基于注意力的显著性图提升

气胸分类的可解释性

“本文已获准在《放射学：人工智能》(https://pubs.rsna.org/journal/ai)上发表，该期刊由北美放射学会 (RSNA) 出版。” © 2023年 RSNA

稿件标题： 基于注意力的显著性图提升

气胸分类的可解释性

文章类型： 原创研究

作者：

- Alessandro Wollek，理学硕士*[1,2],- Robert Graf，理学硕士*[1,2]- Saša Čečatka，医学博士。[3]- Nicola Fink，医学博士。[3]- Theresa Willem，文学硕士。[4]- Bastian O. Sabel，医学博士。[3]- Tobias Lasser 博士 [1,2]

* Alessandro Wollek 和 Robert Graf 对本研究做出了同等贡献。[1] 慕尼黑生物医学工程研究所，慕尼黑工业大学，玻尔兹曼街 11 号，加兴（慕尼黑附近），85748，巴伐利亚，德国[2] 信息学系，慕尼黑工业大学，玻尔兹曼街 3 号，加兴（慕尼黑附近），85748，巴伐利亚，德国[3] 放射科，慕尼黑大学医院，马尔基奥尼尼街 15 号，慕尼黑，81377，巴伐利亚，德国[4] 慕尼黑社会技术学院，慕尼黑工业大学，阿尔齐斯街 21 号，慕尼黑，80333，巴伐利亚，德国

工作源自： 慕尼黑生物医学工程研究所，慕尼黑工业大学，玻尔兹曼街 11 号，加兴（慕尼黑附近），85748，巴伐利亚，德国。

通讯作者：

Alessandro Wollek

- 电话：+49 89 289 10840- 电子邮箱：alessandro.wollek@tum.de - 地址：德国巴伐利亚州加兴（慕尼黑附近）玻尔兹曼街 11 号，85748

经费：

本文研究获得德国联邦卫生部“改善医疗保健中以患者为中心的护理的数字创新计划”资助[, 资助协议编号：2520DAT920]。

数据共享协议：

所有使用的数据集均可公开访问：

- *CheXpert*： https://stanfordmlgroup.github.io/competitions/chexpert/ - *Chest X-Ray 14*： https://www.kaggle.com/datasets/nih-chest-xrays/data - *MIMIC CXR*： https://physionet.org/content/mimic-cxr-jpg/2.0.0/ - VinBigData： https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/ - SIIM ACR： https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation - 代码： https://gitlab.lrz.de/IP/publishing/attention-based-saliency-map

总结陈述：

在胸部X线片气胸分类任务中，视觉变换器的表现与卷积神经网络相当；其生成的基于注意力的显著性图也比梯度加权类激活图具有更强的可解释性。

要点：

- 在三个数据集的胸部X线片气胸分类任务中，视觉变换器（ViT）的表现与传统卷积神经网络（CNN）相近（ViT 的受试者工作特征曲线下面积（AUC）为 0.84 至 0.95，CNN 的曲线下面积为 0.83 至 0.87）。- 在放射科医师进行气胸诊断时，相较于梯度加权类别激活映射（GradCAM），更多病例（47%）被认为可从基于注意力的显著性图中受益（GradCAM 为 39%）。

- 在全部三个数据集和所有评估指标（扰动测试、Sensitivity-n、有效热度比、架构内可重复性和架构间可复现性）上，基于注意力的显著性图对模型决策的解释效果均优于GradCAM。

缩写：

AI	人工智能
CNN	卷积神经网络
ViT	视觉变换器
CXR	胸部X线片
TMME	Transformer-多模态可解释性
GradCAM	梯度加权类激活映射
AUC	受试者工作特征曲线下面积

摘要

目的：以气胸分类为例，研究视觉变换器（ViT）的胸部X线片（CXR）分类性能，以及基于注意力的显著性的可解释性。

材料与方法：在这项回顾性研究中，使用四个公开数据集对视觉变换器进行微调，以进行肺部疾病分类：CheXpert、Chest X-Ray 14、MIMIC CXR 和 VinBigData。采用变换器多模态可解释性和梯度加权类激活映射（GradCAM）生成显著性图。在 Chest X-Ray 14、VinBigData 和 SIIM-ACR 数据集上，通过受试者工作特征曲线下面积分析（曲线下面积，AUC）评估分类性能，并与卷积神经网络（CNN）进行比较。采用正向/负向扰动、Sensitivity-n、有效热度比、架构内可重复性和架构间可复现性评估可解释性方法。在用户研究中，三名放射科医师对160张胸部X线片进行气胸分类，并对是否使用显著性图两种情况下的分类结果进行比较，同时评定显著性图的有用性。

结果： 视觉变换器在胸部X线分类方面的曲线下面积与最先进的卷积神经网络相当：在 Chest X-Ray 14 上分别为 0.95（95%置信区间：0.943，0.950）和 0.83（95%，置信区间 0.826，0.842）；在 VinBigData 上分别为 0.84（95%置信区间：0.769，0.912）和 0.83（95%置信区间：0.760，0.895）；在 SIIM ACR 上分别为 0.85（95%置信区间：0.847，0.861）和 0.87（95%置信区间：0.868，0.882）。两种显著性图方法都揭示出模型对气胸引流管存在明显偏向。放射科医师认为，47%的基于注意力的显著性图和39%的 GradCAM 有用。在所有指标上，基于注意力的方法均优于 GradCAM。

结论：视觉变换器在胸部X线分类中的表现与卷积神经网络相似，其基于注意力的 显著性图对放射科医师更有用，且优于GradCAM。

引言

人工智能（AI）有望通过提升人类的表现和速度来改进医疗流程（1,2），但其黑箱特性给决策支持带来了巨大障碍（3,4）。当提供病变位置时，放射科医师能够最有效地确认或排除基于人工智能的识别结果，例如可能危及生命的气胸；而当相关区域仍需定位时，评估则会变得更加困难且耗时（5）。即使深度学习模型的表现与放射科医师相当（6,7），模型做出的错误决策在性质上也不同于人类专家所做的错误决策（7）。因此，使用此类模型的放射科医师需要适当的方法来了解其优势与局限。

显著性图通过突出显示与模型决策过程相关的位置来解决这一问题（**图 1**）。如果预测结果依据不足，例如仅凭胸管判断气胸，放射科医师便可以有把握地否定模型的预测结果。

当前最先进的图像分类器基于视觉Transformer（ViT）(8,9)。与卷积神经网络相比，ViT利用了在结构上更易于解释的注意力机制，因为该机制能够捕捉空间令牌关系（例如图像块）对于模型预测的重要性。我们使用变换器多模态可解释性（TMME）(10)创建了基于注意力的显著性图。TMME是滚动机制的扩展，即对所有注意力图进行矩阵乘法。通过取最小值或均值来合并注意力头。通过将单位矩阵添加到合并后的注意力矩阵中，考虑注意力层的跳跃连接。此外，根据所选类别，将所有注意力头与其导数相乘。在应用滚动机制之前，将负值设为零。作为基线，我们采用了常用的梯度加权类激活映射（GradCAM）技术(11)。

迄今为止，关于将显著性图用于医疗用途的研究仍然有限(12–18)，这些研究都侧重于比较显著性图与分割结果，而非解释模型作出决策的依据。有必要研究显著性图在胸部X线片（CXRs）上的可靠性，以及其对放射科医师的有用性。据我们所知， 尚未开展关于基于注意力的解释器的医学研究。

本探索性研究旨在以气胸分类为例，评估视觉Transformer用于胸部X线分类的可行性，以及基于注意力的显著性图在临床决策支持中的可解释性。

材料与方法

数据集

本回顾性研究使用了五个公开的胸部X线数据集：CheXpert (19)、Chest X-Ray 14 (20)、MIMIC CXR (21)、VinBigData (22)，以及医学影像信息学会—美国放射学会 (SIIM-ACR) (23) (**图2**)。CheXpert 包含 65,240 名患者的 224,316 张胸部X线片，Chest X-Ray 14 包含 30,805 名患者的 112,120 张胸部X线片，MIMIC CXR 包含 65,379 名患者的 377,110 张胸部X线片，VinBigData 包含 15,000 张带标签的胸部X线片，SIIM-ACR 则包含 2669 张气胸胸部X线片。本研究无需经过机构审查委员会审查，也无需获得患者知情同意。我们使用 CheXpert、Chest X-Ray 14、MIMIC CXR 和 VinBigData 进行训练，使用 CheXpert 的留出集进行验证，并在包含图像标注的数据集上进行测试：Chest X-Ray 14 的留出集、VinBigData 数据集中的 7500/15000 (50%)，以及 SIIM-ACR 的全部数据。所有数据集均在无患者重叠的情况下进行划分，且均可公开获取：

CheXpert：https://stanfordmlgroup.github.io/competitions/chexpert/Chest X-Ray 14：https://www.kaggle.com/datasets/nih-chest-xrays/dataMIMIC CXR：https://physionet.org/content/mimic-cxr-jpg/2.0.0/VinBigData：https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/SIIM-ACR：https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation

模型训练与开发

我们在重叠训练数据集中的以下类别上对ViT变体‘deit_base_distilled_patch16_224’(9)进行了微调：气胸、心脏肥大、肺实变、胸腔积液和肺不张（图2）。我们将最终的softmax层替换为sigmoid函数。我们采用随机梯度下降（SGD），学习率为0.001，批量大小为64，并训练了500轮，期间使用早停。我们对图像进行随机过采样。训练期间，我们采用随机仿射平移（-15°–15°，平移0.05，缩放0.9–1.05）和随机水平翻转。所有图像均按照ImageNet的均值和标准差进行归一化，并调整为224 x 224像素。我们将网络与预训练的DenseNet-121 (24)进行了比较，该模型常用于胸部X线分类（6,25）。我们的模型使用相同的数据和增强方法进行微调，采用学习率为0.001、批量大小为32的随机梯度下降，并训练了30轮，期间使用早停。

显著性图评估指标

正负扰动测试。 我们根据显著性图的重要性，持续移除输入图像中的图像块（*16x16*像素），随后生成新的显著性图。在正向扰动测试中，我们移除最重要的图像块；在负向扰动测试中，我们移除最不重要的图像块。移除图像块时，我们测量模型置信度，并计算曲线下面积。我们遵循了 Chefer等人 (28) 的流程，但没有将移除的图像块涂黑，而是用网络预测置信度为零的图像中的图像块替换它们，从而确保在替换大部分像素后，预测结果收敛至零，而非随机结果。***Sensitivity-n***。 Sensitivity-n 测试 (29) 会在输入图像上应用随机掩码。能够降低置信度的掩码应与优质显著性图相关。我们使用了 200 个随机掩码，每个掩码由*n* 个随机标记（*16x16* 像素）组成。我们将显著性图 × 掩码的 Hadamard积与置信度变化进行比较，计算皮尔逊积矩相关系数 (30)。对得分取平均值后，我们使用不同数量的标记*n*重复测试；这些数量在 1 到标记总数的一半之间按对数分布

标记。我们计算了标记数量与相关性得分之间的曲线下面积。我们使用了被视觉Transformer预测为气胸的测试图像。

有效热度比。我们使用有效热度比（EHR）（29）来表明显著性图突出显示的区域与先验医学知识一致。该测试首先根据阈值生成显著性图的二值掩膜。EHR 是真实标注内阈值区域与完整阈值区域之间的比例。阈值按等距步长计算。我们计算了所有 EHR 和阈值上的曲线下面积。**架构内可重复性和架构间可复现性**。遵循 Arun等人 （12），我们比较了不同模型显著性图之间的相似性。我们比较了 SIIM-ACR 数据集中 1000 张气胸胸部X线片显著性图的平均结构相似性指数（SSIM）。

用户研究

为评估临床实用性，我们邀请了慕尼黑路德维希-马克西米利安大学医院（以下简称“放射科医师”）的一位拥有十年以上经验的经认证的放射科医师（B.S.）、一位四年级放射科住院医师（N.F.）和一位一年级放射科住院医师（S.C.）参与调查。我们比较了有气胸和无气胸图像上的 GradCAM 与 TMME。由于可用的分割标注和研究参与时间有限，我们不得不将研究范围限制在气胸检测。我们从 SIIM-ACR数据集中抽取了 160 张图像，其中 110 张有气胸，50 张无气胸。我们纳入了 70 张 TMME 显著性图、70 张 GradCAM 显著性图、10 张基于分割标注生成的人工显著性图，以及 10 张通过反复应用高斯模糊并与柏林噪声相乘生成的随机显著性图（**表1**）。我们使用直方图分箱对模型预测进行校准（30）。在连续进行的两个部分中，放射科医师先根据胸部X线片和模型预测评估是否存在气胸，然后在额外提供显著性图后再次评估（**图3**）。此外，他们使用 1 至 5 分的量表对显著性图的实用性进行评分（非常不同意–非常同意）。

统计分析

我们使用受试者工作特征（ROC）、敏感度以及最大 F1 分数对应的特异度评估模型性能，并计算 95% 置信区间（采用非参数自助法，在图像级别进行 10,000 次重采样）。我们使用ROC曲线下面积（AUC）比较模型的分类性能。AUC 比较采用 DeLong 等人（32）提出的非参数方法的快速实现版本（31）。由于本分析属于探索性研究且涉及多重比较，因此提供 P 值，但不作显著性结论。我们使用敏感度和特异度衡量放射科医师在有无显著性图辅助时的气胸分类性能。本研究的统计分析使用 NumPy（https://numpy.org）1.21.5 版和 SciPy（https://scipy.org）1.7.3 版完成。

结果

胸部X线分类性能

由于本研究重点关注气胸，因此这里只简要报告其他类别（肺不张、心脏肥大、肺实变和胸腔积液）的性能结果。ROC 曲线及相应的曲线下面积和95%置信区间见补充材料（**图 S. 2: - 图 S. 5:** EH）。对于这些类别，在 Chest X-Ray 14 和 VinBigData 数据集上，视觉Transformer的表现均优于 DenseNet。就气胸分类而言，在 Chest X-Ray 14 上，视觉Transformer取得的曲线下面积高于 DenseNet（曲线下面积，0.95 [95 % CI：0.94，0. 95] vs. 0.83 [95 % CI：0.83，0.84]；p<0.001）。我们没有发现视觉Transformer与DenseNet 在 VinBigData 上的表现存在差异的证据（曲线下面积，0.84 [95% CI：0.77，0.91] vs. 0.83 [95% CI：0.76，0.90]；p = 0.67）。在 SIIM-ACR 数据集上，DenseNet 的气胸分类曲线下面积高于视觉Transformer（曲线下面积，0.87 [95% CI：0.87，0.88] vs. 0.85 [95% CI：0.85，0.86]；p<0.001；**图4**）。F1 分数、敏感度和特异度见**表 2**。

显著性图评估

以下测试用于衡量每种显著性方法在多大程度上能够突出显示对视觉Transformer预测结果至关重要的区域。在所有显著性图测试和数据集上，TMME 的表现均优于 GradCAM（**图 5)**）。

在正向扰动测试中，随着最重要的标记首先被移除，数值越低越好；在 SIIM-ACR 上，TMME 和 GradCAM 的测量值分别为 0.04（95%置信区间：0.036，0.044）和 0.12（95%置信区间：0.110，0.135）；在 Chest X-Ray 14 上分别为 0.03（95%置信区间：0.030，0.037）和 0.12（95%置信区间：0.111，0.137）；在 VinBigData 上分别为 0.02（95%置信区间：0.013，0.037）和 0.09（95%置信区间：0.048，0.135）。

负向扰动测试的结果是，随着最重要的标记最后被移除，数值越高越好；在 SIIM-ACR 上，TMME 为 0.67（95%置信区间：0.65，0.69），GradCAM 为 0.37（95%置信区间：0.34，0.39）；在 Chest X-Ray 14 上分别为 0.55（95%置信区间：0.52，0.57）和 0.25（95%置信区间：0.23，0.27）；在 VinBigData 上分别为 0.73（95%置信区间：0.60-0.85）和 0.55（95%置信区间：0.40-0.67）。TMME 与 GradCAM 对比的 Sensitivity-n（数值越高越好）分别为：在 SIIM-ACR 上为 14（95%置信区间：12.59，15.32）和 6（95%置信区间：4.44，6.57）；在 Chest X-Ray 14 上为 11（95%置信区间：9.85，13.14）和 4（95%置信区间：2.83，5.17）；在 VinBigData 数据集上为 9（95%置信区间：1.21，17.98）和 7（95%置信区间：0.55，13.00）。

在 SIIM-ACR 数据集上，TMME 的 EHR（越高越好）为 0.16（95%置信区间：0.142，0.171），而 GradCAM 为 0.11（95%置信区间：0.099，0.122）；在 Chest X-Ray 14 数据集上，TMME 为 0.26（95%置信区间：0.199，0.318），而 GradCAM 为 0.14（95%置信区间：0.100，0.193）；在 VinBigData 数据集上，TMME 为 0.33（95%置信区间：0.237，0.434），而 GradCAM 为 0.22（95%置信区间：0.126，0.275）。

就架构内可重复性而言，TMME 的平均结构相似性指数评分为 0.57（95%置信区间：0.562，0.578），而 GradCAM 的结构相似性指数评分为0.12（95%置信区间：0.105，0.126）。比较不同模型时，TMME 的架构间可复现性结构相似性指数评分为 0.47 （95%置信区间：0.465，0.481），而 GradCAM 为 0.08（95%置信区间：0.074，0.091）（**图 S. 6**）。

在对生成的显著性图进行目视评估时，我们发现，在某些图像中，GradCAM 和 TMME 显著性图突出显示的是混杂因素（例如胸管），而非气胸（**图 6**）。

用户研究

仅根据正位胸部X线片和视觉Transformer 的预测结果，放射科医师的敏感度为 64%（211/330；95% CI：59 %，69 %），特异度为 84%（126/150；95% CI：78 %，90 %）。在此基础上进一步显示显著性图后，敏感度提高至 65%（216/330；95% CI：60 %，70 %），特异度仍为 84%（126/150；95% CI：78 %，90 %）。排除模型预测错误的病例后，查看显著性图前的敏感度为 77%（162/210；95% CI：72 %，83 %），查看后为 79%（165/210；95% CI：73 %，84 %）；查看显著性图前的特异度为 93%（56/60；95% CI：86 %，98 %），查看后为 93%（56/60；95% CI：87 %，99 %）。显示 GradCAM 显著性图并未提高敏感度或特异度。TMME 显著性图将敏感度从 61%（83/135；95% CI：53 %，70 %）提高至 64%（86/135；95% CI：55 %，71 %）（**图 S. 7**）。

基于真实分割标注生成的人工显著性图，在 63%（19/30）的真阳性病例中被评为有用，而在大多数假阴性病例中被评为无用（26/30，87%）（**表 3**）。在真阳性预测中，TMME 和 GradCAM 分别有 47/90（52%）和 46/90（51%）被评为有用。总体而言，TMME 被评为有用的频率（99/210，47%）高于 GradCAM（82/210，39%），尤其是在模型错误预测中（TMME：41/90，45%；GradCAM：27/90，30%）。

讨论

在本研究中，我们在多个胸部X线数据集上训练了视觉Transformer，以生成基于注意力的显著性图，并评估其相较于 GradCAM 的表现，以及它们在协助放射科医师检测气胸方面的实用性。我们发现，尽管训练数据量有限，视觉Transformer 在胸部X线分类任务上的表现仍与卷积神经网络（CNN）相近（视觉Transformer 的曲线下面积为 0.84 至 0.95，CNN 为 0.83 至 0.87）。鉴于视觉Transformer 比 CNN 需要更多训练数据，这一发现值得进一步研究。

随着训练数据增多，我们预计视觉Transformer的表现将显著优于基于 CNN 的架构，这一点已在其他计算机视觉任务中得到证实（**图 S. 8**）。

此外，我们发现，基于注意力的 TMME 显著性图在所有指标上的表现均优于 GradCAM。放射科医师使用 TMME 显著性图检测气胸时，其有用性与使用 GradCAM 或不使用显著性图时存在差异，这进一步证实了这一结果，表明与 GradCAM 相比，基于注意力的显著性图更有助于临床决策支持。

我们认为，在某些情况下，GradCAM 和 TMME 的结果大相径庭，GradCAM 的可视化结果无法用于解读，这归因于两种方法的内部机制不同：GradCAM 使用单个层，而 TMME 则结合网络多个层的注意力，因此更加稳健。

显著性图可能会引发人们过度解读视觉结果的倾向（33），因此必须进行系统评估。利用显著性图可以发现模型或数据偏差，例如胸管可能暗示气胸。文献此前也观察到了这一现象（34），这凸显了提高此类模型可解释性的重要性。随后可以利用这些洞见来提升数据质量并改进架构选择。

尽管结果表明，放射科医师受到模型错误预测的影响，但显著性图提高了阅片者敏感度，并可用于防止确认偏误（35）。研究后讨论中，参与者表示，显著性图有助于他们更好地评估模型的优缺点，例如模型在预测气胸概率时对是否存在胸腔引流管的关注。尽管本研究的参与者人数较少，但基于注意力的显著性图能够提高阅片者敏感度，这一发现促使我们开展进一步研究。

我们的研究存在若干局限。首先，由于除气胸以外的肺部疾病缺乏分割掩膜，无法使用有效热度比对显著性图进行精确分析。在这些情况下，我们使用了现有但精度较低的边界框。其次，我们认为，模型性能（尤其是气胸分类性能）及由此生成的显著性图受到可用数据量较少的限制，即使汇集了多个机构的数据也是如此（**图 S. 8**）。已有研究表明，只要数据量充足，视觉变换器的表现就会大幅优于基于 CNN 的架构 (8)。

第三，尽管我们的研究表明，显著性图能够提高放射科医师气胸分类的敏感度，但研究结果受样本数量和参与者人数限制。要更全面地评估显著性图对放射科医师诊断性能的影响，需要纳入更多参与者，并采用结构化阅片以减轻确认偏误。这也有助于评估显著性图对不同专业水平的放射科医师是否有用。第四，遵循 Dosovitiskiy 等人 (8) 的方案，我们将放射影像调整为 224 x 224 像素，这可能导致细微的气胸无法检测到。

总之，我们探讨了基于视觉变换器的显著性图在胸部X线分类中的潜在用途。我们表明，视觉变换器取得了与传统卷积神经网络相近的结果。基于注意力的显著性图 TMME 的表现优于基线 GradCAM，支持在胸部X线分类中使用视觉变换器。放射科医师认为，在更多情况下，基于注意力的显著性图比 GradCAM 更有用。我们表明，显著性图能够帮助人工智能开发者和放射科医师更好地理解模型，从而有助于检测模型偏差。

致谢

本文研究获得德国联邦卫生部“改善医疗保健中以患者为中心的护理的数字创新计划”资助[资助协议编号：2520DAT920]。

参考文献

1. Rudin C. 停止为高风险决策解释黑箱机器学习模型，转而使用可解释模型。《Nature Machine Intelligence》。自然出版集团；2019；1(5)：206–215。doi: 10.1038/s42256-019-0048-x。 2. Reyes M, Meier R, Pereira S, 等. 放射学中人工智能的可解释性：挑战与机遇。《放射学：人工智能》。北美放射学会；2020；2(3)：e190043。doi: 10.1148/ryai.2020190043。 3. Doshi-Velez F, Kim B. 迈向严谨的可解释机器学习科学。arXiv预印本 arXiv:170208608。2017；doi: 10.48550/arXiv.1702.08608。 4. Alvarado R. 我们是否应该用深度学习取代放射科医师？鸽子、错误与医疗人工智能的可信度。《生物伦理学》。Wiley Online Library；2022；36(2)：121–133。doi: 10.1111/bioe.12959。 5. Brar MS, Bains I, Brunet G, Nicolaou S, Ball CG, Kirkpatrick AW. 隐匿性气胸究竟是真正隐匿，还是仅仅被漏诊：再议。《创伤杂志》。2010；69(6)：1335–1337。doi: 10.1097/TA.0b013e3181f6f525。 6. Rajpurkar P, Irvin J, Ball RL, 等. 用于胸部X线片诊断的深度学习：将CheXNeXt算法与执业放射科医师进行回顾性比较。《公共科学图书馆·医学》。公共科学图书馆，旧金山（加利福尼亚州，美国）；2018；15(11)：e1002686。doi: 10.1371/journal.pmed.1002686。 7. Majkowska A, Mittal S, Steiner DF, 等. 使用深度学习模型解读胸部X线片：采用放射科医生裁定的参考标准和人口调整评估进行评估。《放射学》。北美放射学会；2020；294(2)：421–431。doi: 10.1148/radiol.2019191293。 8. Dosovitskiy A, Beyer L, Kolesnikov A, 等. 一张图像胜过 16×16 个词：用于大规模图像识别的Transformer。国际学习表征会议。2020。 9. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. 通过注意力机制实现训练数据高效型图像Transformer与蒸馏。国际机器学习会议。机器学习研究论文集；2021。第 10347–10357 页。 10. Chefer H, Gur S, Wolf L. 用于解读双模态及编码器-解码器Transformer的通用注意力模型可解释性。IEEE/CVF国际计算机视觉会议论文集。2021。第 397–406 页。 11. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM：通过基于梯度的定位，从深度网络生成视觉解释。IEEE国际计算机视觉会议论文集。2017。第 618–626 页。 12. Arun N, Gaw N, Singh P, 等. 评估显著性图在医学影像中定位异常的可信度。《放射学：人工智能》。北美放射学会；2021；3(6)：e200267。doi: 10.1148/ryai.2021200267。 13. Jin W, Fatehi M, Abhishek K, Mallya M, Toyota B, Hamarneh G. 胶质瘤影像中的人工智能：挑战与进展。《神经工程学杂志》。IOP出版社；2020；17(2)：021002。doi: 10.1088/1741-2552/ab8131。

14. Jin W, Li X, Hamarneh G. 一张图并非适用于所有情况：评估多模态医学图像上的显著性图解释。arXiv预印本arXiv:210705047。2021；doi: 10.48550/arXiv.2107.05047。15. Eitel F, Soehler E, Bellmann-Strobl J, et al. 使用逐层相关性传播揭示卷积神经网络在常规磁共振成像上诊断多发性硬化的决策。NeuroImage: Clinical。Elsevier；2019；24:102003。doi: 10.1016/j.nicl.2019.102003。16. Ayhan MS, Kümmerle LB, Kühlewein L, et al. 用于理解眼科学中深度神经网络的显著性图的临床验证。Medical Image Analysis。Elsevier；2022；102364。doi: 10.1016/j.media.2022.102364。17. Zhang J, Petitjean C, Yger F, Ainouz S. 用于从超声图像估计胎儿头围的回归卷积神经网络可解释性。医学图像计算中的可解释与高效标注学习。Springer；2020。p. 73–82。18. Young K, Booth G, Simpson B, Dutton R, Shrapnel S. 深度神经网络还是皮肤科医生？用于临床决策支持的医学图像计算与多模态学习中机器智能的可解释性。Springer；2019。p. 48–55。19. Irvin J, Rajpurkar P, Ko M, et al. CheXpert：具有不确定性标签和专家比较的大型胸部X线片数据集。AAAI人工智能会议论文集。2019。p. 590–597。doi: 10.1609/aaai.v33i01.3301590。20. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. ChestX-ray8：医院规模的胸部X线数据库，以及常见胸部疾病弱监督分类和定位的基准。2017 IEEE计算机视觉与模式识别会议（CVPR）。2017。p. 3462–3471。21. Johnson AE, Pollard TJ, Greenbaum NR, et al. MIMIC-CXR-JPG：一个大型公开可用的带标注胸部X线片数据库。arXiv预印本arXiv:190107042。2019；doi: 10.48550/arXiv.1901.07042。22. Nguyen HQ, Lam K, Le LT, et al. VinDr-CXR：一个带有放射科医生标注的开放胸部X光片数据集。arXiv预印本arXiv:201215029。2020；doi: 10.48550/arXiv.2012.15029。23. Filice RW, Stein A, Wu CC, et al. 使用NIH胸部X线数据集上的机器学习标注进行气胸标注众包。Journal of digital imaging。Springer；2020；33(2):490–496。doi: 10.1007/s10278-019-00299-9。24. Huang G, Liu Z, van der Maaten L, Weinberger KQ. 密集连接卷积网络。2017。p. 4700–4708。25. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet：利用深度学习在胸部X光片上实现放射科医生级肺炎检测。arXiv预印本arXiv:171105225。2017；doi: 10.48550/arXiv.1711.05225。26. Chefer H, Gur S, Wolf L. 超越注意力可视化的Transformer可解释性。IEEE/CVF计算机视觉与模式识别会议论文集。2021。p. 782–791。27. Ancona M, Ceolini E, Öztireli C, Gross M. 迈向更好地理解深度神经网络的基于梯度的归因方法。arXiv预印本arXiv:171106104。2017；doi: 10.48550/arXiv.1711.06104。

28. Schober P, Boer C, Schwarte LA. 相关系数：恰当的使用与解读。《麻醉与镇痛》。Wolters Kluwer；2018；126(5):1763–1768。doi: 10.1213/ANE.0000000000002864。29. Zhang Y, Khakzar A, Li Y, Farshad A, Kim ST, Navab N. 通过识别具有预测信息的输入特征进行细粒度神经网络解释。《神经信息处理系统进展》。2021；34。30. Zadrozny B, Elkan C. 从决策树和朴素贝叶斯分类器获得校准概率估计。ICML。CiteSeer；2001。p. 609–616。31. Sun X, Xu W. 用于比较相关受试者工作特征曲线下面积的DeLong算法快速实现。《IEEE信号处理快报》。2014；21(11):1389–1393。doi: 10.1109/LSP.2014.2337313。32. DeLong ER, DeLong DM, Clarke-Pearson DL. 比较两条或多条相关受试者工作特征曲线下面积：一种非参数方法。《生物统计学》。JSTOR；1988；837–845。33. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. 显著性图的合理性检验。《神经信息处理系统进展》。2018；31。34. Rueckel J, Trappmann L, Schachtner B, et al. 干扰性胸腔管和胸膜裂开范围对人工智能气胸检测在胸部中的影响

t

放射影像。《放射学研究》。2020；55(12)：792–798。doi：10.1097/RLI.0000000000000707 .

35. Lee CS、Nagy PG、Weaver SJ、Newman-Toker DE。导致放射学诊断错误的认知与系统因素。《美国放射学杂志》。《美国伦琴射线学会》；2013；201(3)：611–617。doi：10.2214/AJR.12.10375。

表格

表1：用户研究中测试图像的分布

模型预测结果 / 显著性图	True 正例	假 阴性	True 阴性	假 阳性
GradCAM	30	15	10	15
TMME	30	15	10	15
人工	10	10	0	0

注。—GradCAM = 梯度加权类激活映射，TMME = 变换器多模态可解释性

表 2： 视觉Transformer（ViT）与 DenseNet 在

胸部X线片气胸分类中的性能比较

数据集	模型	F1分数	敏感度	特异度
SIIM-ACR				
	ViT	***66 %** [65 % - 68 %]	***83 %** [82 %-85 %]	81 % [80 %-82 %]
	DenseNet	64 % [63 %- 66 %]	75% [71 %-78%]	***84 %** [82 %-86 %]
Chest X-Ray 14				
	ViT	***67 %** [65 %-68 %]	***71 %** [68 %- 73 %]	***95 %** [95 %-96 %]
	DenseNet	44 % [43 %-46 %]	55% [50 %- 63%]	89 % [86 %-91 %]
VinBigData				
	ViT	***40 %** [26 %-56 %]	***26 %** [13 %-40 %]	**100 %** [100 %-100 %]
	DenseNet	19% [11 %-33%]	19 % [5 %-30 %]	100 % [98 %-100 %]

注.— 最大化的 F1分数以及由此得到的敏感度和特异度。括号中的数据为 95%置信区间。两个模型中数值较高
SIIM-ACR = 医学影像信息学会-美国放射学会

表 3： C 用于评估显著性图对检测气胸有用性的累计投票数
在用户研究中由三位放射科医师进行。

ViT 预测	方法	无用	两者皆非	有用
假阴性				
	人工	26	0	4
	GradCAM	23	7	15
	TMME	*18	5	*22
假阳性				
	GradCAM	29	4	12
	TMME	*22	4	*19
真阴性				
	GradCAM	20	1	9
	TMME	*12	7	*11
真正例				
	人工	*2	9	19
	GradCAM	37	7	46
	TMME	34	9	*47

注.— Most 有用且至少不有用投票数为标记带有 *。GradCAM = 梯度-加权类别激活映射，TMME = 变换器多模态可解释性，ViT = 视觉 transforme r

表格

表1： 用户研究测试图像的分布.17

表 2： 视觉Transformer（ViT）与 DenseNet 在气胸分类中的性能比较.18

表 3： 用户研究中三位放射科医师对显著性图有用性的累计投票.19

图

图 1： 可解释的肺部疾病分类流程示意图.21

图2： 本研究使用的数据集、数据划分和类别.22

图3： 研究中供放射科医师使用的用户界面.22

图4： 气胸分类的受试者工作特征（ROC）曲线及ROC曲线下面积.22

图5： 气胸预测的显著性图性能比较.23**图 6：** 来自 SIIM ACR 的视觉Transformer气胸预测图像及 TMME可视化.24

补充图

图 S. 1: 基于正确证据和错误证据的气胸预测显著性图.25

图 S. 2： 五个类别的受试者工作特征曲线.27

图 S. 3： 所有 5 个类别的正负扰动测试.29

图 S. 4： 所有 5 个类别的 Sensitivity-n 测试.30

图 S. 5： 所有 5 个类别的 EHR 测试.31

图 S. 6： 架构内可重复性和架构间可复现性.32**图 S. 7：** 用户研究的 ROC 曲线和模型曲线下面积.33

图 S. 8：增加训练数据对 SIIM ACR 测试数据集上气胸分类的影响.34

图注

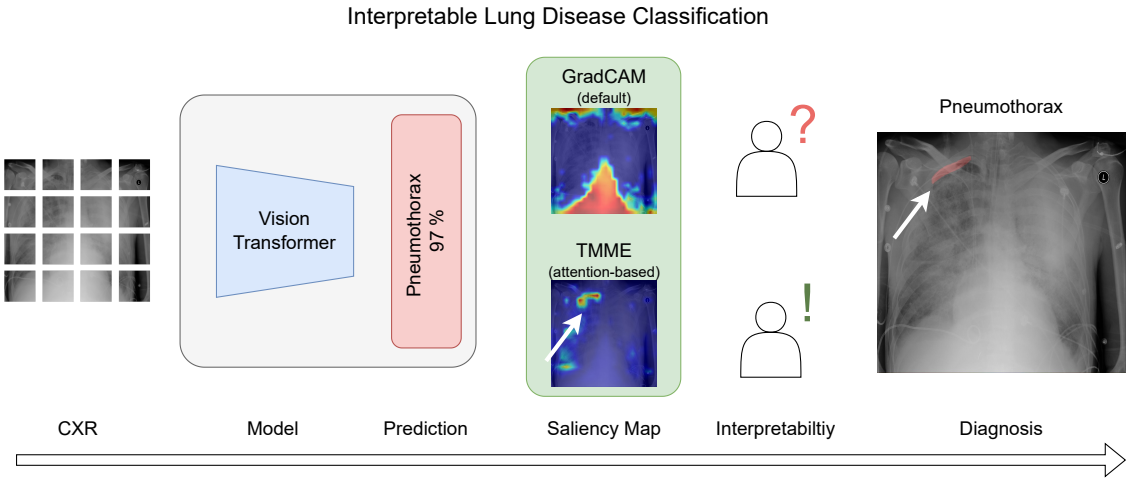


图 1：本研究采用的可解释肺部疾病分类流程示意图。使用视觉Transformer网络对胸部X线片（CXR）进行分类后，会生成基于注意力的显著性图——Transformer 多模态可解释性（TMME），以帮助放射科医师解读模型的预测结果。将基于 TMME 的显著性图与传统的梯度加权类激活图（GradCAM）进行了比较。

Datasets

Training

CheXpert		Chest X-Ray 14		MIMIC CXR		VinBigData	
Default Training Set		Default Training Set		Default Training Set		7500 / 15000 (50 %)	
223414 CXRs		86524 CXRs		353592 CXRs		7500 CXRs with Bounding Boxes	
Pneumothorax	22593	Pneumothorax	2637	Pneumothorax	15177	Pneumothorax	53
Cardiomegaly	35087	Cardiomegaly	1707	Cardiomegaly	67718	Cardiomegaly	1129
Consolidation	42525	Consolidation	2852	Consolidation	19741	Consolidation	172
Pleural Effusion	97815	Pleural Effusion	8659	Pleural Effusion	80401	Pleural Effusion	507
Atelectasis	67115	Atelectasis	8280	Atelectasis	76236	Atelectasis	85

Validation

CheXpert	
Default Validation Set	
234 CXRs	
Pneumothorax	8
Cardiomegaly	68
Consolidation	33
Pleural Effusion	67
Atelectasis	80

Test

Chest X-Ray 14				VinBigData		SIIM ACR	
Default Test Set				Rem. 7500 / 15000 (50 %)		All	
25596 CXRs		with Bounding Boxes		7500 CXRs with Bounding Boxes		2669 CXRs with Segmentations	
Pneumothorax	2665	Pneumothorax	98	Pneumothorax	43	Pneumothorax	618
Cardiomegaly	1069	Cardiomegaly	146	Cardiomegaly	1171		
Consolidation	1815	Consolidation	0	Consolidation	181		
Pleural Effusion	4658	Pleural Effusion	153	Pleural Effusion	525		
Atelectasis	3279	Atelectasis	180	Atelectasis	101		

图2：本研究使用的数据集、数据划分和类别。CXR = 胸部X线片，SIIM-ACR = 医学影像信息学会-美国放射学会

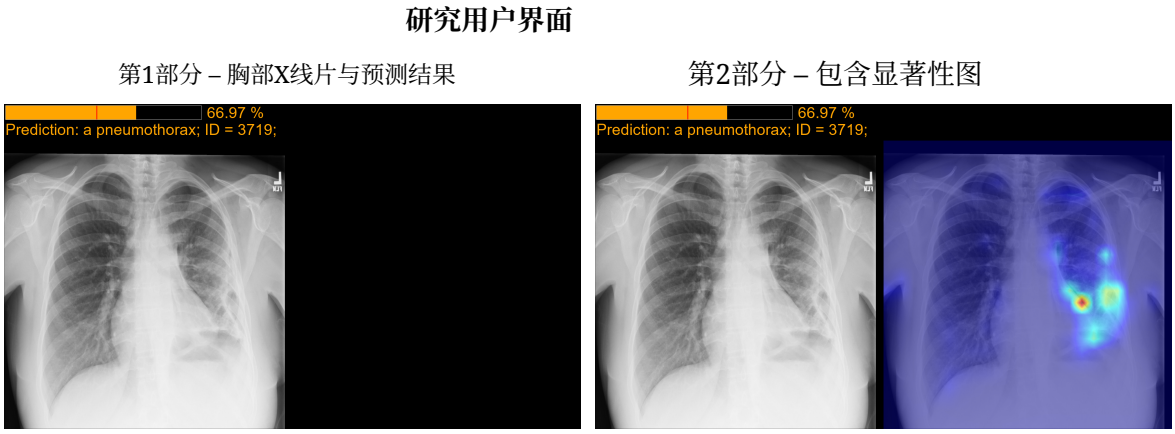


图3：研究中供放射科医师使用的用户界面。首先，向他们展示了胸部

放射影像（CXR），其中包括存在和不存在气胸的影像，以及视觉Transformer（ViT）的预测分数。第二步中，还额外显示了显著性图。在这两个环节中，放射科医师都需要检测是否存在气胸，然后判断显著性图是否（主观上）有助于检测。

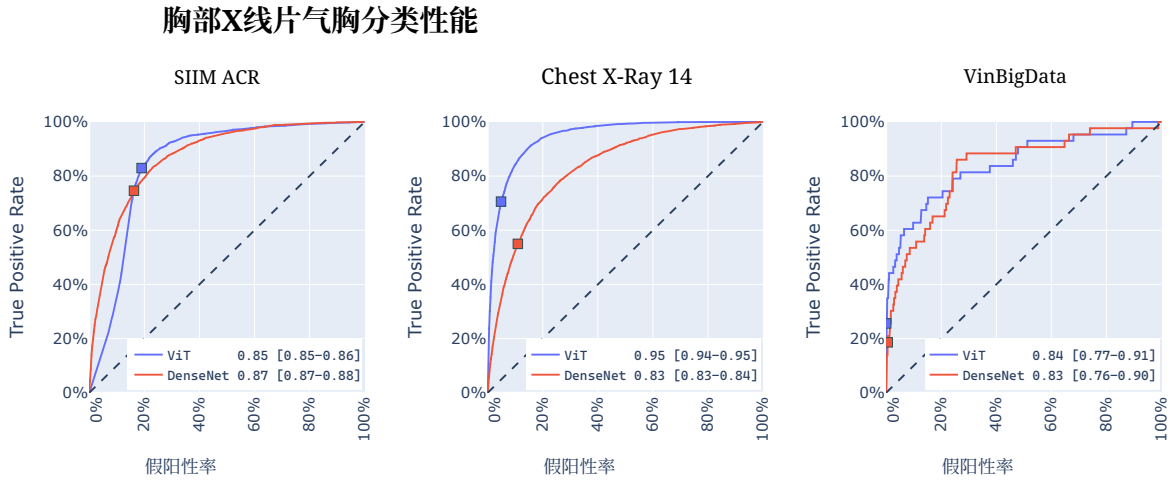


图4：受试者工作特征（ROC）曲线及ROC曲线下面积

[95% 置信区间]： 视觉Transformer（ViT）和DenseNet模型的气胸分类。

ROC曲线根据留出集计算。圆点表示最大化的F1分数。

SIIM-ACR = 医学影像信息学会-美国放射学会

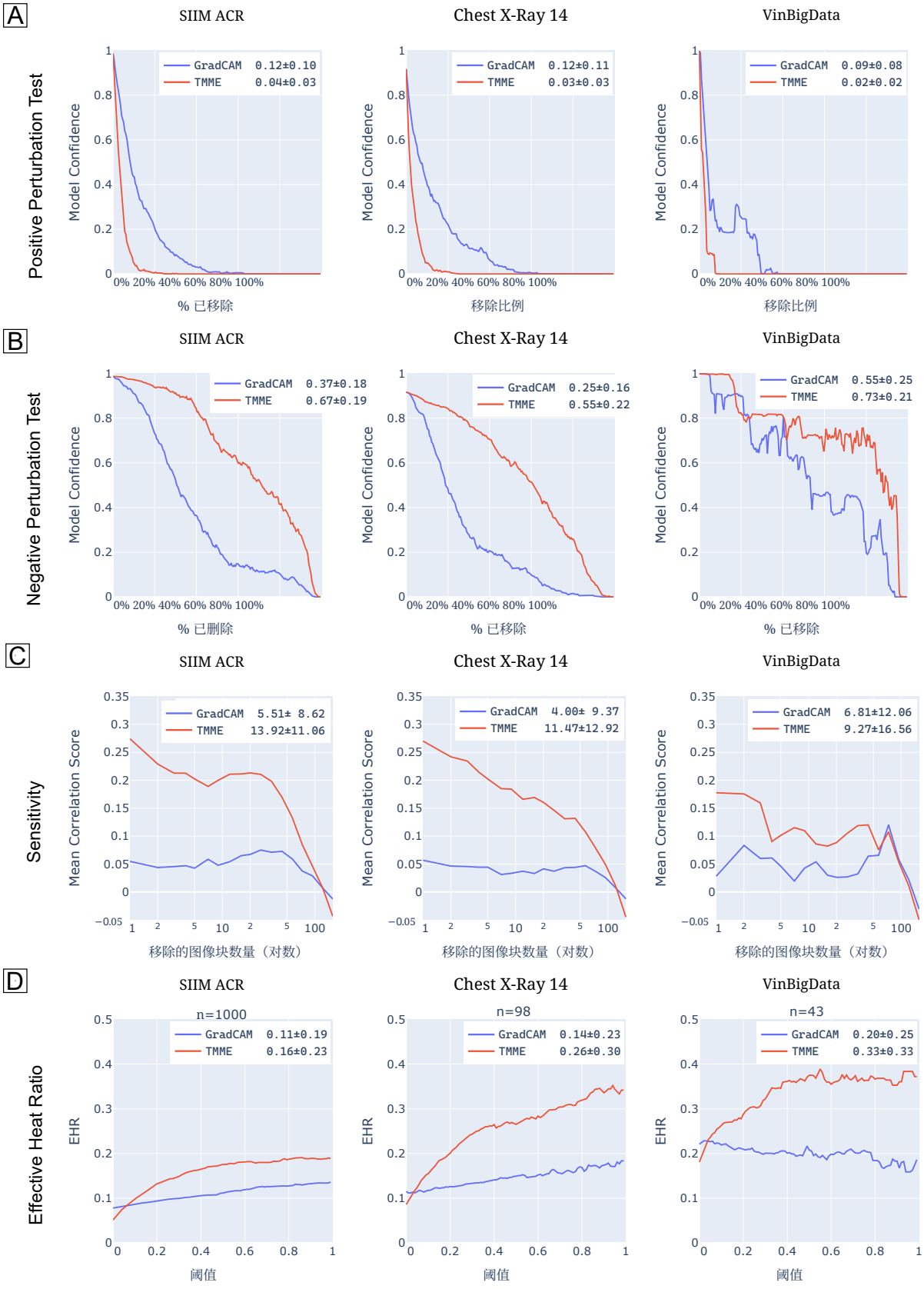


图 5： 气胸预测的显著性图性能比较。在所有 I

测试和数据集中，变换器多模态可解释性（TMME）的表现显著

优于梯度加权类激活映射（GradCAM）；显著性指标值 ±标准差值为相应曲线下的积分，并在各图中给出。**A:**正向扰动测试；值越低越好。**B:** 负向扰动测试；值越高越好。**C:** Sensitivity-n；值越高越好。**D:** 有效热度比（EHR）；值越高越好。所有类别的结果见**图 S. 2:** - **图 S. 5:** EH。SIIM-ACR =医学影像信息学会-美国放射学会

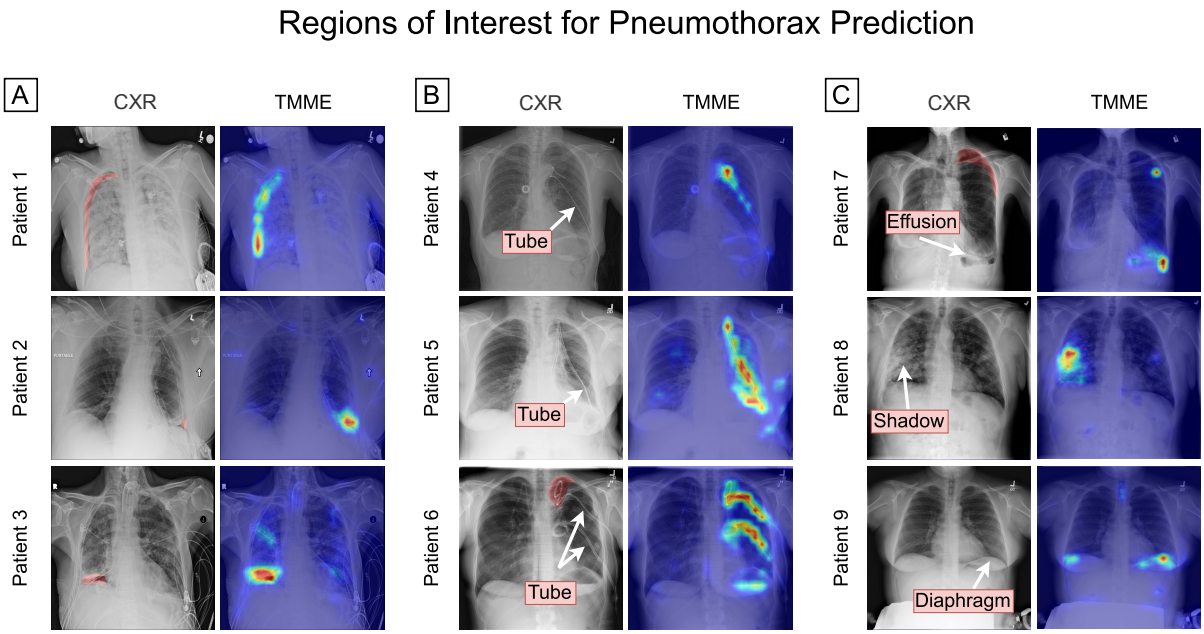


图 6: SIIM-ACR 中视觉变换器气胸预测图像，使用

Transformer 多模态可解释性（TMME）可视化。真实标注中的气胸分割以红色突出显示。没有气胸的假阳性病例未进行图像修复。**A:** TMME 突出显示气胸的示例。**B:** 突出显示胸管的示例。**C:** 根据其他病变（多发性积液、肺部阴影）或胸膈进行气胸预测。胸部X线片 = 胸部X线片，SIIM-ACR = 医学影像信息学会-美国放射学会

补充材料

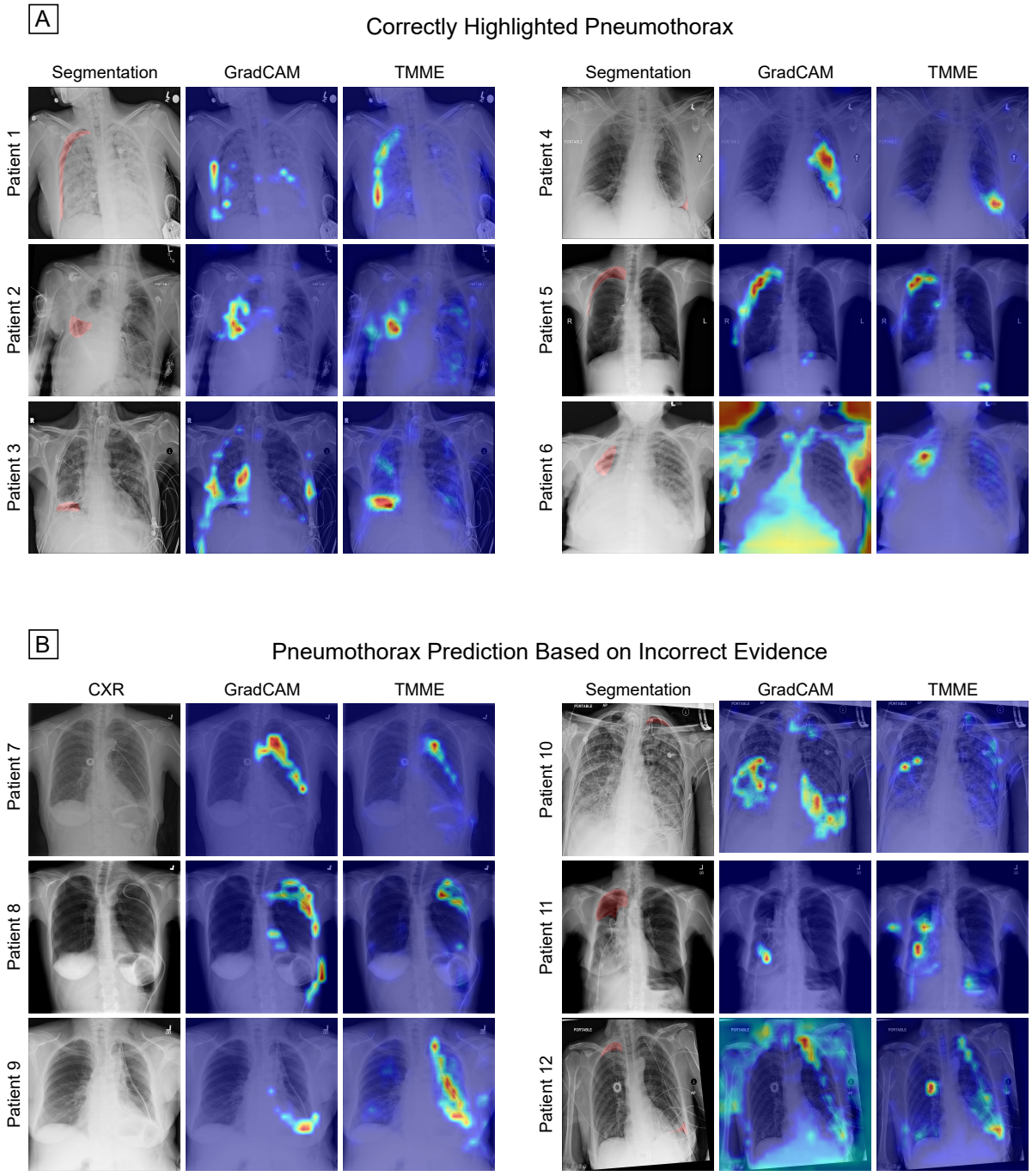


Figure S. 1: 基于临床正确且.....的气胸预测结果显著性图 d

错误证据。**A**：TMME 突出显示气胸的真正例示例。**B**：TMME 突出显示气胸错误证据的假阳性（左）和真正例（右）。我们随机抽取了 50 个真正例预测结果和 25 个假阳性预测结果来生成这些图像。

胸部X线分类性能对比

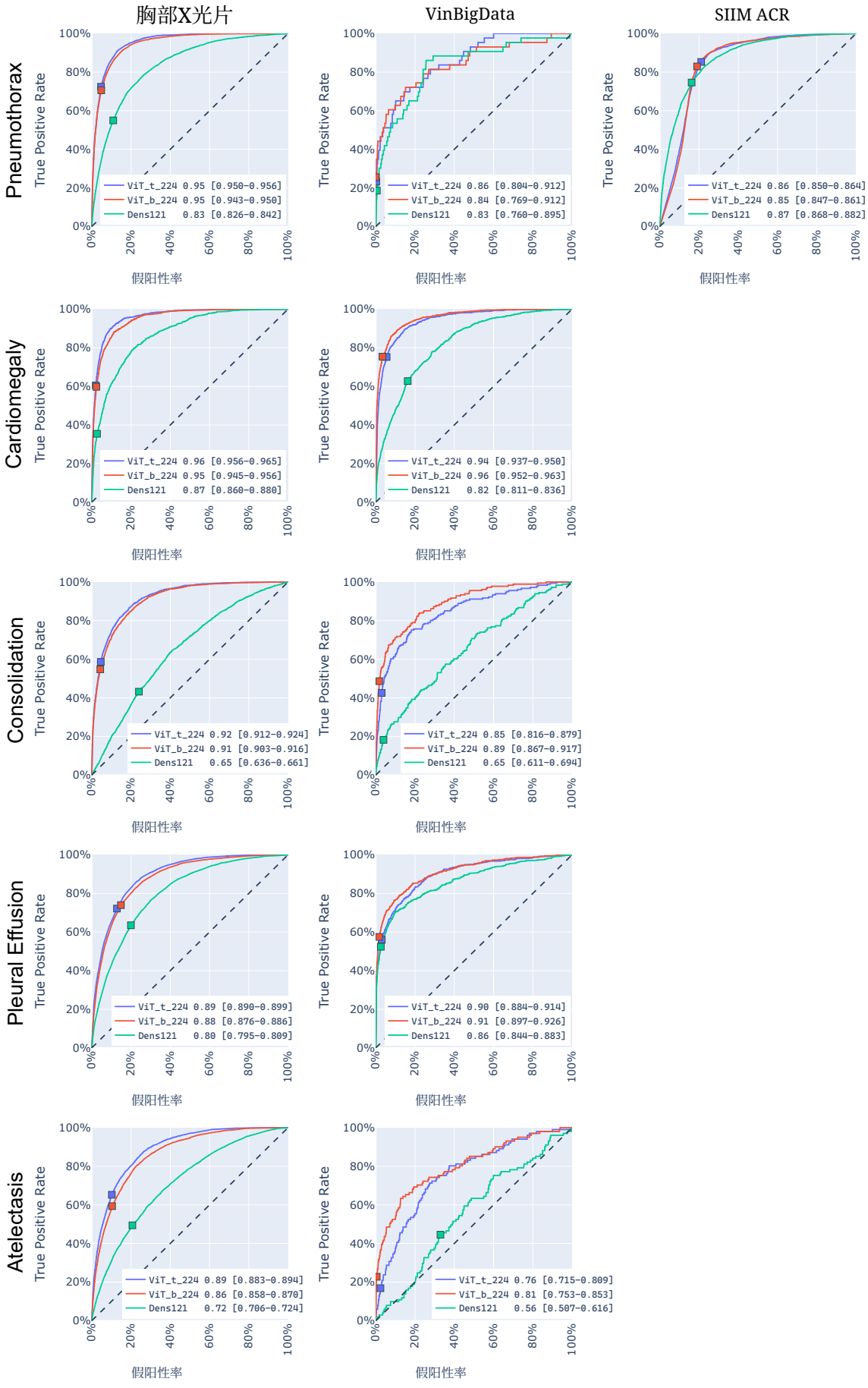
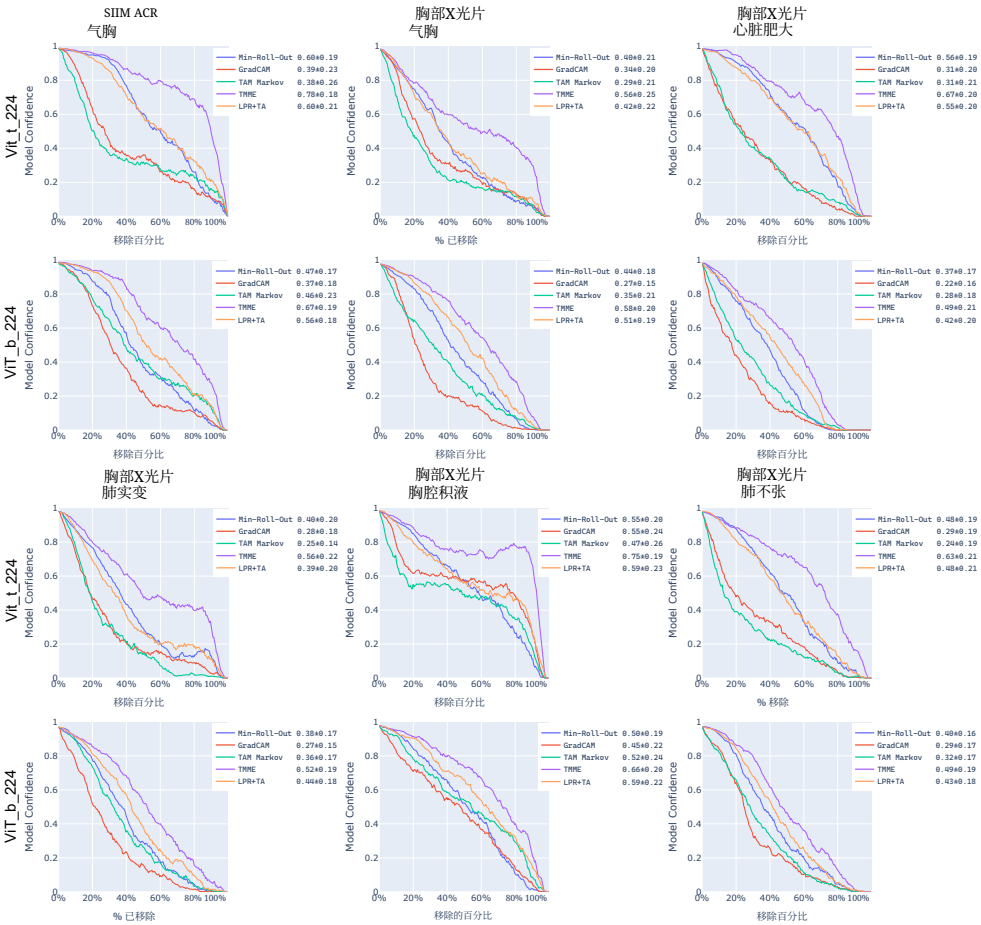

胸部X光片
VinBigData
SIIM ACR
Pneumothorax
Cardiomegaly
Consolidation
Pleural Effusion
Atelectasis
True Positive Rate
假阳性率
ViT_t_224 0.95 [0.950-0.956]
ViT_b_224 0.95 [0.943-0.950]
Dens121 0.83 [0.826-0.842]
ViT_t_224 0.86 [0.804-0.912]
ViT_b_224 0.84 [0.769-0.912]
Dens121 0.83 [0.760-0.895]
ViT_t_224 0.86 [0.850-0.864]
ViT_b_224 0.85 [0.847-0.861]
Dens121 0.87 [0.868-0.882]
ViT_t_224 0.96 [0.956-0.965]
ViT_b_224 0.95 [0.945-0.956]
Dens121 0.87 [0.860-0.880]
ViT_t_224 0.94 [0.937-0.950]
ViT_b_224 0.96 [0.952-0.963]
Dens121 0.82 [0.811-0.836]
ViT_t_224 0.92 [0.912-0.924]
ViT_b_224 0.91 [0.903-0.916]
Dens121 0.65 [0.636-0.661]
ViT_t_224 0.85 [0.816-0.879]
ViT_b_224 0.89 [0.867-0.917]
Dens121 0.65 [0.611-0.694]
ViT_t_224 0.89 [0.890-0.899]
ViT_b_224 0.88 [0.876-0.886]
Dens121 0.80 [0.795-0.809]
ViT_t_224 0.90 [0.884-0.914]
ViT_b_224 0.91 [0.897-0.926]
Dens121 0.86 [0.844-0.883]
ViT_t_224 0.89 [0.883-0.894]
ViT_b_224 0.86 [0.858-0.870]
Dens121 0.72 [0.706-0.724]
ViT_t_224 0.76 [0.715-0.809]
ViT_b_224 0.81 [0.753-0.853]
Dens121 0.56 [0.507-0.616]

图 S. 2： 五个类别的受试者工作特征曲线（ROC）：

气胸、心脏肥大、肺实变、胸腔积液和肺不张。SIIM-ACR数据集仅包含气胸标签。ViT_b_244 （红色）是正文中报告的ViT。在原型开发期间，我们还测试了较小的版本ViT_t_244（蓝色）。ROC曲线下面积和95%的置信区间已在图例中列出。SIIM- ACR = 医学影像信息学会-美国放射学会。

显著性图比较：负向扰动



显著性图比较：正向扰动

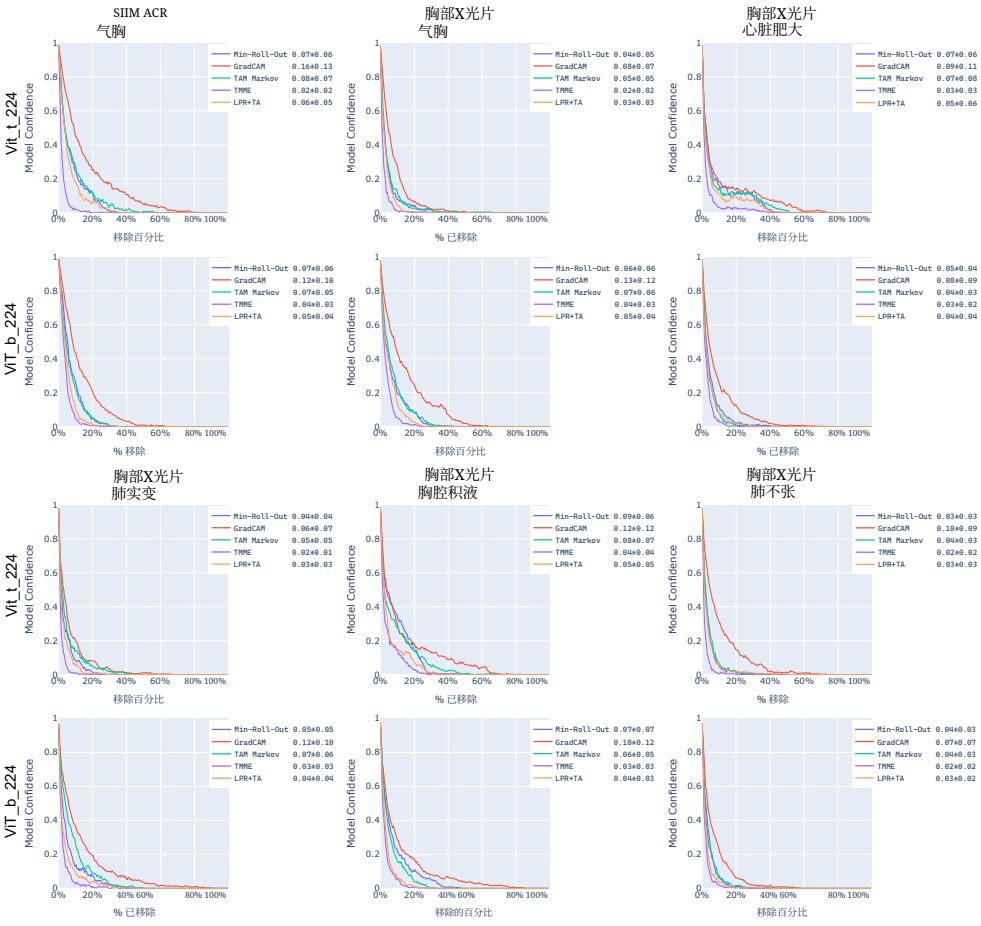


图 S. 3：对全部 5 个类别进行正负扰动测试：气胸、心脏肥大、肺实变、胸腔积液和肺不张。ViT_b_244 与正文中报告的 ViT 相同，ViT_t_244 则是在原型开发过程中使用的较小版本。在原型开发过程中，我们还研究了其他表现较差的基于注意力的显著图方法：min-roll-out、马尔可夫转移注意力图（TAM）和逐层相关性传播变换器归因（LRP+TA）。图例中报告了受试者工作特征曲线下面积和标准差。结果是对 250 个真阳性示例取平均所得。

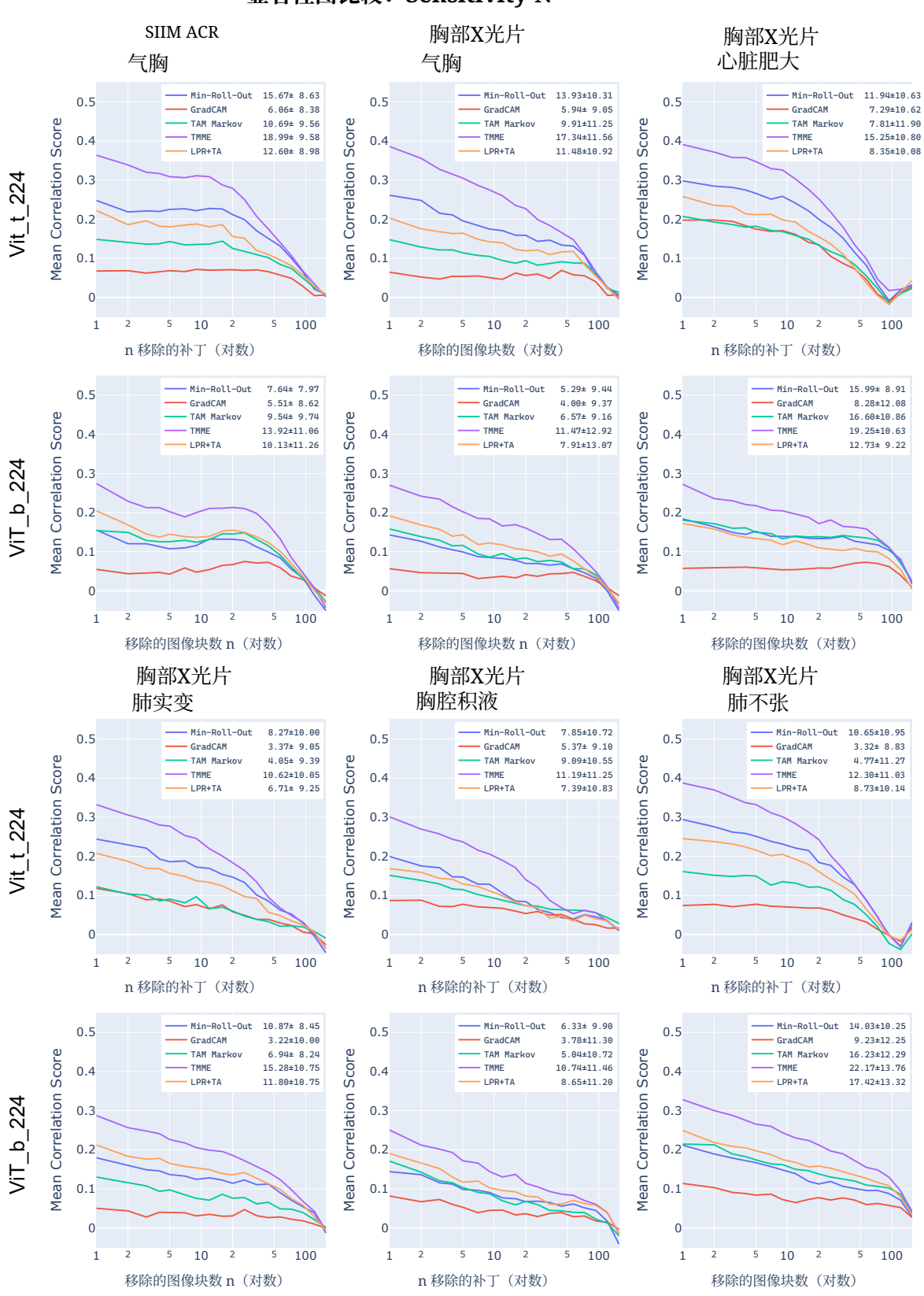


图 S. 4: 对全部 5 个类别进行的 Sensitivity-n 测试：气胸、心脏肥大、实变 n, 胸腔积液和肺不张。ViT_b_244 与正文中报告的 ViT 相同，ViT_t_244 是

原型开发期间使用的较小版本。在原型开发过程中，我们研究了其他表现较差的基于注意力的显著性图方法：min-roll-out、转移注意力图（TAM）马尔可夫方法，以及逐层相关性传播变换器归因（LRP+TA）。图例中报告了受试者工作特征曲线下面积和标准差。

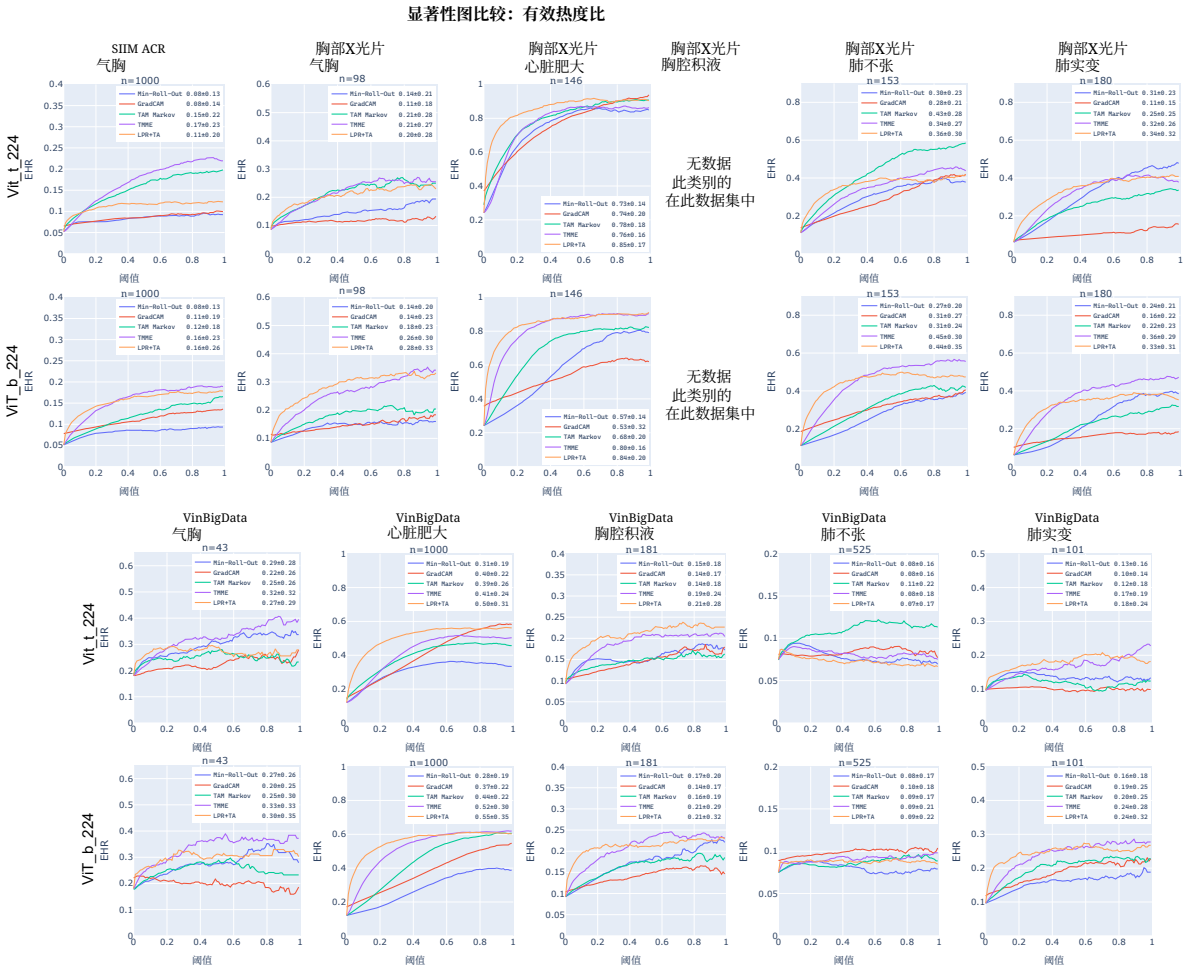


图 S. 5：对全部 5 个类别进行 EHR 测试：气胸、心脏肥大、肺实变、胸膜积液和肺不张。ViT_b_244 是文中报告的同一款 ViT，而 ViT_t_244 是原型开发期间使用的较小版本。在原型开发期间，我们还研究了其他性能较差的基于注意力的显著图方法：min-roll-out、转移注意力图（TAM）Markov 和逐层相关性传播变换器归因（LRP+TA）。图例中报告了受试者工作特征曲线下面积和标准差。结果是对 n 个分割样本取平均值，其中 n 的数值见图示。

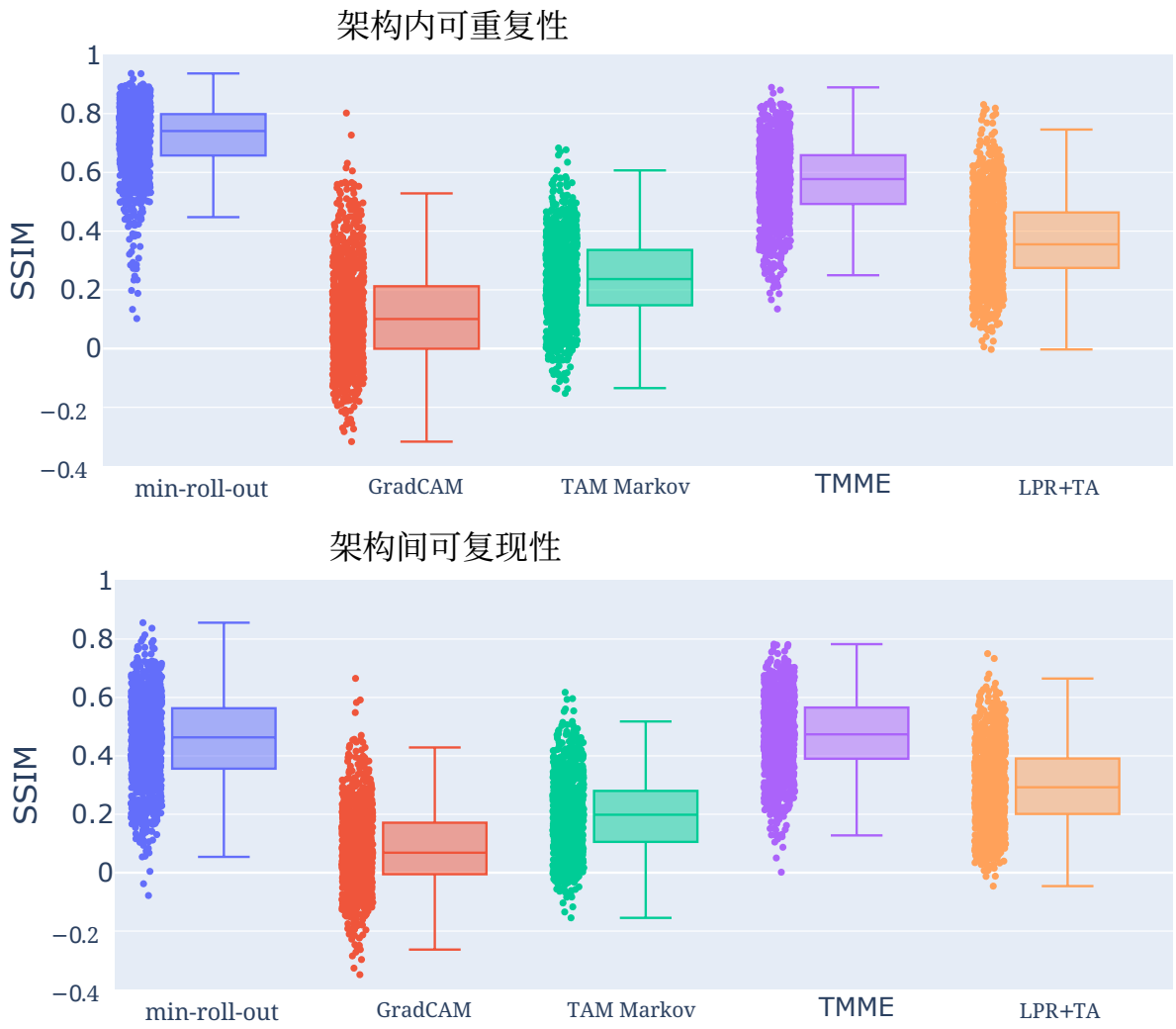


图 S. 6：架构内可重复性和架构间可复现性。

架构内可重复性是同一 ViT（ViT_b_244）在不同训练会话中生成的两张显著性图的 SSIM 评分。Min-Roll-Out：0.71（95%置信区间：0.71，0.72）；GradCAM：0.116（95%置信区间：0.11，0.13）；TAM Markov：0.24（95%置信区间：0.23，0.25）；TMME：0.57（95%置信区间：0.56，0.58）；LPR+TA：0.37（95%置信区间：0.36，0.38）；架构间可复现性是两种不同架构生成的两张显著性图的 SSIM 评分：文中报告的 ViT（ViT_b_244）和较小的 ViT（ViT_t_244）。Min-roll-out：0.46（95%置信区间：0.45，0.47）；GradCAM：0.082（95%置信区间：0.074，0.091）；TAM Markov：0.20（95%置信区间：0.19，0.21）；TMME：0.47（95%置信区间：0.47，0.48）；LPR+TA：0.30-（95%置信区间：0.29，0.31）。在原型开发期间，我们还研究了其他性能较差的基于注意力的显著图方法：min-roll-out、

转移注意力图（TAM）马尔可夫，以及逐层相关性传播变换器归因（LRP+TA）。SSIM = 结构相似性指数度量。

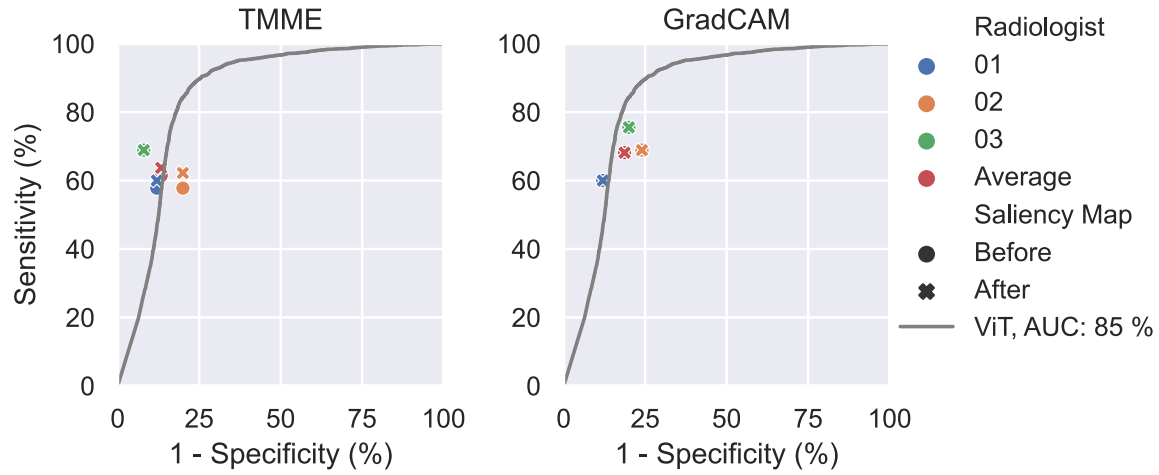


图 S. 7： 用户的受试者工作特征曲线（ROC）和模型曲线下面积

研究。 视觉Transformer模型（灰色）以及放射科医师在未使用（圆圈）和使用（叉号）额外显著性图进行气胸分类时的表现。平均读者表现以红色圆圈/叉号表示。AUC = ROC曲线下面积。

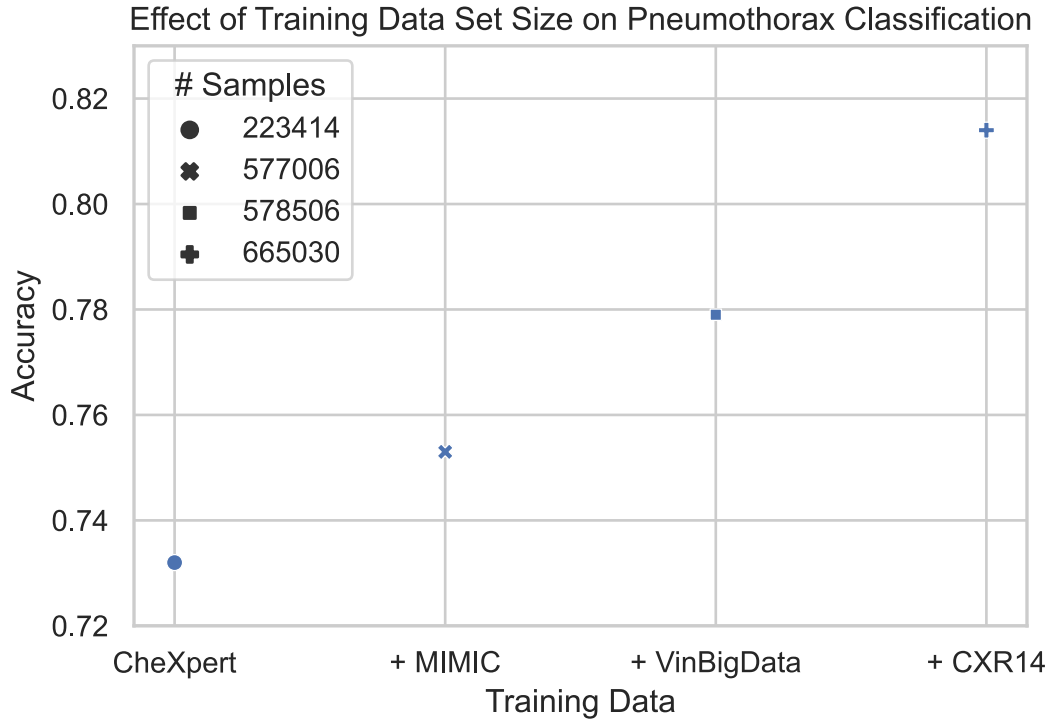


图 S. 8：增加训练数据对 SIIM ACR 测试数据集上气胸分类的影响。分类阈值根据验证数据上各自最佳的 F1 分数计算得出。所有网络均以相同的迭代次数进行训练。CXR14 = Chest X-Ray 14，SIIM-ACR = 医学影像信息学会-美国放射学会。