomyopathy	0.958	1.000	0.231
Hilum	1.000	1.000	0.584
Osteopenia	0.892	1.000	0.308
Scoliosis	0.762	0.896	0.616
Macro FNR	0.923	0.979	0.400
Micro FNR	0.974	0.989	0.580