arXiv:2601.00355v1 [eess.IV] 1 Jan 2026

病灶关注对基于人工智能的黑色素瘤分类性能的影响

Tanay Donde 伊利诺伊大学厄巴纳-香槟分校计算机科学系
tanaydonde@gmail.com

摘要

黑色素瘤是皮肤癌中致死性最高的亚型，早期且准确地检测出这种疾病能够显著改善患者的预后 [1]。尽管机器学习模型，尤其是卷积神经网络（CNNs），在自动化黑色素瘤分类方面展现出巨大潜力，但由于对病灶区域的关注不一致，其诊断可靠性仍受到影响 [2]。本研究分析了病灶注意力与诊断性能之间的关系，涉及掩蔽图像、边界框检测和迁移学习。我们采用多种可解释性方法和敏感性分析方法，探究模型将注意力对齐至病灶区域的程度，以及这种对齐与精确率、召回率和 F1 分数之间的相关性。结果表明，更多关注病灶区域的模型取得了更好的诊断性能，这显示出可解释人工智能在医学诊断中的应用潜力。本研究为未来开发更准确、更可靠的黑色素瘤分类模型奠定了基础。

1 引言

皮肤癌是全球最常见的癌症，而黑色素瘤是其中致死性最高的一种。早期准确检测对于改善患者预后至关重要。若在早期确诊黑色素瘤，患者的五年生存率可超过 99%。然而，一旦癌症扩散至远处器官，这一比例就会降至约 32% [1]。

深度学习模型，尤其是卷积神经网络（CNNs），在自动化医学图像分析方面展现出巨大潜力 [2]。尽管前景广阔，许多模型仍缺乏可靠性，因为它们具有“黑箱”特性，无法清晰解释其预测结果 [3]。在黑色素瘤诊断中，可靠模型的一个关键因素是病灶注意力——即模型对皮肤镜图像中病灶区域的关注程度。

使用广泛应用的可解释性方法 Grad-CAM，我们观察到，基线卷积神经网络模型通常无法关注病灶区域，反而聚焦于无关的背景区域 [4]。图 1 展示了黑色素瘤分类过程中模型注意力与病灶未对齐的若干示例。

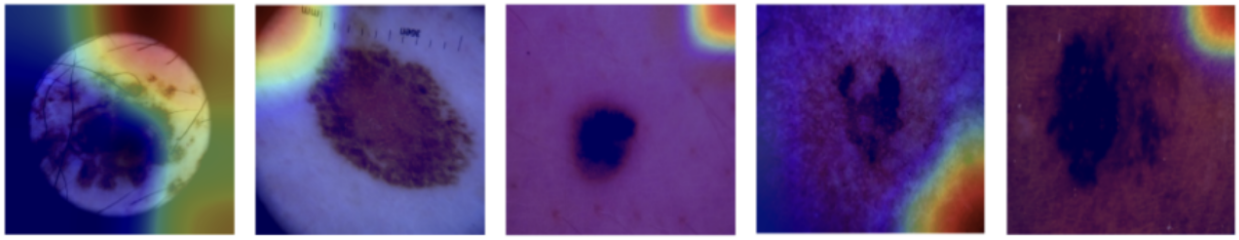


图 1：基线 InceptionV3 模型的 Grad-CAM 结果，展示了在黑色素瘤分类过程中，模型在不同示例中的注意力均未对齐，偏离了病灶区域。

此外，我们还观察到一些案例：模型虽然正确预测为黑色素瘤，但其注意力图却几乎完全未与病灶对齐。图 2 展示了一个正确分类的黑色素瘤案例，其中模型的注意力集中在远离感兴趣区域的位置。

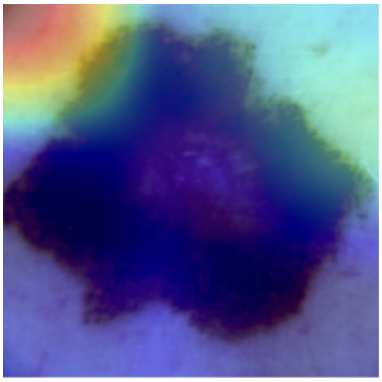


图 2：对一个预测正确的黑色素瘤病例进行 Grad-CAM 解释，显示模型的注意力未对齐，聚焦于远离病灶的区域。

这引出了我们的核心问题：模型对病灶区域的关注增加，是否对应着更好的诊断性能？直观来看，这似乎是一个显而易见的问题。提高模型对病灶区域的关注，理应自然地提升诊断准确率。然而，较大的模型往往会基于与任务无关的特征或上下文线索做出决策，使这种关系变得不那么直接。例如，在黑色素瘤分类中，模型可能会学习将肤色等背景特征与特定类别联系起来，而不是关注实际病灶。

回答这一问题对于提升人工智能驱动的黑色素瘤分类模型的准确率和透明度都至关重要。如果病灶关注的增加确实与更好的性能相关，那么这可能为设计兼具准确性和可靠性的模型提供一条路径，使其能够应用于医疗环境。通过持续关注病灶区域，这些模型可以提高可解释性，缓解人们对人工智能在医学影像中“黑箱”本质的担忧。此外，这些发现还可以指导可解释性驱动的训练技术的开发，使模型能够更有效地学习具有诊断相关性的特征。

为检验这一假设，我们设计了实验，以评估增强对病灶区域的关注是否能提升诊断性能。这些实验包括边界框检测、预训练，以及使用掩膜图像数据集进行迁移学习，在这些数据集中，病灶区域被单独提取出来。

除这些实验外，我们还采用了多种可解释性和敏感性分析方法，包括 Grad-CAM、Sobol' 和 RISE（随机输入采样解释），以定性评估模型对病灶区域的关注程度 [4, 5, 6]。这些方法共同构成了一个全面的评估框架，用于评估模型与病灶区域的对齐程度如何影响诊断准确率，并为开发更具可解释性且与诊断更相关的人工智能模型提供了见解。

2 相关工作

已有多项研究探索了人工智能模型在皮肤病变分析中的各种应用，包括病灶检测、定位和分类。尽管这些领域已取得显著进展，但人们仍未充分了解，提升病灶关注程度是否与诊断性能的改善直接相关。下文将考察相关工作，以说明本研究的背景，并展示其对人工智能驱动的皮肤病变分析所作的独特贡献。

Adegun 和 Viriri 全面综述了最先进的技术，讨论了皮肤病变分析中的预处理、分割和分类方法 [7]。虽然他们的综述将分割技术纳入病灶分析流程，但并未通过实验检验分割如何影响诊断准确率，也未比较基于分割的方法与不采用分割的方法。相比之下，我们的研究直接探讨了这种关系，揭示了病灶分割在黑色素瘤分类中的作用，并提供了新的见解。

将病灶定位整合到黑色素瘤检测中的尝试展现出了一定潜力。Taghizadeh 和 Mohammadi 提出了一种两步流程，使用 YOLOv3 进行病灶检测，并使用 SegNet 进行分割，取得了较高的定位准确率 [8]。虽然他们的研究证明了病灶定位的可行性，但其重点仅限于分割任务，并未评估病灶关注对黑色素瘤分类诊断准确率的影响。

Getamesay Haile Dagnaw、Meryam El Mouhtadi 和 Musa Mustapha 最近的研究探讨了视觉 Transformer（ViTs）与卷积神经网络（CNNs）在皮肤癌分类中的整合，并采用可解释性方法来解读模型决策 [9]。他们的研究展示了可解释人工智能在识别临床诊断相关特征方面的作用，进一步凸显了透明度对于人工智能驱动的医疗解决方案的重要性。虽然这一方法与我们对模型可解释性的关注一致，但我们的研究有所不同，重点在于病灶注意力与诊断性能之间的关系。此外，我们还研究了掩膜图像数据集、边界框检测和迁移学习等具体技术，以增强病灶关注，并对模型注意力对齐进行了更细致的分析。这些差异凸显了我们研究的独特贡献：探讨人工智能驱动的黑色素瘤分类中一个独特且鲜少研究的方面。

尽管取得了这些进展，但关于提高病灶关注是否与皮肤镜成像诊断准确率的提升相关，现有证据仍然有限。为弥补这一研究空白，我们采用多种可解释性方法评估模型注意力，以探究两者之间的关系。本研究通过深入理解模型推理，为推动人工智能驱动的黑色素瘤分类更加可解释、可靠这一更广泛的努力作出了贡献。

3 方法

为探究对病灶区域的关注度提高是否与诊断准确率提升相关，我们开展了一系列实验。这些实验探索了多种增强模型对病灶关注度的技术，例如基于分割的方法和目标检测任务，最终旨在理解病灶关注与模型表现之间的关系。实验的具体细节将在以下章节中介绍。所有实验均测试了多种超参数配置，并根据F1分数选出表现最佳的模型。Grad-CAM、Sobol'和RISE仅用于将各项实验的最终模型与基线模型进行比较，以定性评估病灶关注。所有实验均在Kaggle平台上使用Keras和TensorFlow实现，并在条件允许时利用GPU加速 [10, 11]。在预处理过程中，所有输入图像均被归一化至范围 $[-1, 1]$，以提高数值稳定性和训练效率。

3.1 数据集

本研究在所有实验中使用了两个公开数据集，即 ISIC-2019 和 HAM10000 [12, 13, 14]。

ISIC-2019 数据集是一个汇总数据集，包含来自多个子数据集的 25,331 张皮肤镜图像，其中包括 HAM10000。对于基线模型，这 25,331 张图像被划分为 75% 的训练集（18,999 张图像）和 25% 的验证集（ 6,332 张图像）。验证图像采用随机方式选取。为解决黑色素瘤（ MEL ）与非黑色素瘤（非 MEL）类别之间的不平衡问题，所有实验均对训练集采用过采样，以确保两个类别具有相同的样本数量。

HAM10000 数据集是 ISIC-2019 的组成部分，包含 10,015 张皮肤镜图像及其诊断标签。原始数据集不包含分割掩码，但一个由 HAM10000 衍生的 Kaggle 数据集提供了这些掩码 [15]。这些掩码最初由全卷积网络（FCN）生成，随后由皮肤科医生使用 FIJI 的自由手绘选择工具进行审核和修正。本研究使用这些经过整理的掩码开展基于分割的实验，并生成边界框。

HAM10000 和 ISIC-2019 均包含多个皮肤病灶类别，如表 1 [16] 所示。

类别名称	HAM10000 图像数量	ISIC-2019 图像数量
AKIEC	327	867
BCC	514	3,323
BKL	1,099	2,624
DF	115	239
MEL	1,113	4,522
NV	6,705	12,875
VASC	142	628
SCC	–	253
总计	**10,015**	**25,331**

表1：HAM10000 和 ISIC-2019 数据集中皮肤病变图像的类别分布，展示了每种病变类型的可用图像数量。

HAM10000 包含七个类别：日光性角化病和上皮内癌（AKIEC）、基底细胞癌（BCC）、良性角化样病变（BKL）、皮肤纤维瘤（DF）、黑色素瘤（MEL）、黑色素细胞痣（NV）以及血管性病变（VASC）。ISIC-2019 在此基础上增加了一个类别，即鳞状细胞癌（SCC）。图 3 [17]展示了 ISIC-2019 数据集的一个片段。

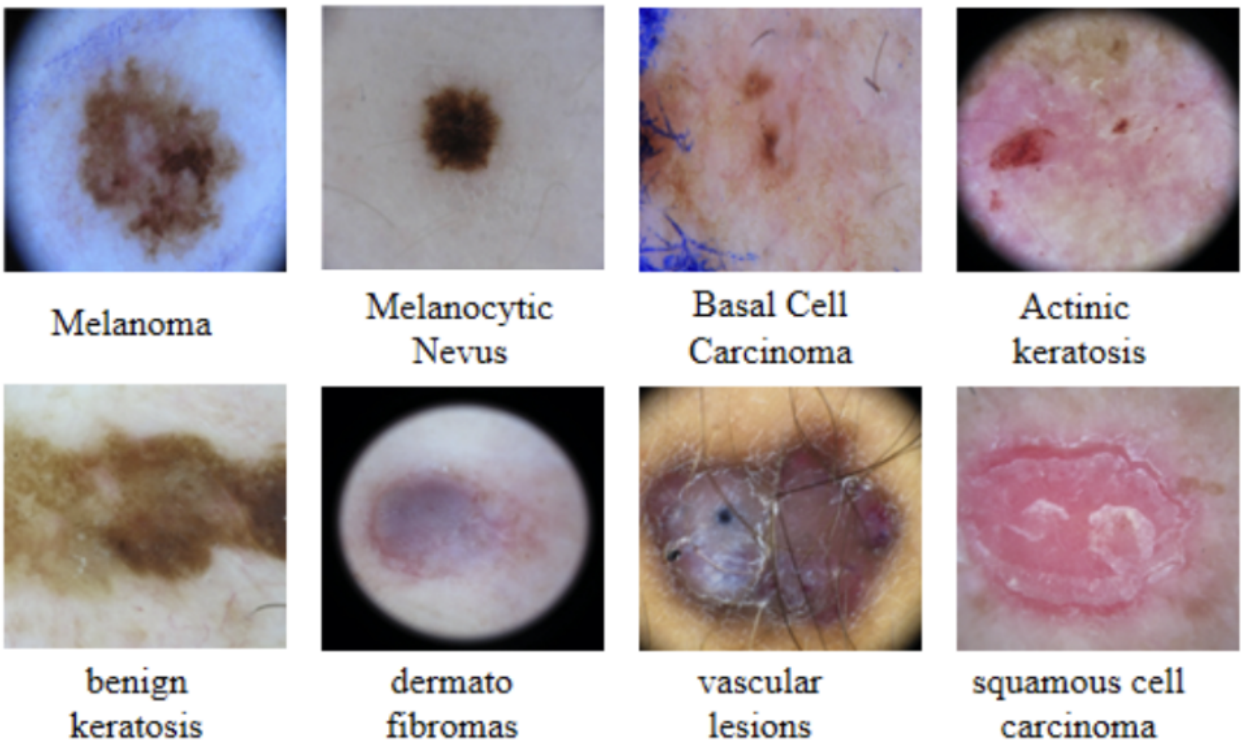


图 3：ISIC-2019 数据集中各类别的代表性图像。

本研究关注二分类，因此将这些类别归为黑色素瘤和非黑色素瘤两类。

3.2 评估指标

我们结合定量和定性指标对模型性能进行了评估：

- **精确率：**真阳性预测占所有阳性预测的比例，用于衡量模型避免假阳性的能力。
- **召回率：**真阳性预测占所有实际阳性病例的比例，反映模型对阳性样本的敏感度。
- **F1分数：** 精确率和召回率的调和平均数。由于类别不平衡，该指标用于在每次实验中选出表现最佳的模型。
- **可解释性方法：** 采用 Grad-CAM、Sobol' 和 RISE 定性评估模型注意力与病灶区域之间的对齐情况。这些方法仅应用于每项实验中表现最佳的模型，并与基线进行比较。

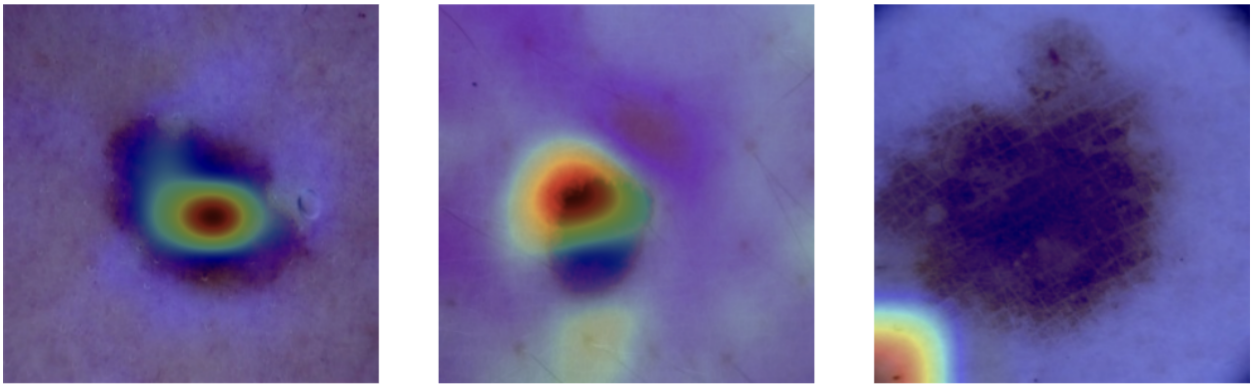


图 4：展示三种注意力对齐类别的视觉示例：完全对齐（左）、部分对齐（中）和未对齐（右）。

使用 Xplique 库生成热力图，该库为每种方法提供了实现 [18]:

- xplique.attributions.GradCAM 用于 Grad-CAM
- xplique.attributions.SobolAttributionMethod 用于 Sobol'
- xplique.attributions.Rise 用于 RISE

通过使用不透明度为 $\alpha = 0.5$ 的 jet 色彩映射，将生成的热力图叠加到输入图像上进行可视化，确保热力图和底层病灶均清晰可见。

使用可解释性热力图评估病灶关注可能涉及含糊不清的解读。“位于病灶上”这一概念本质上并不精确，因为注意力可能以不同程度与病灶重叠，或突出显示无关区域。由于数据集限制和预处理差异，统一的注意力与掩码定量重叠指标并不适用于所有模型。因此，我们采用受控的定性分析，并为视觉评估定义了清晰且一致的类别（见图 4）：

- **完全对齐：** 热力图主要集中在病灶上，其他区域没有受到显著关注。
- **部分对齐：** 热力图与病灶重叠，但也明显超出病灶边界，或突出显示了无关区域。
- **未对齐：** 热力图与病灶只有极少重叠，或完全没有重叠，注意力集中在病灶区域之外。

为确保定性评估的一致性，所有视觉评估均由同一人依据这些预定义类别完成。

3.3 基线模型

为选择合适的基线架构，我们评估了三种广泛使用的模型：ResNet50 [19],BEiT v2 [20], 和 InceptionV3 [21]。ResNet50 是一种卷积神经网络，利用残差连接实现对非常深的模型进行有效训练。BEiT v2 是一种专为大规模图像建模设计的基于 Transformer 的架构，但计算开销较大。InceptionV3 是一种卷积神经网络，通过并行卷积路径高效提取多尺度特征。本研究采用了 tensorflow.keras.applications [11] 提供的 InceptionV3 和 ResNet50 实现。BEiT v2 模型则来自 keras cv attention models 库 [22]。

每个模型都在 ISIC-2019 数据集上经过多轮训练和评估，用于区分黑色素瘤与非黑色素瘤的二分类。ResNet50 的最佳 F1分数为 0.661，BEiT v2 和 InceptionV3 的最佳 F1分数分别为 0.716 和 0.718。尽管 BEiT v2 的表现与 InceptionV3 相当，但其训练时间明显更长，因此不太适用于本研究开展的大规模实验。因此，考虑到性能出色且训练效率较高，研究选择 InceptionV3 作为基线架构。

基线 InceptionV3 模型采用预训练的 ImageNet 权重进行初始化，并使用 Adamax 优化器进行微调 [23]。训练过程中，研究采用了指数学习率衰减计划，并使用基于验证集的学习率回调机制动态调整学习率。

该基线模型作为对照，用于后续涉及不同数据集配置和训练策略的实验。

3.4 YOLO 模型

在本实验中，我们探究了在分类前加入目标检测，能否通过使模型在预测前定位病灶来提升诊断性能。为此，我们采用了 YOLO（You Only Look Once）框架，这是一种实时目标检测架构，旨在单次前向传播中联合执行目标定位和分类 [24]。

YOLO 模型在 HAM10000 数据集上进行了训练和评估，用于病灶检测和分类。该数据集被划分为 80% 的训练集和 20% 的验证集。训练前生成边界框，以调整数据集使其适用于目标检测。首先，对每张病灶图像进行填充，以保持一致的空间尺寸。接着，利用 HAM10000 数据集提供的相应分割掩码，确定紧密包围病灶的边界框。最后，将该边界框应用于填充后的图像，以生成检测目标。图 5 展示了这一边界框生成过程。

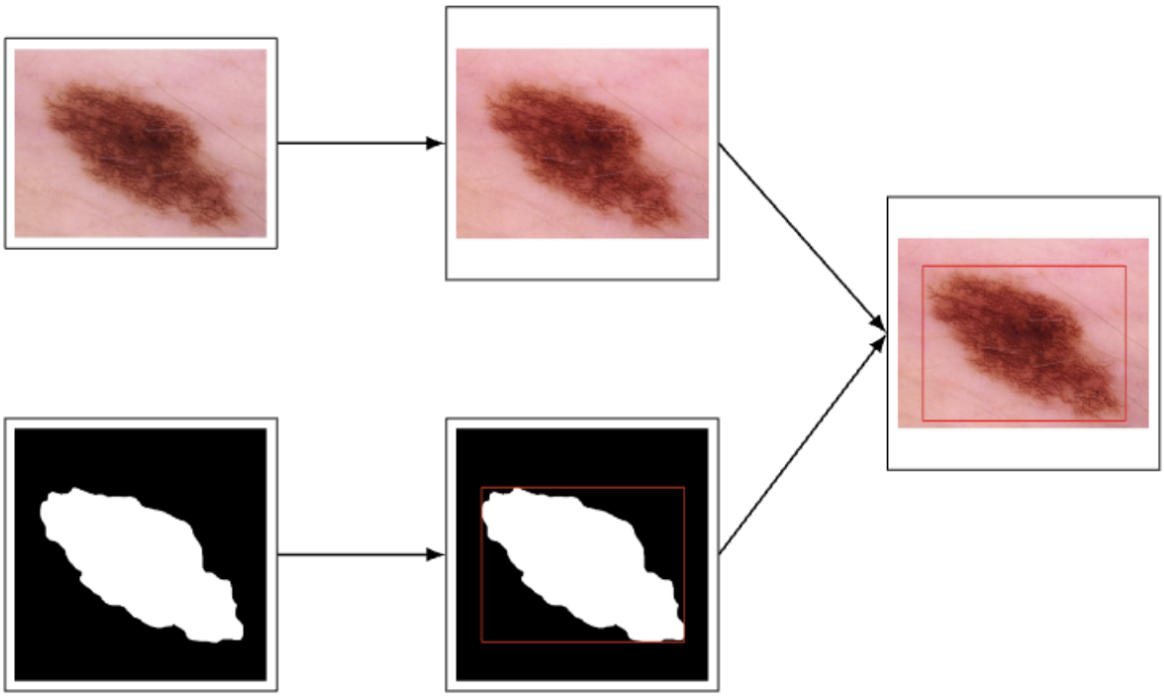


图 5：展示如何从 HAM10000 数据集的病灶分割掩膜生成边界框的流程图。

在实现过程中，我们采用了 Ultralytics YOLOv8 框架，该框架因其在目标检测任务中的稳健性和高效性而被选用 [25]。

3.5 MaskedImage实验

在掩膜图像实验中，我们研究了明确增强模型对病灶区域的关注是否会带来诊断性能的提升。为此，我们生成了掩蔽图像，其中病灶被单独分离，并移除了背景信息。

该流程从原始病灶图像开始，先进行填充以确保空间尺寸一致。接着，应用对应的分割掩码来分离病灶区域。病灶以外的所有像素均替换为统一的白色背景，从而得到最终的掩蔽图像。图6展示了这一掩膜处理流程的各个步骤。

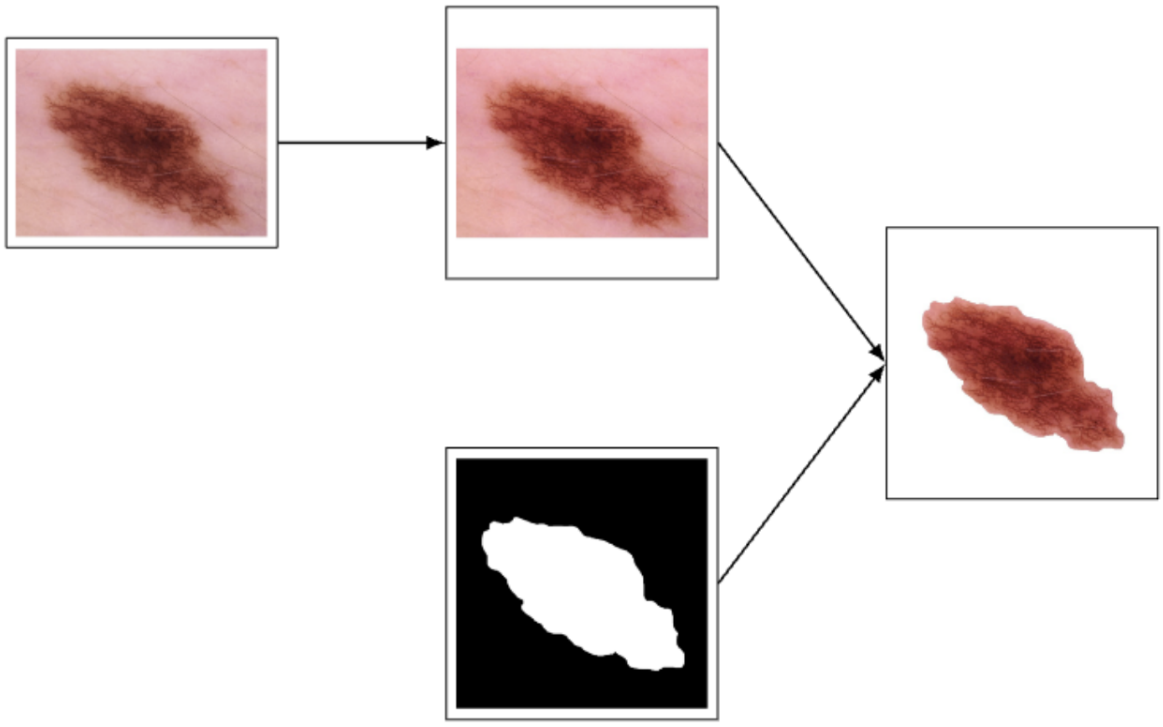


图6：展示病灶分割实验中生成掩蔽图像流程的流程图。

3.5.1 从掩膜图像到常规图像的迁移

为研究病灶分割的影响，我们使用预先训练的 ImageNet 权重初始化了一个新的 InceptionV3 模型，并仅使用来自 HAM10000 的掩蔽图像对其进行训练；这些图像中的病灶已被分离，背景则替换为白色。数据集按 80% 训练集和 20% 验证集进行划分。在此过程中，我们测试了多种学习率和超参数配置，以找出表现最佳的模型。确定最佳配置后，我们将该模型的权重迁移至常规 ISIC-2019 数据集，进行进一步训练。在第二阶段，我们再次尝试了多种学习率和超参数配置，并选取表现最佳的方案作为最终模型。迁移过程中冻结了不同的层，以保留从掩蔽图像中学到的知识。我们确保 ISIC-2019 验证集中不包含任何属于 HAM10000 掩蔽训练集的图像。此步骤对于确保评估公平性至关重要，因为在训练中用过的图像上进行测试，无法准确衡量模型的性能。这些模型使用 F1分数进行评估。

模型训练或超参数调优过程中未使用可解释性方法。相反，在确定表现最佳的最终模型后，才应用这些方法，并定性比较该模型与原始基线模型对病灶区域的关注情况。

3.5.2 组合掩蔽与常规数据集

为评估掩蔽图像与常规图像的综合影响，我们构建了一个包含 15,316 张常规图像和 10,015 张掩蔽图像的数据集。模型采用不同比例的常规图像和掩蔽图像进行训练，具体而言，常规图像占 50%、75% 和 90%，其余部分为掩蔽图像。我们评估了两种训练策略：

- **方法1：** 模型采用基于 ImageNet 预训练的 InceptionV3 权重进行初始化，并在组合数据集上训练了 30 个周期。测试了多种超参数配置，包括不同的学习率，并根据 F1分数 选择表现最佳的模型。

- **方法2：** 以在常规 ISIC-2019 数据集上训练的 InceptionV3 基线模型作为起点。随后，该模型在组合数据集上进行了 15 个周期的微调。与方法1相同，研究探索了多种超参数配置，并依据性能指标选出最佳配置。

对于方法2，研究特别注意确保验证集中不包含任何曾纳入 InceptionV3 基线模型训练集的图像。

训练结束后，根据 F1分数，从两种方法中选出表现最佳的模型进行进一步分析。仅对该模型应用 Grad-CAM、Sobol’ 和 RISE 来评估注意力焦点，并将结果与原始基线模型所得结果进行比较。

4 结果与讨论

本节总结了研究期间训练和评估的各种模型的性能。

4.1 基线模型性能

基线 InceptionV3 模型在 ISIC-2019 数据集上训练了 15 个周期，准确率达到 90.22%，F1分数为 0.743 （精确率：0.8096，召回率：0.6868 ）。我们采用 Grad-CAM 、 Sobol’ 和 RISE，分析模型对 40 张图像中病灶区域的注意力；这些图像包括 20 例真实黑色素瘤和 20 例真实非黑色素瘤。通过这种分层抽样，我们得以针对不同真实类别标签，对注意力对齐情况进行平衡分析。结果汇总于表 2。

可解释性方法	对齐	真实黑色素瘤	真实非黑色素瘤	总计
Grad-CAM	完全对齐	4	0	4
	部分对齐	7	12	19
	未对齐	9	8	17
Sobol'	完全对齐	10	3	13
	部分对齐	10	13	23
	未对齐	0	4	4
RISE	完全对齐	13	2	15
	部分对齐	6	8	14
	未对齐	1	10	11

表 2：基线 InceptionV3 模型针对 40 张平衡样本图像（20 例真实黑色素瘤和 20 例真实非黑色素瘤）生成的 Grad-CAM、Sobol' 和 RISE 注意力图的对齐情况。对齐类别包括完全对齐、部分对齐和未对齐。各类别中的数量表示相应可解释性方法和真实类别标签下的图像数量。

Grad-CAM、Sobol' 和 RISE 的结果存在显著差异，这可归因于各可解释性方法背后的不同机制。Grad-CAM 是一种基于梯度的方法，主要关注影响模型输出的最显著区域。相比之下，Sobol' 采用敏感性分析来评估不同输入区域的贡献，这或许能提供更广泛的视角，但缺乏细粒度定位。另一方面，RISE 利用随机扰动生成热力图，能够捕捉全局模式，但可能会分散对特定病灶区域的关注。这些方法上的差异解释了三种技术在对齐程度上的不同表现。

4.2 YOLO 模型性能

YOLO 模型在 HAM10000 数据集上进行训练，以检测病灶位置并将其分类为黑色素瘤或非黑色素瘤。该模型的准确率达到 85.13%，但 F1 分数明显偏低，仅为 0.348。尽管较高的准确率乍看之下表明模型表现良好，但与 F1 分数之间的差距说明，模型的预测严重偏向多数类（非黑色素瘤），实际上未能识别出黑色素瘤阳性病例。由于 InceptionV3 基线模型在完全相同的数据集上取得了高得多的 F1 分数（0.743），因此这一失效不能仅归因于类别不平衡。相反，这些结果表明，YOLO 的架构难以捕捉黑色素瘤诊断所需的细粒度纹理差异，而 InceptionV3 等标准卷积神经网络则保留了这些差异。鉴于该模型未能学到稳健的判别信号，因此未对 YOLO 架构开展进一步的可解释性分析（Grad-CAM、Sobol' 和 RISE）。

4.3 掩膜图像实验结果

4.3.1 从掩膜图像到常规图像的迁移结果

表现最佳的模型最初仅使用掩蔽图像进行训练，随后在常规 ISIC-2019 数据集上进行微调，准确率达到 90.19%，F1分数为 0.734（精确率：0.831，召回率：0.657）。在迁移学习过程中，仅解冻网络的前 200 层和最后几层时，模型取得了最优表现。

与基线模型相比，该方法的 F1分数和召回率略有下降，但精确率小幅提升。

为比较病灶注意力与基线模型，我们选取了40张图像进行分层抽样，其中包括20例真实黑色素瘤病例和20例真实非黑色素瘤病例。我们采用Grad-CAM、Sobol'和RISE，定性评估模型注意力与病灶区域的对齐情况。通过这项分析，我们得以评估微调相较于基线模型对病灶关注的影响。结果汇总于表3。

可解释性方法	对齐	真实黑色素瘤	真实非黑色素瘤	总计
Grad-CAM	完全对齐	3	6	9
	部分对齐	5	11	16
	未对齐	12	3	15
Sobol'	完全对齐	11	4	15
	部分对齐	8	13	21
	未对齐	1	3	4
RISE	完全对齐	5	4	9
	部分对齐	13	10	23
	未对齐	2	6	8

表3：对于由掩蔽到常规模型生成的 40 幅图像（20 幅真实黑色素瘤图像和 20 幅真实非黑色素瘤图像）的平衡样本，Grad-CAM、Sobol' 和 RISE 注意力图的对齐情况。数值的解读方式与表 2 相同。

对于真实黑色素瘤病例，结果显示，特征关注减弱与诊断敏感性降低之间存在明显相关性。掩膜到常规模型在所有可解释性方法中都表现出更严重的注意力未对齐（其中 Grad-CAM 最为明显，未对齐实例从 9 个增加到 12 个）。这种关注度的下降与召回率的降低直接对应（从 0.687 降至 0.657），证实当模型无法关注诊断性病灶特征时，就无法正确识别病理，进而造成漏诊（假阴性）。

相反，真实非黑色素瘤病例显示，关注更加集中与预测精确率提高相关。模型对这些图像的对齐情况显著改善，Grad-CAM、Sobol' 和 RISE 中的未对齐实例分别从 8 个降至 3 个、从 4 个降至 3 个，以及从 10 个降至 6 个。对实际病灶更集中的关注，使模型能够正确排除良性病例，而不被背景噪声干扰，从而减少假阳性。因此，错误减少推动精确率从 0.810 提高到 0.831。

4.3.2 掩膜与常规数据集组合结果

使用 ImageNet预训练权重进行训练集

在此方法中，我们使用 ImageNet 预训练权重初始化了一个新的 InceptionV3 模型，并在由掩蔽图像和常规图像组成的组合数据集上训练了 30 个周期。我们测试了三种数据集组成，常规图像分别占 50%、75% 和 90%，其余为掩蔽图像。主要结果，包括准确率、精确率、召回率和 F1分数，汇总于表4。

常规图像占比	准确率（%）	精确率（%）	召回率 (%)	F1分数
50	87.12	83.80	45.32	0.588
75	90.19	87.97	59.67	0.711
90	90.31	84.48	64.96	0.734

表4：方法1的性能指标，展示了组合数据集中常规图像占不同比例时的准确率、精确率、召回率和F1分数。

使用皮肤癌专用预训练权重进行训练

在此方法中，我们以在 ISIC-2019 常规数据集上预训练的基线 InceptionV3 模型作为起点。该模型在由掩蔽图像和常规图像组成的组合数据集上进行了微调。测试的数据集组成包括常规图像占 50%、75% 和 90%，其余部分为掩蔽图像。关键结果，包括准确率、精确率、召回率和 F1分数，汇总于表5中。

常规图像占比	准确率（%）	精确率（%）	召回率 (%)	F1分数
50	90.78	86.84	54.98	0.673
75	**91.87**	**85.30**	**71.90**	**0.780**
90	91.09	94.03	51.22	0.663

表5：方法1的性能指标，展示了组合数据集中不同常规图像占比下的准确率、精确率、召回率和F1分数。

组合数据集最佳性能模型分析

通过比较方法1和方法2中表现最佳模型的F1分数，选出了最佳模型。方法2中常规图像占比为75%的模型（表5中以粗体标出）被选为表现最佳的模型。该模型的准确率达到91.87%，F1分数为0.780（精确率为0.853，召回率为0.719），在所有定量指标（精确率、召回率和F1分数）上均优于基线模型。

为评估表现最佳模型的注意力对齐情况，使用 Grad-CAM、Sobol’ 和 RISE 可解释性方法分析了40张图像（20张黑色素瘤图像和20张非黑色素瘤图像）。这项定性评估的结果汇总于表6。

可解释性方法	对齐	真实黑色素瘤	真实非黑色素瘤	总计
Grad-CAM	完全对齐	7	1	8
	部分对齐	3	12	15
	未对齐	10	7	17
Sobol'	完全对齐	15	7	22
	部分对齐	5	12	17
	未对齐	0	1	1
RISE	完全对齐	12	7	19
	部分对齐	5	8	13
	未对齐	3	5	8

表6：组合数据集模型针对 40 张图像（20 张真实黑色素瘤图像和 20 张真实非黑色素瘤图像）的平衡样本生成的 Grad-CAM、Sobol' 和 RISE 注意力图的对齐情况。数值的解释方式与表 2 中所述相同。

对于真实黑色素瘤病例，组合数据集模型显示，注意力对齐的改善与诊断敏感性呈正相关。该模型在三种可解释性方法中的两种方法上对齐程度有所提高，其中 Grad-CAM 的完全对齐实例从4个增至7个，Sobol' 的完全对齐实例从10个增至15个。尽管 Grad-CAM 和 RISE 中未对齐实例略有增加，但完全对齐的注意力图显著增多，表明模型定位病灶特征的能力总体有所提升。注意力聚焦的整体改善与召回率的提升（从0.6868增至0.719）直接对应，证实了模型对相关病理特征关注度的提高有助于其正确分类真实黑色素瘤病例。

此外，真实非黑色素瘤病例表现出更清晰的关注焦点与更高的预测精确率之间存在全面相关性。该模型在所有可解释性方法上的对齐情况均显著改善，具体表现为 Grad-CAM（从 0 增至 1）、Sobol'（从 3 增至 7）和 RISE（从 2 增至 7）的完全对齐实例增加。此外，未对齐实例全面减少，其中 RISE（从 10 降至 5）和 Sobol'（从 4 降至 1）的降幅最为显著。关注焦点更加集中，表明模型更有效地忽略背景噪声，从而正确识别良性病例并减少假阳性。这一机制上的改进与精确率从 0.8096 提升至 0.853 直接对应。

5 局限性

尽管本研究提供了大量证据，表明病灶注意力与诊断性能相关，但仍有若干局限性值得考虑。

首先，对注意力对齐的定性评估依赖目视检查和预先设定的类别（完全对齐、部分对齐、未对齐）。尽管采用了统一的标准，与交并比（IoU）或像素级敏感度评分等定量指标相比，这一过程本质上仍具有主观性。未来的研究可使用更精细的自动化对齐度量来验证这些发现。

其次，可解释性分析仅针对由 40 张图像（20 张黑色素瘤图像和 20 张非黑色素瘤图像）组成的分层抽样进行。样本量受到定性评估需手动进行这一因素的限制，因为该评估需要经过详细的人工审查，以对注意力图的对齐情况进行分类

并比较多种可解释性方法。尽管该样本经过平衡，以确保具有公平的代表性，但可能无法充分反映整个验证数据集中的模型行为方差。

第三，YOLO 架构未能取得有竞争力的 F1 分数，这凸显了在该领域应用纯目标检测框架的局限性。正如结果所示，YOLO 针对空间定位进行优化，可能削弱了其学习黑色素瘤鉴别所需的细粒度纹理特征的能力。这表明，尽管定位很重要，但它无法取代标准卷积神经网络分类器的纹理特征提取能力。

最后，尽管采用过采样来缓解类别不平衡问题，但底层数据集（HAM10000 和 ISIC-2019）仍严重偏向非黑色素瘤病例。这种固有的不平衡持续给稳健诊断模型的训练带来挑战，也可能影响临床部署场景中假阴性的基线概率。

6 结论

本研究探讨了模型对病灶区域的注意力与黑色素瘤分类模型诊断性能之间的关系。通过开展涉及边界框检测、掩膜图像数据集和迁移学习的实验，我们观察到，对病灶区域关注度更高的模型与诊断准确率提升之间存在明显关联。尤其是，组合数据集模型的可解释性图与病灶区域的对齐程度更高，同时在所有关键指标上均优于基线模型。

这些结果表明，迫使模型聚焦于临床相关特征，是提升诊断可靠性的可行途径。除了单纯关注性能指标之外，我们的研究结果还凸显了可解释性方法在指导模型开发方面发挥的关键作用。随着人工智能日益融入临床工作流程，确保模型依赖可解释特征（例如病灶本身，而非背景噪声）对于建立信任至关重要。归根结底，本研究强调了注意力聚焦训练方法的必要性，并展示了可解释性工具如何不仅作为审查工具，还能作为设计可信医疗人工智能的积极组成部分。

致谢

我谨向Ivan Felipe Rodriguez博士致谢，感谢他给予的宝贵指导、对实验设计的指导，以及在准备本文期间提供的反馈。

参考文献

[1] Skin Cancer Foundation。《皮肤癌事实与统计数据》，2024年。访问日期：2024年12月28日。

[2] J. Ker、L. Wang、J. Rao 和 T. Lim。《深度学习在医学图像分析中的应用》。*IEEE Access*，6:9375–9389，2018年。[3] Cynthia Rudin。停止为高风险决策解释黑箱机器学习模型，改用可解释模型。*Nature Machine Intelligence*，1:206–215，2019年。

[4] Ramprasaath R. Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh 和 Dhruv Batra。Grad-CAM：通过基于梯度的定位，为深度网络提供可视化解释。*arXiv* 预印本 *arXiv:1610.02391*，2019年。[5] Thomas Fel、R´emi Cadene、Mathieu Chalvidal、Matthieu Cord、David Vigouroux 和 Thomas Serre。看看方差！使用基于 Sobol 的敏感性分析，提供高效的黑盒解释。*arXiv* 预印本 *arXiv:2111.04138*，2021年。[6] Vitali Petsiuk、Abir Das 和 Kate Saenko。RISE：用于解释黑盒模型的随机输入采样。*arXiv* 预印本 *arXiv:1806.07421v3*，2018年。[7] Adewole Adegun 和 Serestina Viriri。用于皮肤病变分析和黑色素瘤癌症检测的深度学习技术：前沿技术综述。《人工智能评论》，54:811–841，2021年。[8] Soroush Taghizadeh 和 Mohammad Mohammadi。基于深度迁移学习，使用微调后的 YOLOv3 和 SegNet 快速准确地检测和分割黑色素瘤皮肤癌。*arXiv* 预印本 *arXiv:2210.05167v2*，2023年。[9] Getamesay Haile Dagnaw、Meryam El Mouhtadi 和 Musa Mustapha。使用视觉变换器和可解释人工智能进行皮肤癌分类。《医学人工智能杂志》，7:14，2024年。[10] Fran¸cois Chollet 等。Keras，2015年。[11] Mart´ın Abadi 等。TensorFlow：面向异构系统的大规模机器学习。*arXiv* 预印本 *arXiv:1603.04467*，2016年。[12] Noel C. F. Codella、David Gutman、M. Emre Celebi、Brian Helba、Michael A. Marchetti、Stephen W. Dusza、Aadi Kalloo、Konstantinos Liopyris、Nabin Mishra、Harald Kittler 和 Allan C. Halpern。面向黑色素瘤检测的皮肤病变分析：2017年国际生物医学成像研讨会（ISBI）的一项挑战，由国际皮肤成像协作组织（ISIC）主办。*arXiv* 预印本 *arXiv:1710.05006v3*，2018年。[13] Philipp Tschandl、Cliff Rosendahl 和 Harald Kittler。HAM10000 数据集：包含常见色素性皮肤病变的多源皮肤镜图像的大型集合。《科学数据》，5:180161，2018年。[14] Carlos Hern´andez-P´erez、Marc Combalia、S´ara Podlipnik、Noel C. Codella、Veronica Rotem-berg、Allan C. Halpern、Ofer Reiter、Cristina Carrera、Adriana Barreiro、Brian Helba、Susana Puig、Ver´onica Vilaplana 和 Josep Malvehy。BCN20000：真实环境中的皮肤镜病变。《科学数据》，11(1):641，2024年。[15]Philipp Tschandl。HAM10000 病灶分割。https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000，2020年。Kaggle 数据集，访问日期：2025年1月1日。[16] Long Hoang、Sang-Hyun Lee、Eun-Ji Lee 和 Kyung-Ryong Kwon。面向智慧医疗的新型轻量级深度学习框架：多类别皮肤病变分类。《应用科学》，12(5):2677，2022年。[17] S. Li、P. Kou、M. Ma、H. Yang、S. Huang 和 Z. Yang。半监督学习在图像分类中的应用：有标签与无标签数据融合研究。*IEEEAccess*，12:27331–27343，2024年。

[18] T. Fel、L. Hervier、D. Vigouroux、A. Poche、J. Plakoo、R. Cadene、M. Chalvidal、J. Colin、T. Boissin、L. Bethune、A. Picard、C. Nicodeme、L. Gardes、G. Flandin 和 T. Serre。Xplique：深度学习可解释性工具箱。*arXiv* 预印本 *arXiv:2206.04394*，2022年。[19] K. He、X. Zhang、S. Ren 和 J. Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，第 770–778 页，2016年。[20] Z. Peng、W. Huang、S. Guo、B. Du 和 Y. Wei。BEiT v2：利用向量量化视觉分词器进行掩码图像建模。2022年。[21] C. Szegedy、V. Vanhoucke、S. Ioffe、J. Shlens 和 Z. Wojna。重新思考用于计算机视觉的 Inception 架构。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，第 2818–2826 页，2016年。[22]Leon D. Garse。keras-cv-attention-models。https://github.com/leondgarse/keras_cv_attention_models，2024年。GitHub 仓库，访问日期：2025年1月1日。[23] Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2017年。[24] Joseph Redmon、Santosh Divvala、Ross Girshick 和 Ali Farhadi。You Only Look Once：统一的实时目标检测。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，第 779–788 页，2016年。[25]Glenn Jocher 等。Ultralytics。https://github.com/ultralytics/ultralytics，2023年。GitHub 仓库，访问日期：2025年1月1日。