在为监督分类任务训练人工智能（AI）模型时，隐含的假设是训练数据集提供的标签代表了真实情况。然而，大规模数据集的注释是有效训练深度模型所必需的，这些注释获取成本高昂，并且由于人为错误、主观或模糊的标记任务、或不准确的自动标记系统等因素，可能不完全可靠。错误标记的数据—— “标签噪声” ——已被证明会对泛化能力和训练动态产生负面影响，这是一个常常被忽略但可能广泛存在的问题。医学成像数据集可能尤其容易受到标签噪声的影响，因为在标注者之间的变异性、使用语言模型从自由文本放射学报告中自动提取标签，以及许多基于成像的诊断任务固有的复杂性是主要贡献因素。

标签噪声可能存在于整个数据集中，但也可能系统性地影响某些数据子集。例如，在资源匮乏的环境中使用低质量扫描仪 [11] 可能会产生更难以识别病变的图像。在主要服务于特定人口群体的医疗中心中，由经验较少的放射科医生进行的注释可能会导致影响该群体的诊断错误率更高 [18] 。人类决策偏见可能导致数据集中某些性别或种族亚群的误诊或漏诊率更高 [10] 。一个从放射学报告中提取疾病标签的语言模型可能在不同语言撰写的报告上表现不准确 [8] 。在上述任何一个例子中，在此类数据集上训练的人工智能模型可能会学习到图像特征和标签之间的不一致映射，这种映射系统性地影响特定子群体，可能导致数据中子群体之间的性能差异。

虽然在识别和解决医学影像数据集中标签噪声的问题上取得了一些进展 [14,16] ，但对于系统性子群标签噪声（即标签偏差）作为潜在公平性问题的影响却很少有关注 [13] 。在此背景下，我们的工作旨在研究子群标签偏差对深度学习模型的影响，使用乳腺 X 光片数据中的组织密度分类作为具有临床重要性应用的例子。具体来说，我们（1）考察在模拟的子群标签偏差下训练的深度学习模型的特征空间，说明子群大小和可分性如何影响学习到的表示，以及（2）展示取决于是否使用无偏验证集来定义分类阈值后，在子群分类性能上显著的差异。

1 材料和方法

2.1 数据集。我们在乳腺组织密度分类任务中调查了标签偏差，该任务使用了 EMory BrEast 成像数据集（EMBED）中的全场数字乳房 X 光照片，这些照片是由四所机构医院在七年期间获取的 [7] 。这个数据集经过筛选以去除左右不匹配、局部压缩和放大图像。仅使用来自被识别为女性且被分配单一 BI-RADS 组织密度标签为 A 、B 、C 或 D 的患者的内斜侧和头尾位图像。密度标签 A 对应最低密度（即，主要为脂肪组织），而 D 对应最高密度（即，主要为致密组织）。对于该分类任务，我们将密度标签二值化为 $\texttt{0} := \{A, B\}$ 和 $\texttt{1} := \{C, D\}$ ，并进行欠采样以平衡两个类。数据集进一步经过筛选，仅包括三个最大的成像系统制造商：Hologic（`HOLO` ）、GE Medical Systems（`GEMS` ）和 Fujifilm（`FUJI` ）。`HOLO` 构成了数据集的绝大多数，其次是 `GEMS` 和 `FUJI` （见表 1 ）。

2.2 子群标签偏差。我们将子群标签偏差定义为影响单个子群的系统性误标。这通过将一个子群中的图像的二值组织密度标签从类别 `1` 改为类别 `0` 来模拟（例如，来自某个特定制造商）。然而，我们只对组织密度类别为 C 的患者引入了标签偏差，因为我们认为这种方法更切合实际地用于潜在模糊案例（B / C ）的误标，而不是在类别极端（A / D ）。此外，由于数据集中许多患者由于后续检查而有多张图像，我们假设标签偏差影响到来自某个患者的所有图像。因此，在所有实验中，子群标签偏差被表示为在分配给某个子群的 30 % 的患者中，二

Table 1. 属于每个制造商子组的图像数量。

Manufacturer	Total	Percent of Dataset	Tissue Density			
			A	B	C	D
Hologic (HOLO)	170,995	89.0	18,367	67,124	76,720	8,784
GE Medical Systems (GEMS)	15,965	8.3	636	7,210	7,590	529
Fujifilm (FUJI)	5,130	2.7	425	2,283	2,260	162

值类别标签 `1` 改为 `0` ，这些患者的组织密度为 C 。我们分析了应用于三个制造商子群中每一个的标签偏差，以及在“伪子群”中，以评估子群可分性 [9] 在标签偏差情景中的作用（见第 3.1 节）。

2.3 模型和训练。使用 ResNet-18 [5] 将图像分类为二元组织密度类别。数据集被随机分为 60 % /20 % /20 %，分别用于训练、验证和测试，确保不同分组之间没有病人重叠。图像通过遮盖非组织区域和将像素值归一化到 [0,1] 的范围进行预处理。训练的数据增强包括伽马校正、亮度/对比度抖动和随机仿射变换。模型以学习率 1×10^{-5} 和类平衡的批量采样进行训练。在验证集中，经过 10 个周期后，具有最高接收器操作特征曲线下面积 (AUC) 的模型被评估。

2.4 特征检查。对于在每个标签偏置场景上训练的模型，我们按照 Glocker 等人描述的方法对学习到的特征进行了探索。简而言之，将测试集通过训练好的模型，提取倒数第二层的特征。使用主成分分析（PCA）进行降维。我们使用第一主成分模式（PC1）的核密度估计（KDE）图对这些特征进行可视化，因为它最能预测组织密度。

2.5 分类性能。我们计算了测试集的子组特定的真实（TPR）和假阳性率 (FPR)，使用一个由验证集上 10% 整体 FPR 定义的阈值和真实（清洁的）类别标签。由于验证数据在实际情况下也可能受到标签偏差的影响，我们比较了使用清洁标签和有偏标签选择的阈值之间的子组性能（如第 2.2 节所述）。

2 结果

3.1 子群可分性。在这项工作中，子群可分性被定义为深度学习模型检测图像属于哪个子群的能力 [9] 。为了确定子群的可分性，我们使用第 2.3 节中描述的相同方法训练了一个模型来对子群进行分类。每个制造商子群（`HOLO` 、`GEMS` 和 `FUJI` ）的" 一对其余" AUC 为 1.00，表明这些子群是完全可分的。为了与不可分情况进行比较，我们创建了三个大小相等的“伪子群”，其中子群标签是随机分配到数据集中的每个图像，对应于伪子群 `1`，`2`，和 `3` 。训练来分类这些子群的模型仅能达到 0.50 的随机水平 AUC 值，表明这些子群是不可分的（即，模型无法区分这三个伪子群）。

3.2 亚群标签偏差对学习特征的影响。模型在具有干净标签的数据集上进行训练，代表基线，以及在对伪亚群 `1`（`PS1`）和每个制造商亚群分别施加标签偏差的数据集上进行训练。图 1 中展示了在每个标签偏差场景（`PS1` , `FUJI` , `GEMS` 和 `HOLO` ）中，各亚群的学习特征的第一个主成分 (PC1) 的 KDE 图，其中叠加了干净标签基线的相应亚群特征分布。值得注意的是，即使模型仅训练以预测二元密度类别，在干净标签场景下，特征空间沿 PC1 自然地组织成与组织密度类别 A , B , C 和 D 相对应的相对对称的序列表现。

不可分的子群。当标签偏置应用于密度类别 C 在 PS1 中时，整个类别 1 的特征（即，密度类别 C 和 D ）发生位移并更加紧密地聚集到类别 0 附近。从图 1 A 中可以看到，与干净标签基线相比，类别 1 组织密度类别在分隔这两类的边界附近的浓度更高。在这种情况下，当标签偏置应用于一个不可分的子群（PS1 ）时，这种特征偏移在所有伪子群（PS1 、PS2 、PS3 ）中是一致的（参见图 1 A 子图）。

可分离子群。图 1 B 和 1 C 分别展示了标签偏倚对 FUJI 和 GEMS 的影响，二者都是完全可分离但处于少数的子群。在这里可以看到，特征偏移主要出现在具有标签偏倚的子群中。HOLO 的特征空间与干净标签的基准相似，而另一个少数子群的特征空间经历了一些小扰动，但没有经历像具有标签偏倚的子群那样显著的特征偏移。

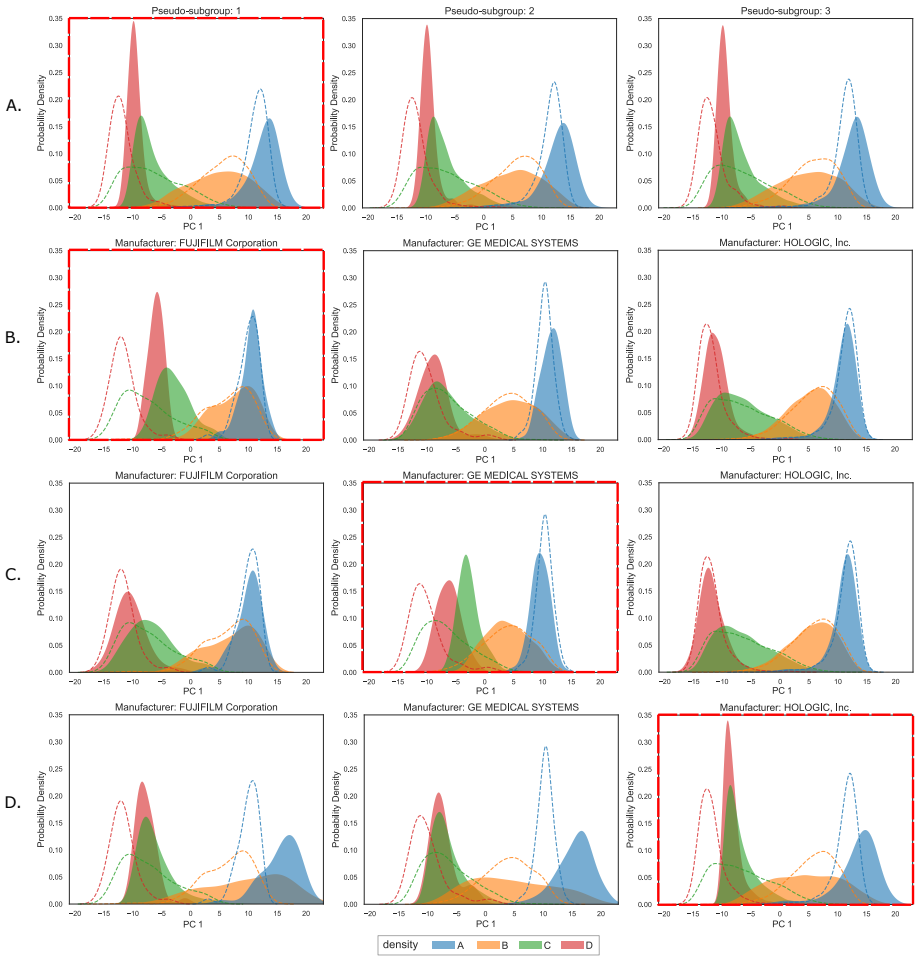

Fig. 1. 在每个标签偏差情境下特征空间的第一个主成分（PC1）上的组织密度类别分布。行 A：伪子组 1（`PS1` ）中的标签偏差，行 B：Fujifilm（`FUJI` ）中的标签偏差，行 C：GE Medical Systems（`GEMS` ）中的标签偏差，行 D：Hologic（`HOLO` ）中的标签偏差。虚线轮廓显示了从使用干净标签训练的模型中指定子组的组织密度类别分布。红色框架表示存在标签偏差的子组。

相反，当标签偏差应用于 `HOLO` 时，这是一个占据数据集大部分的可分离子群，全类别中带标签偏差的类别在 PC1 上的特征偏移在所有子群中都很明显（参见图 1 D）。特别是，虽然 `HOLO` 在接近类 `0` 的地方（以最锋利的峰值表示）具有最高浓度的类 `1` 特征，`GEMS` 和 `FUJI` 也出现 C 和 D 类别的特征向类 `0` 移动。在所有子群中，两个类 `1` 密度类别重叠，而不是来自清洁标签基线的序数模式。还可以看出，与其他标签偏差情景相比，类 `0` 组织密度类别的分布更为广泛。

3.3 子组标签偏差对性能的影响。在每个标签偏差场景中，三个制造商子组的 TPR 和 FPR 显示在表 2 中。重要的是，使用干净标签的验证集来定义操作点阈值，会导致与使用带有偏差标签的验证集不同的子组性能值，这取决于带有标签偏差的子组的可分性和大小。

Table 2. 在每个标签偏差场景下，每个子组的真正例率（TPR）和假正例率（FPR）。这些阈值是在验证集的总体假正例率为 10 % 时计算的，使用的是干净标签或有偏标签。括号中的值表示相对于干净基线的百分比变化。

Subgroup with Label Bias	Validation Set Labels	TPR			FPR		
		Non-Separable Subgroups					
		PS1	PS2	PS3	PS1	PS2	PS3
Clean (Baseline)		0.916	0.913	0.915	0.106	0.104	0.104
PS1	Clean	0.910 (-00.7 %)	0.912 (-00.1 %)	0.912 (-00.3 %)	0.106 (+00.0 %)	0.108 (+03.8 %)	0.104 (+00.0 %)
PS1	Biased	0.826 (-09.8 %)	0.828 (-09.3 %)	0.826 (-08.6 %)	0.042 (-60.4 %)	0.045 (-56.7 %)	0.048 (-53.8 %)
		Separable Subgroups					
		FUJI	GEMS	HOLO	FUJI	GEMS	HOLO
Clean (Baseline)		0.923	0.914	0.915	0.069	0.170	0.100
FUJI	Clean	0.750 (-18.7 %)	0.907 (-00.8 %)	0.914 (-00.1 %)	0.013 (-81.2 %)	0.158 (-07.1 %)	0.106 (+06.0 %)
GEMS	Clean	0.941 (+02.0 %)	0.780 (-14.7 %)	0.914 (-00.1 %)	0.092 (+33.3 %)	0.087 (-48.8 %)	0.103 (+03.0 %)
HOLO	Clean	0.938 (+01.6 %)	0.943 (-03.2 %)	0.898 (-01.9 %)	0.069 (+00.0 %)	0.228 (+34.1 %)	0.096 (-04.0 %)
FUJI	Biased	0.728 (-21.1 %)	0.904 (-01.1 %)	0.910 (-00.5 %)	0.013 (-81.2 %)	0.153 (-10.0 %)	0.102 (+02.0 %)
GEMS	Biased	0.931 (+00.9 %)	0.726 (-20.6 %)	0.901 (-01.5 %)	0.076 (+10.1 %)	0.068 (-60.0 %)	0.089 (-11.0 %)
HOLO	Biased	0.844 (-08.6 %)	0.828 (-09.4 %)	0.518 (-43.4 %)	0.015 (-78.3 %)	0.080 (-52.9 %)	0.012 (-88.0 %)

不可分离子群。使用干净标签在训练期间对 `PS1` 应用标签偏差来计算阈值时，所有子群的 TPR 和 FPR 与干净标签的基线非常相似。然而，当使用有偏标签计算此阈值时，所有子群的 TPR 下降了近 0.10，FPR 下降了 0.05 至 0.07。

可分离的子群。当标签偏差影响少数可分离子群（即，`FUJI` 或 `GEMS` ）时，无论阈值是在干净的还是在有偏标签的验证集上设置的，TPR/FPR 都会对受影响的子群减少。对于没有标签偏差的子群，表现相对保持在基线附近，尽管通常当使用有偏标签的验证集计算阈值时会略微低于基线。相反，清洁和有偏验证集之间的子群表现最大差异来自标签偏差影响到 `HOLO` ，即多数可分离子群。使用干净标签验证集时，`HOLO` 的 TPR 和 FPR 仅比基线略微减少。相比之下，当使用验证集中的有偏标签来计算分类阈值时，`HOLO` 的 TPR 降至 0.518，而 `GEMS` 和 `FUJI` 的 TPR 也减少了约 0.10，较使用干净标签时。所有可分离子群的 FPR 也大幅下降。

3 讨论

我们证明了，在二分类任务中训练的深度学习模型中，子组标签偏置的影响取决于受影响子组的可分性和相对大小。一般而言，标签偏置导致特征沿着 PC1 发生偏移，其中受标签偏置影响的类别的特征向另一个类别偏移。当标签偏置影响不可分的子组时，这种特征偏移发生在所有其他不可分的子组中。相反，当标签偏置应用于可分的少数子组时，这种特征偏移主要在受影响的子组中显现，可能是因为模型学习到这种“噪声”将图像映射到标签的方式仅与该特定子组的特征相关。然而，当标签偏置应用于多数可分子组时，所有子组都经历了特征

偏移，类似于不可分的情况。这可能是因为在这种情况下噪声标签在训练过程中被观察到的频率要高得多，从而有效地导致模型将这种噪声映射与整个数据集相关联。

我们推测，在 PC1 中观察到的特征变化可能也导致了模型高维决策空间中最佳类别分离阈值的变化。图 2 包含一个简单的玩具例子，说明了由标签偏差引起的特征和阈值变化如何帮助解释表格 2 中展示的小组性能的影响。

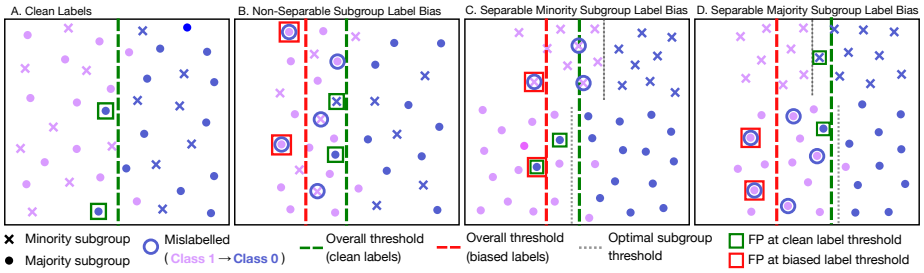

Fig. 2. 一个玩具示例说明了当模型的操作点阈值是通过验证集上的整体 FPR 选择时，标签偏倚引起的特征偏移如何影响子群的性能。

例如，考虑标签偏差影响非可分子群的情况（图 2 B）。在这种情况下，即使特征空间（以及相应的最优类别决策阈值）发生了变化，使用干净标签选择操作点阈值仍会带来较高的分类性能。然而，使用带有偏差标签的验证集会将选定的操作点阈值向类 1 移动，因为总的 FPR 计算会考虑到错误标记的图像。这将导致所有子群的 TPR 和 FPR 降低，类似于我们观察到的标签偏差影响 PS1 的情况。

在一个少数可分子群中存在标签偏差（如图 2 C），我们观察到该子群中的特征偏移尤为明显。然而，由于分类阈值是基于总体假阳性率（FPR），这一阈值主要由多数子群决定。因此，当使用干净的验证集时，少数子群的真正率（TPR）和假阳性率会下降，因为该子群的最佳阈值相比总体阈值会更接近类别 0 。使用带偏差的标签进行验证时，总体阈值会稍微更接近类别 1 ，导致所有子群的 TPR 和 FPR 进一步降低，正如我们观察到的，当标签偏差影响 GEMS 和 FUJI 时的情况一样。

在标签偏差影响主要子群的情况下（图 2 D），分类阈值会向类 0 移动，但对于主要群体的表现将与基线相似。假设对于主要子群的最佳阈值相较于少数子群的移动更大，少数群体的 TPR 和 FPR 将会增加。另一方面，使用有偏差的标签来计算整体的 FPR 会使操作点阈值在类 1 方向上移动得更远，以考虑主要子群中错误标记的图像。这将导致 TPR 和 FPR 戏剧性地下降，特别是在标签偏差应用于 HOLO 时，我们观察到了这种情况。

4 结论

医学影像数据集中的子群标签偏差的影响尚未被充分研究，但在 AI 公平性方面具有重要意义。已有研究表明，医学影像数据集中的社会人口统计学子群可能具

有不同的可分离性，例如，视网膜眼底影像中的性别可能具有相对较低的可分离性 [9] ，而胸部 X 光片中自我报告的种族子群则几乎完全可分离 [3] 。此外，在相对子群规模上常常存在显著差异，特别是在自我报告的种族类别中 [1] 。我们的结果表明，子群的可分离性和规模对带标签偏差下的群体公平性有不同的影响，因此在未来的研究中，考虑这两个因素都至关重要。从临床可用性的角度来看，当需要定义分类阈值时，我们的工作强调了理解用于验证模型的数据集是否受标签偏差影响的重要性。特别是，用偏见标签定义操作点可能会对一个或所有子群的性能产生重大影响，尤其是如果受影响的子群是绝大多数或子群无法分离。因此，未来的工作不仅应该进一步在受控场景中评估标签偏差的影响，还应该继续致力于开发可靠检测医学影像数据集中的标签偏差的方法。

局限性。重要的是要注意，我们假设数据集中的组织密度标签在初始时是干净的（即，没有显著的现存标签偏差）。此外，数据集中可能存在其他可以分离的子群体，这可能引入了交互效应。我们仅探讨了特征空间的第一主成分，因此，确认玩具示例中显示的趋势对应于模型的高维决策空间是未来工作的下一步。我们研究了由硬件相关差异定义的可分离子群体，并且仅考虑了一个方向上（例如，从类 `1` 到类 `0` ）静态量（30 %）的标签偏差。未来的工作应调查影响两个类别的不同程度的标签偏差，以及观察到的趋势是否适用于社会人口子群体中的标签偏差。

可重复性。用于数据处理、实验和分析的代码可以在 https://github.com/biomedia-mira/mammo-label-bias 上获得。

致谢。 E.A.M.S. 和 N.D.F. 感谢来自加拿大自然科学与工程研究委员会、Killam 信托、Alberta Innovates、加拿大研究主席项目以及 Calgary Foundation 的 River Fund 的支持。M.R. 由伦敦帝国学院院长博士奖学金和谷歌博士奖学金资助。R.M. 由欧盟 Horizon Europe 研究和创新项目资助，资助协议编号为 101080302。B.G. 从皇家工程院获得支持，作为其 Kheiron/RAEng 研究主席的一部分。

利益披露。 B.G. 是 DeepHealth 的兼职员工。没有其他竞争利益。

References

1. Chen, R.J., Wang, J.J., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., Mahmood, F.: Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering 7 (6), 719–742 (Jun 2023), publisher: Nature Publishing Group
2. Frenay, B., Verleysen, M.: Classification in the Presence of Label Noise: A Survey. IEEE Transactions on Neural Networks and Learning Systems 25 (5), 845–869 (May 2014)
3. Gichoya, J.W., Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L.C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.C., et al.: AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health 4 (6), e406–e414 (2022)
4. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Algorithmic encoding of protected characteristics in chest x-ray disease detection models. eBioMedicine 89 , 104467 (Mar 2023)

5. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (Jun 2016), iSSN: 1063-6919
6. Jain, S., Smit, A., Truong, S.Q., Nguyen, C.D., Huynh, M.T., Jain, M., Young, V.A., Ng, A.Y., Lungren, M.P., Rajpurkar, P.: VisualCheXbert: addressing the discrepancy between radiology report labels and image labels. In: Proceedings of the Conference on Health, Inference, and Learning. pp. 105–115. CHIL '21, Association for Computing Machinery, New York, NY, USA (Apr 2021)
7. Jeong, J.J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilies, G., Smith, G., Woo, M., McAdams, C.R., Newell, M.S., Banerjee, I., Gichoya, J., Trivedi, H.: The emory breast imaging dataset (embed): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. Radiology: Artificial Intelligence 5 (1), e220047 (Jan 2023)
8. Jones, C., Castro, D.C., De Sousa Ribeiro, F., Oktay, O., McCradden, M., Glocker, B.: A causal perspective on dataset bias in machine learning for medical imaging. Nature Machine Intelligence pp. 1–9 (Feb 2024), publisher: Nature Publishing Group
9. Jones, C., Roschewitz, M., Glocker, B.: The role of subgroup separability in group-fair medical image classification. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention –MICCAI 2023. p. 179–188. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2023)
10. Markowitz, D.M.: Gender and ethnicity bias in medicine: a text analysis of 1.8 million critical care records. PNAS Nexus 1 (4), pgac157 (Sep 2022)
11. Mollura, D.J., Culp, M.P., Pollack, E., Battino, G., Scheel, J.R., Mango, V.L., Elahi, A., Schweitzer, A., Dako, F.: Artificial Intelligence in Low- and Middle-Income Countries: Innovating Global Health Radiology. Radiology 297 (3), 513–520 (Dec 2020), publisher: Radiological Society of North America
12. Nichyporuk, B., Cardinell, J., Szeto, J., Mehta, R., Falet, J.P., Arnold, D.L., Tsaftaris, S.A., Arbel, T.: Rethinking generalization: The impact of annotation style on medical image segmentation. Machine Learning for Biomedical Imaging 1 , 1–37 (2022)
13. Petersen, E., Holm, S., Ganz, M., Feragen, A.: The path toward equal performance in medical machine learning. Patterns 4 (7), 100790 (Jul 2023)
14. Shi, J., Zhang, K., Guo, C., Yang, Y., Xu, Y., Wu, J.: A survey of label-noise deep learning for medical image analysis. Medical Image Analysis 95 , 103166 (Jul 2024)
15. Tong Xiao, Tian Xia, Yi Yang, Chang Huang, Xiaogang Wang: Learning from massive noisy labeled data for image classification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2691–2699. IEEE, Boston, MA, USA (Jun 2015)
16. Wei, Y., Deng, Y., Sun, C., Lin, M., Jiang, H., Peng, Y.: Deep learning with noisy labels in medical prediction problems: a scoping review. Journal of the American Medical Informatics Association 31 (7), 1596–1607 (Jul 2024)
17. Yang, F., Zamzmi, G., Angara, S., Rajraman, S., Aquilina, A., Xue, Z., Jaeger, S., Papagiannakis, E., Antani, S.K.: Assessing Inter-Annotator Agreement for Medical Image Segmentation. IEEE access : practical innovations, open solutions 11 , 21300–21312 (2023)
18. Zhang, L., Wen, X., Li, J.W., Jiang, X., Yang, X.F., Li, M.: Diagnostic error and bias in the department of radiology: a pictorial essay. Insights into Imaging 14 , 163 (Oct 2023)