基于多目标优化的医学影像可解释性增强研究*

李海芳[1,2,3]，唐　超[3]，岳　鑫[3]，张　强[1,2]
（1. 大连理工大学　计算机科学与技术学院，辽宁　大连　116024；
2. 大连理工大学　社会计算与认知智能教育部重点实验室，辽宁　大连　116024；
3. 新疆师范大学　计算机科学技术学院，新疆　乌鲁木齐　830054）

摘　要：针对医学影像场景的解释需求，提出一种基于多目标粒子群优化的解释增强方法，通过优化解释生成过程来提升解释质量与临床可读性。该方法在 LIME（局部与模型无关的解释）框架中引入多目标搜索机制，实现了高解释保真性与高区域稀疏性的自适应权衡，并获得了帕累托最优的解释结果。为验证方法有效性，以膝关节 X 光影像为实验对象，基于公开膝骨关节炎数据集在典型卷积神经网络上进行了实验评估。实验结果显示，保真性最高可提升 18%，稀疏性最大可降低 22%，展现出更高的聚焦性、稳定性，为基于人工智能的医疗影像可信诊断提供了可行技术路径。

关键词：膝骨关节炎；医学影像可解释性；多目标粒子群优化；LIME；可信赖医疗人工智能
中图分类号：TP309.7　　**文献标志码：**A　　**DOI：**10.19358/j.issn.2097-1788.2026.04.008

中文引用格式：李海芳，唐超，岳鑫，等. 基于多目标优化的医学影像可解释性增强研究[J]. 网络安全与数据治理，2026，45(4)：59-67.
英文引用格式：Li Haifang，Tang Chao，Yue Xin，et al. Multi-objective optimization for enhanced explainability in medical imaging models [J]. Cyber Security and Data Governance，2026，45(4)：59-67.

Multi-objective optimization for enhanced explainability in medical imaging models

Li Haifang[1,2,3]，Tang Chao[3]，Yue Xin[3]，Zhang Qiang[1,2]
（1. School of Computer Science and Technology，Dalian University of Technology，Dalian 116024，China；
2. Key Laboratory of Social Computing and Cognitive Intelligence（Dalian University of Technology），Ministry of Education，Dalian 116024，China；3. School of Computer Science and Technology，Xinjiang Normal University，Urumqi 830054，China）

Abstract：To address the need for reliable interpretability in medical imaging，this study proposes a multi-objective particle swarm optimization-enhanced explanation framework that improves explanation quality and clinical readability by optimizing the LIME（Local and Model-Agnostic Explanations）process. The proposed method incorporates a multi-objective search strategy into the LIME pipeline，enabling an adaptive trade-off between explanatory fidelity and regional sparsity，and producing Pareto-optimal explanation outcomes. Experiments conducted on knee X-ray images from a publicly available knee osteoarthritis dataset using representative convolutional neural networks demonstrate that the method increases fidelity by up to 18% and reduces sparsity by up to 22%，resulting in more focused and stable explanations. These results indicate that the proposed framework offers a feasible and effective pathway toward trustworthy AI-driven medical image interpretation.

Key words：knee osteoarthritis；medical image explainability；multi-objective particle swarm optimization；LIME；trustworthy medical artificial intelligence

0　引言

随着深度学习模型与计算资源的快速发展，数据驱动方法已在视觉与医学影像分析中取得显著进展。然而，由于深度神经网络在结构与训练过程中的高度复杂性，其决策机制往往呈现“黑箱”特征，使得模型预测结果难以被人类理解与复核。在医疗影像等高风险应用场景中，这种不可解释性会直接影响系统的

*基金项目：国家重点研发计划（2024YFA1012700）；教育部人文社科项目（25YJCZH119）

可信性、可审计性与临床可用性，并进一步牵涉到人机协作流程与责任归属等系统性问题[1-2]。相关研究指出，可信医疗 AI 的解释机制不应仅停留在单一算法层面的可视化呈现，而需与临床工作流、交互方式与问责框架形成闭环，以支撑真实场景下的信任建立与可控使用[3]。

在医学影像领域，膝骨关节炎（Knee Osteoarthritis，KOA）是较为常见的退行性关节疾病，其 X 光影像中的关节间隙变窄、软骨下骨硬化及骨赘形成等特征是临床分级和诊断的关键依据。近年来，KOA 的自动诊断与分级任务的处理广泛应用卷积神经网络，并在多个公开数据集上获得了较高的识别精度[4]。但多数方法仍以“预测输出”为主，缺乏对模型预测依据的清晰解释，临床医生难以评估模型是否关注了合理的病灶结构，进而限制了其在辅助诊断场景中的应用效果。

围绕可解释人工智能（Explainable Artificial Intelligence，XAI），现有研究主要可以分为事后解释与事前可解释建模两大类。其中，事后解释方法通过可视化或归因分析对黑盒模型的局部决策进行近似刻画，代表性方法包括基于梯度的 Grad-CAM 与 Grad-CAM++[5-6]、模型无关的 LIME [7]与 SHAP [8]，以及积分梯度方法[9]。因无需修改原有模型结构，此类方法在医学影像中应用较为广泛。然而，多项研究表明，显著图与归因结果对模型参数、输入扰动与解释超参数等高度敏感，可能产生不稳定甚至错误的解释；仅凭“看起来合理”的可视化难以保证解释的可信性[10-11]，因此需要综合系统化的量化评估与鲁棒性检验加以规范[12-13]。

在 KOA X 光影像场景中，上述挑战更为突出。一方面，关键特征通常呈现局部化、细粒度分布，传统 LIME 的超像素扰动与默认参数配置往往难以与医学结构精确对齐，容易导致解释区域过散、过大或覆盖无关特征；另一方面，LIME 以单一局部拟合优度作为优化目标时，可能出现解释区域“膨胀”，即通过激活大量非关键区域来提升拟合度，进而削弱解释的稀疏性与定位能力。这类不稳定现象已在多项医学影像解释可信性评估研究中被观测到[11]。

从方法论角度看，可信医学影像解释通常需同时满足两个核心属性：一是高保真性，即解释结果应忠实反映模型的局部决策行为；二是高稀疏性，即解释区域应聚焦于有限且具有临床意义的关键区域，以提升可读性与诊断效率[12-14]。二者之间存在内在张力：单纯追求拟合精度往往引入冗余高亮，而过度稀疏则可能损失关键区域信息。为此，本文引入多目标优化视角，将解释生成过程从“单点参数设定”调整为“可控权衡的参数搜索”。群智能优化方法因参数量少、全局搜索能力强，且对复杂目标具有良好适应性，已成为处理多目标权衡问题的经典方法[15]。

基于以上背景，本文以 KOA X 光影像解释为研究对象，提出一种基于多目标粒子群优化的事后可解释性增强方法（MOPSO-LIME）。该方法在不改变原有诊断模型结构的前提下，将解释过程形式化为“保真性-稀疏性”的双目标优化问题，通过自动搜索 LIME 参数空间，在两类目标之间实现可控平衡，进而获得更稳定、聚焦且具备临床可读性的解释结果。本文的主要贡献包括：

（1）提出一种面向 KOA 医学影像的多目标解释优化框架，在固定诊断模型条件下对 LIME 解释参数进行整体优化；

（2）构建以局部拟合优度与解释区域稀疏性为核心的可量化评价指标，为解释质量评估与优化提供统一标准；

（3）在公开 KOA 数据集及多种主流网络结构上验证方法有效性，实验结果表明该方法在不降低诊断性能的前提下提升了解释稳定性与临床可读性。

1 相关工作

事后可解释方法因其模型无关性与较低的应用门槛，在医学影像领域得到了广泛研究与应用。基于梯度的可视化方法（如 Grad-CAM 与 Grad-CAM++），通过分析特征图对预测结果的梯度响应，实现对模型关注区域的粗粒度定位[5-6]；模型无关方法如 LIME[7]与 SHAP [8]则通过局部扰动与替代模型近似刻画黑盒决策边界。尽管上述方法为医学影像分类与病灶定位提供了直观解释手段，但其解释可靠性问题亦逐渐成为研究焦点。

Adebayo 等通过“sanity check”实验平台说明多类显著图方法在参数扰动或随机初始化下仍可能输出看似合理的解释，表明视觉一致性并不能直接等价于解释可信性[10]。在医学影像场景中，Arun 等进一步从定位有效性、扰动敏感性与可重现性多维度评估显著图方法，发现其在异常区域定位任务中普遍存在稳定性不足的问题[11]。

针对上述挑战，研究逐渐从定性展示转向解释质量的系统化量化评估。Nauta 等对 XAI 领域的大量研究进行了系统综述，提出应从紧凑性、完整性、稳定性与定位能力等多维属性对解释方法进行评估，并强

调定量化与可重复评估在解释可信研究方面的重要性[12]。Hedström 等进一步提出 Quantus 工具箱，集成多类扰动与一致性指标，为不同解释方法的统一比较提供了可复现的评估框架[13]。在医学影像场景中，相关研究进一步指出，仅关注解释可视化效果难以全面反映模型行为，解释的稳定性、公平性与跨样本一致性等同样是解释可信性的重要组成部分。高宇系统分析了医学图像分类任务中解释方法在不同模型、不同样本分布下的稳定性与公平性问题，强调应从解释结果的可重复性与可验证性角度对现有方法进行综合评估[16]。更进一步，面向医疗实际落地的研究指出，解释机制需要与临床工作流和责任边界相匹配，才能在真实场景中形成可追溯、可问责的信任链条[3]。这些工作共同为医学影像解释结果的比较分析与可信性验证奠定了研究基础。

除显著图方法外，部分研究通过引入概念层或结构先验提升解释的语义一致性与临床可沟通性。TCAV 方法通过人类可理解概念对模型预测进行定量检验，实现从像素级归因向概念级解释的提升[17]。在医学影像场景中，Patrício 等提出基于概念的一致性解释框架用于皮肤病变分析[18]；Zhang 等通过联合视觉证据与诊断报告生成，实现模型判读路径的显式呈现[19]。此外，弱监督学习与多实例学习方法在降低标注依赖的同时，也为病灶证据定位提供了可行途径[20]。不过，上述方法多依赖额外标注信息、特定网络结构或任务定制模块，其适用前提与实现成本与模型无关的统一解释框架存在差异。

与上述研究侧重于语义对齐、证据定位或结构一致性的解释形式不同，本文关注于解释过程本身的优化建模。在保持原有黑盒诊断模型结构不变的前提下，本文将医学影像解释质量需求形式化为“保真性 – 稀疏性”的双目标优化问题，并通过多目标粒子群优化对 LIME 参数进行自动搜索与平衡。该方法不依赖额外标注或结构修改，能够在多种网络架构下生成更稳定、聚焦且具备临床可读性的解释结果，为 KOA 医学影像解释提供一种可量化、可复核的增强路径。

2　多目标优化解释模型分析思路

2.1　基于可解释性优化的整体框架

为实现深度诊断模型的可信解释与临床友好展示，本研究设计了一套端到端的可解释性增强流程，将深度特征学习与多目标解释优化统一纳入同一框架。整体流程从原始影像输入开始，依次完成模型训练、预测生成、解释构建、解释质量评价与优化，最终输出兼具保真性与稀疏性的可视化解释结果。该框架强调：

（1）模块化设计，支持不同骨关节炎诊断网络与解释方法替换；

（2）闭环式可解释性控制，解释结果与模型推理过程动态联动；

（3）定量 – 可视化双重输出，确保解释既科学可信又便于医生判读。

在此基础上，本文提出的 MOPSO-LIME 机制实现了对解释过程本身的自动调节与优化，为医学影像模型的透明化与临床落地提供了一种可行性参考。

图 1 展示了本研究的整体流程框架。模型采用端到端的深度学习架构，并在预测阶段引入多目标解释优化机制，形成从数据输入、模型推理到解释生成与可视化输出的完整闭环。数据集按训练/验证/测试划分（训练集约 80%），训练阶段分别对 VGG16、ResNet-50 与 MobileNetv2 三种网络结构进行参数学习，并通过验证集进行超参数调优与早停，最终保存性能最优的网络权重。

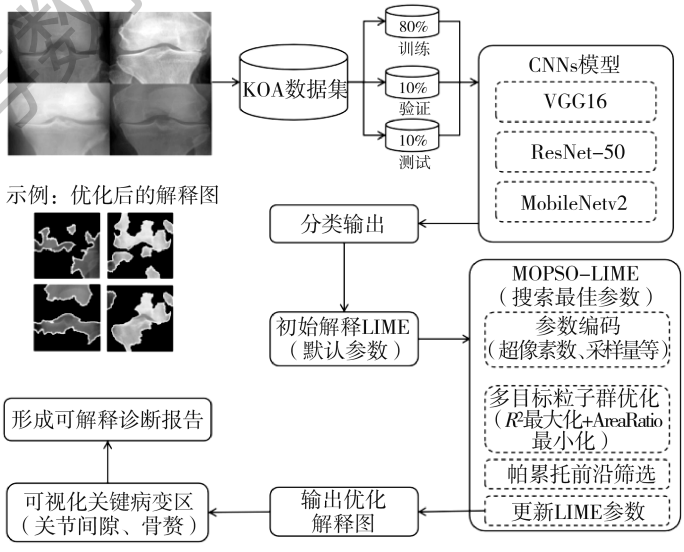

图 1　基于深度学习的 KOA 诊断与 MOPSO-LIME 解释流程

图 2 给出了 MOPSO-LIME 的核心优化流程。方法将 LIME 的关键解释超参数（如核宽度、采样量、特征选择数量与正则化强度）编码为可搜索变量，并以局部拟合优度 R^2 与解释区域稀疏性 AreaRatio 为双优化目标。模型优化过程中，二者存在天然权衡关系。过分追求高 R^2 会导致解释区域膨胀，产生冗余大面积高亮，使解释失去聚焦性；过度强调稀疏性则可能丢失关键判别特征，导致解释与模型真实推理过程偏离。

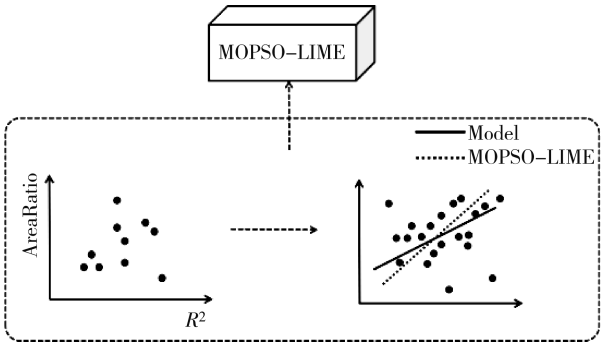

图 2　MOPSO-LIME 多目标解释优化机制示意图

为解决这一矛盾，本研究通过多目标粒子群优化自动搜索最佳参数组合，生成兼顾高保真与低冗余的解释热图。最终输出可视为解释质量的帕累托最优解，使解释更符合临床关注模式，帮助医生快速聚焦疑似病变区域，提高解释的可信度与诊断辅助价值。

2.2　保真性优化目标与过程

为了确保解释的高保真性，需要优化 LIME 模型中的关键参数，使得生成的解释能够充分拟合原始模型的局部决策。

（1）保真性优化目标

保真性优化的目标是最大化解释模型的 R^2 值。R^2 用来度量解释结果与原始模型决策的一致性，其值越高，解释结果越精确地反映了模型的实际决策过程。

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{1}$$

其中，y_i 为真实标签，$\hat{y}_i$ 为模型预测值，$\bar{y}$为真实标签的平均值，n 为样本数量。

（2）保真性优化过程

首先，通过 PSO 自动调整 LIME 的关键参数，如核宽度（kernel width）、样本数量（num samples）等。

其次，在局部扰动样本的基础上进行线性拟合，优化使得该线性模型的预测与原始复杂模型的决策行为高度一致。

最后，通过不断调整粒子群参数，使得模型的 R^2 值尽可能高，以此提高解释的精度。

（3）粒子群优化过程

对于粒子群优化，每个粒子代表一个参数组合。假设粒子群中的每个粒子位置为 $p = (p_1, p_2, \cdots, p_m)$，其中每个参数对应 LIME 的一个超参数。

粒子适应度函数如式（2）所示：

$$f_{\text{fitness}}(p) = R^2(\hat{y}(p), y) \tag{2}$$

其中，$\hat{y}(p)$ 是使用当前粒子参数组合生成的 LIME 模型的输出，y 是原始模型的输出，目标是最大化该适应度函数。

优化步骤如下：

（1）初始化粒子群，随机选择 LIME 的超参数组合。

（2）通过 LIME 生成局部解释，并计算 R^2 值。

（3）通过粒子群优化算法更新粒子的速度和位置，寻找到最优参数组合。

（4）迭代优化，直到达到最佳的 R^2 值为止。

2.3　稀疏性优化目标与过程

优化稀疏性的目标是通过减少不相关区域的干扰，突出模型决策中最关键的部分，从而提升解释结果的可读性与聚焦程度，使其更符合临床判读习惯。

（1）稀疏化目标

最小化 AreaRatio，即将解释结果尽可能集中于模型决策中最相关的区域。AreaRatio 用于衡量解释区域在整幅图像中的面积占比，是刻画解释结果空间稀疏性与聚焦程度的定量指标，主要反映解释复杂度与可读性特征。其定义如式（3）所示：

$$\text{AreaRatio} = \frac{\text{area of the highlighted region}}{\text{total area of the image}} \tag{3}$$

其中，highlighted region 表示解释结果中被模型关注的区域，total area 为图像的总面积。较大的 AreaRatio 表明解释区域更加集中，有助于突出模型判别所依赖的关键信息，避免解释区域在空间上过度扩展。

（2）优化过程

①通过粒子群优化（PSO）调整 LIME 的关键参数，在保证高 R^2 值的前提下，搜索解释区域覆盖范围较小且聚焦性更强的参数组合；

②在局部扰动与线性拟合过程中，对解释结果的空间分布进行约束，促使每个解释区域主要包含与模型决策高度相关的特征，减少无关区域的干扰；

③通过多代粒子群的迭代更新，使解释结果在保真性与稀疏性之间达到平衡，最终获得空间分布更集中、结构更清晰的解释区域。

2.4　多目标优化模型

高保真性与高稀疏性往往存在内在张力。为实现二者的联合最优与可调平衡，本研究引入多目标粒子群优化，以帕累托最优机制同时优化解释的保真性与稀疏性。

（1）适应度函数

与将两目标线性加权的传统方式不同（会失去帕

累托优势），本研究采用向量形式的双目标适应度函数，如式（4）和式（5）所示：

$$\max f_1(\boldsymbol{p}) = R^2(\hat{y}(\boldsymbol{p}), y) \tag{4}$$

$$\min f_2(\boldsymbol{p}) = \text{AreaRatio}(\boldsymbol{p}) \tag{5}$$

其中：$\boldsymbol{p}$ = [kernel_width，num_samples，num_superpixels，num_features，distance_metric] 为待优化的 LIME 参数向量；R^2 表征解释模型对原模型局部决策的拟合度；AreaRatio 衡量解释区域的聚焦程度。两目标共同约束，使得解释结果既准且简、既忠实于模型又贴近临床观察模式。

（2）优化流程

算法整体流程如图 3 所示。

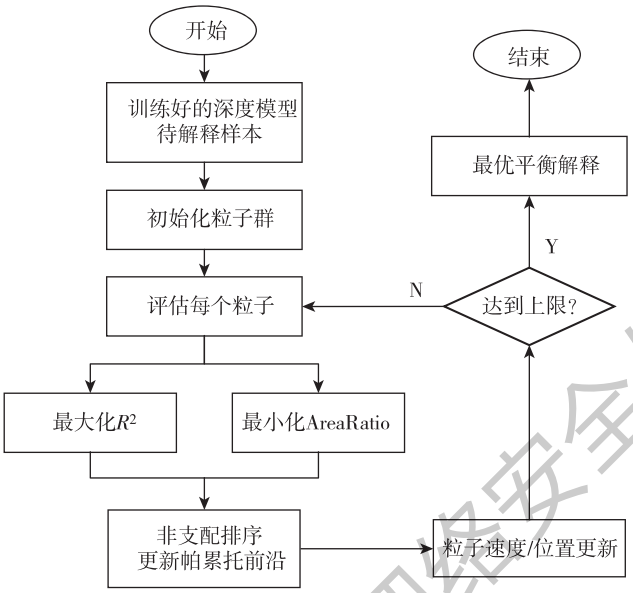

图 3　多目标优化流程图

MOPSO-LIME 多目标解释优化算法的伪代码如算法 1 所示。

算法 1　MOPSO-LIME 多目标解释优化算法

输入：预训练模型 $f(\cdot)$；待解释样本 x；粒子数 N；最大迭代次数 T

输出：帕累托最优解释集 S^*

初始化：

随机初始化粒子群 $P = \{p_1, p_2, \cdots, p_N\}$

// 每个粒子代表一组 LIME 参数

初始化粒子速度 $V = \{v_1, v_2, \cdots, v_N\}$

计算每个粒子的初始适应度（R^2，AreaRatio）

记录个体最优 p_{best} 与全局帕累托集 S^*

for t = 1 to T do

for 每个粒子 p_i do

1）使用参数 p_i 调用 LIME，对样本 x 生成局部扰动

2）拟合线性解释模型，计算 $R^2(p_i)$

3）计算 AreaRatio（p_i）

4）更新粒子适应度 $F_i = (R^2(p_i)$，AreaRatio（p_i））

end for

5）对粒子群执行非支配排序，筛选帕累托前沿解集

6）更新每个粒子的 p_{best}

7）根据群体分布更新全局帕累托集 S^*

8）更新粒子速度与位置：

$v_i = w \cdot v_i + c_1 \cdot \text{rand}() \cdot (p_{\text{best}} - p_i) + c_2 \cdot \text{rand}() \cdot (\text{leader} - p_i)$

$p_i = p_i + v_i$

end for

在所提出的框架中，多目标粒子群优化主要用于解释参数空间的搜索与配置阶段，其计算开销与粒子规模及迭代次数相关。该阶段旨在获得一组在保真性与稀疏性之间具有良好权衡关系的帕累托最优参数组合，用于刻画不同解释质量取向下的候选解集。

在参数配置确定后，具体样本的解释过程可直接调用选定参数生成解释结果，其计算流程与原始 LIME 方法保持一致，不引入额外的推理复杂度。基于上述特性，本文方法侧重于解释质量评估、模型行为分析及临床辅助决策等场景下的解释结果生成与分析。

3　实验与结果分析

3.1　实验说明

本实验旨在系统验证所提出的 MOPSO-LIME 框架在膝骨关节炎 X 光影像解释中的有效性与稳健性。实验数据采用公开 KOA 影像集，按照正常、中度与重度三个病程等级进行标注。为了构建可解释分析的黑箱模型，选择三种具有代表性的卷积神经网络结构作为分类器：VGG16、MobileNetv2 与 ResNet-50。数据集按照训练集、验证集和测试集划分，其中约 80% 用于训练，验证集用于超参数调优与早停策略控制，测试集用于最终性能评估与解释分析。

表 1 展示了三类深度学习模型在膝骨关节炎影像分类任务中的基线性能。评价指标包括精确率、召回率和综合性能指标，分别用于反映模型的误报控制能力、漏报控制能力及整体诊断表现。

表 1　三种分类模型的平均性能对比　（%）

模型	正常	中度	重度
	精确率	召回率	综合性能指标
VGG16	83. 45	77. 83	80. 67
MobileNetv2	82. 16	76. 34	77. 96
ResNet-50	84. 58	88. 96	85. 98

注：综合性能指标为宏平均 F1-score（Macro-F1），由各类别精确率与召回率计算得到，用于综合评价多分类模型的整体性能

从结果来看，ResNet-50 在三个等级任务中均表现最佳，其召回率（88. 96%）显著高于其他模型，说明其能够更有效地识别真实病患，降低漏诊风险，这一特性在临床环境尤为重要。与此同时，其精确率与综合性能指标也保持最高水平，表明其在兼顾误诊与漏诊方面达成良好平衡。相比之下，VGG16 和 MobileNetv2 在召回率上较 ResNet-50 存在 10% ~ 12% 的差距，其中 MobileNetv2 受限于轻量化结构，在性能与模型体积的权衡下表现略有下降，但依旧具备应用价值。

总体而言，三种模型均具备稳定的分类能力，为后续开展可解释性分析提供了可靠基础。特别是 ResNet-50，其优异的诊断性能使其成为评估解释质量的代表性“黑箱”模型，有助于进一步检验 MOPSO-LIME 方法在提升临床信任与模型透明度方面的潜力。

3. 2　多目标优化结果

为系统评估所提出 MOPSO-LIME 框架的解释增强能力，本节从保真性与稀疏性两个维度对比分析原始 LIME 与优化后的解释表现。分别计算局部拟合优度 R^2 以衡量解释对模型决策的逼近程度，并采用 AreaRatio 描述解释区域的聚焦性与临床可读性。

表 2 对比了原始 LIME 与 MOPSO-LIME 在不同 KOA 分级及三类模型上的解释性能。可以观察到，MOPSO-LIME 在所有模型与分级条件下均取得显著提升，表现出良好的稳定性与泛化性。

表 2　基于 MOPSO-LIME 的可解释性指标对比结果

Model	正常		中度		重度	
	R^2	AreaRatio	R^2	AreaRatio	R^2	AreaRatio
VGG16 + LIME	0. 643 9	0. 710 3	0. 523 1	0. 654 2	0. 586 3	0. 697 7
MobileNetv2 + LIME	0. 526 9	0. 673 9	0. 513 6	0. 624 2	0. 634 2	0. 710 6
ResNet-50 + LIME	0. 679 8	0. 747 9	0. 612 5	0. 723 1	0. 712 9	0. 756 2
VGG16 + MOPSO-LIME	0. 758 3	0. 850 7	0. 753 5	0. 864 2	0. 743 2	0. 832 6
MobileNetv2 + MOPSO-LIME	0. 737 7	0. 832 6	0. 721 9	0. 824 4	0. 750 1	0. 849 8
ResNet-50 + MOPSO-LIME	0. 853 6	0. 906 4	0. 842 6	0. 894 2	0. 834 2	0. 891 3

整体来看，保真性指标 R^2 与稀疏性指标 AreaRatio 均显著提升：

（1）VGG16：R^2 提升 17. 8% ~44. 1%，AreaRatio 提升 19. 8% ~32. 1%；

（2）MobileNetv2：R^2 提升 16. 5% ~ 40. 6%，AreaRatio 提升 18. 2% ~35. 7%；

（3）ResNet-50：R^2 提升 18. 9% ~37. 6%，AreaRatio 提升 18. 5% ~28. 4%。

其中 ResNet-50 在原始 LIME 方案中已表现优异，说明其特征表示较为稳健。MOPSO-LIME 进一步提升其解释一致性与区域聚焦能力，使其解释更加符合临床视觉关注模式。

VGG16 与 MobileNetv2 在未优化条件下解释质量相对较低，但经过 MOPSO 调节后均实现显著提升，证明该方法对不同复杂度的网络均适用，并可有效增强轻量网络的解释性能。

从病情分级维度看，原始 LIME 在中度和重度病例中性能下降，而 MOPSO-LIME 在各等级病变上均保持高性能，说明该方法对复杂病灶特征具有更强适应性与鲁棒性。

总体而言，MOPSO-LIME 可显著提升解释与模型内部逻辑的一致性，并有效突出关键病变区域，在不改变原有诊断模型结构的前提下，实现解释可信度和临床友好性双提升，为 KOA 自动辅助诊断提供了更具可信价值的解释机制。

3. 3　与主流可解释性方法的对比分析

为系统评估 MOPSO-LIME 在医学影像解释任务中的性能，本文选取多类具有代表性的事后可解释性方法作为对比基线，包括梯度归因方法、集成归因方法以及模型无关的局部解释方法，具体为基于梯度定位

的 Grad-CAM 与 Grad-CAM ++[5-6]、积分梯度法 Integrated Gradients[9]、基于 Shapley 值的 SHAP[8]以及原始 LIME[7]。上述方法在医学影像可解释性研究中被广泛采用，能够从不同归因范式刻画模型的关注区域，为解释效果的横向比较提供基础参照。

为避免不同特征提取主干网络结构带来的潜在干扰，所有方法均统一在本文分类性能最优的 ResNet-50 模型上生成解释结果。评估指标采用本文提出的保真性指标 R^2 与稀疏性指标 AreaRatio，并进一步引入临床相关性评分，用于衡量解释区域与放射科医生实际关注病灶（如关节间隙变窄与骨赘）的空间一致性。结果如表 3 所示。

表 3　基于 ResNet-50 模型 MOPSO-LIME 与主流可解释方法性能对比

病理分级	方法	R^2（均值 ± 标准差）	AreaRatio（均值 ± 标准差）	临床相关性评分
正常	Grad-CAM	0.75 ± 0.04	0.72 ± 0.05	3.5
	Grad-CAM ++	0.77 ± 0.03	0.74 ± 0.04	3.7
	SHAP	0.73 ± 0.05	0.70 ± 0.06	3.7
	Integrated Gradients	0.74 ± 0.04	0.72 ± 0.03	3.8
	LIME	0.72 ± 0.06	0.78 ± 0.07	3.0
	MOPSO-LIME	0.88 ± 0.02	0.88 ± 0.02	4.3
中度	Grad-CAM	0.70 ± 0.06	0.68 ± 0.08	3.1
	Grad-CAM ++	0.72 ± 0.05	0.70 ± 0.07	3.4
	SHAP	0.66 ± 0.07	0.63 ± 0.09	2.9
	Integrated Gradients	0.73 ± 0.02	0.71 ± 0.07	3.6
	LIME	0.61 ± 0.09	0.63 ± 0.09	2.7
	MOPSO-LIME	0.83 ± 0.03	0.89 ± 0.04	4.0
重度	Grad-CAM	0.68 ± 0.07	0.66 ± 0.09	3.0
	Grad-CAM ++	0.70 ± 0.06	0.69 ± 0.08	3.3
	SHAP	0.69 ± 0.08	0.69 ± 0.08	2.8
	Integrated Gradients	0.71 ± 0.05	0.74 ± 0.06	3.5
	LIME	0.67 ± 0.10	0.73 ± 0.11	2.8
	MOPSO-LIME	0.82 ± 0.03	0.82 ± 0.03	4.1

注：临床相关性评分基于解释区域与膝骨关节炎典型影像学征象的空间一致性及可理解性进行 1 ~ 5 分评估（5 分最高），取平均值

如表 3 所示，MOPSO-LIME 在正常、中度及重度三类病例中均取得最高的 R^2 与 AreaRatio 值，且标准差显著低于其他方法，这表明 MOPSO-LIME 生成的解释在保持高保真性的同时具有更强的聚焦性与稳定性。与 Grad-CAM 系列方法相比，MOPSO-LIME 能更准确地定位关节间隙变窄、骨赘边缘等 KOA 典型病变区域，而 Grad-CAM 在部分样本中出现解释区域扩散或受深层梯度噪声影响的情况。SHAP 与 Integrated Gradients 作为像素级归因方法，尽管在局部细节呈现上具有一定优势，但在本任务中可能存在解释碎片化或高亮区域分散的问题，使其整体可读性与空间一致性相对较弱。

在临床相关性方面，MOPSO-LIME 获得最高评分（平均 4.1 分），显著优于其他方法，这表明其解释热图与真实病灶（如关节间隙、骨赘）的空间重合度更高，更符合放射科医生的诊断习惯。综合量化指标与临床评分的综合表现，可以看出 MOPSO-LIME 不仅在内部保真性与稀疏性方面实现显著提升，也在与 Grad-CAM、SHAP、Integrated Gradients 等主流方法的比较中展现出更优的解释质量，进一步证明其在医学影像场景中具有更高的可读性、可靠性与临床应用潜力。

3.4　典型病例可视化

在定量指标验证解释性能提升的基础上，本节进一步展示典型病例的可视化解释结果，以直观对比原始 LIME 与 MOPSO-LIME 在关注区域、稀疏性与临床可信性上的差异。通过展示不同严重程度 KOA 样本及不同架构模型的解释热图，本节旨在观察解释区域与关键病变征象（如关节间隙变窄、骨赘形成等）的空间一致性，从视觉层面验证所提方法的临床可读性与病灶定位能力。

图 4 展示了不同模型在典型膝骨关节炎病例上的解释热图，可进一步验证所提出方法的视觉表达能力与临床一致性。从结果可以看到，MOPSO-LIME 能显著提升各模型解释区域的聚焦度与稳定性。

对于 VGG16（图 4 第二行），热力图整体覆盖范围较大，但能够明确突出关节间隙狭窄等关键区域，解释区域集中性明显提升，且噪声区域较原始 LIME 方法显著减少，与其较高的 AreaRatio 和 R^2 指标一致，说明优化后模型在视觉可解释性上获得有效增强。

MobileNetv2（图 4 第三行）由于模型结构轻量化，原始解释较为分散且一致性略弱。经 MOPSO 优化后，热图聚焦能力有所改善，重要区域更加凸显，但仍可观察到一定背景干扰，反映其解释性能受限于基础特征表达能力。

ResNet-50（图 4 第四行）整体表现最佳，优化后解释区域最为集中且紧凑，精确覆盖骨赘形成与关节间隙变窄等典型病变部位，热图稳定性与清晰度显著优于其他模型。这与其在 R^2、AreaRatio 及分类性能上的最优表现一致，说明性能更强的基模型更易生成临

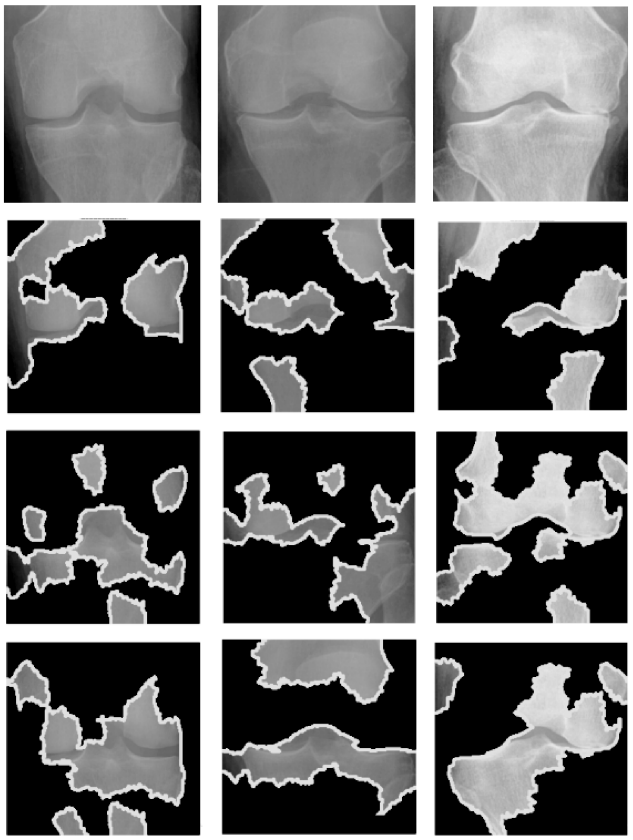

图 4　典型膝骨关节炎病例的 MOPSO-LIME 解释可视化结果

床可信的解释。

综上，MOPSO-LIME 在不同网络架构中均显著提升了解释结果的可视化质量，并能够更加有效地聚焦膝骨关节炎病灶区域，尤其在高性能网络（如 ResNet-50）中更为突出，体现出良好的的模型适配性和解释稳定性。该方法通过离线多目标优化对解释参数进行系统搜索与配置，侧重于解释质量的精细控制与稳定性提升，适用于模型行为分析与解释可信性评估等应用场景。通过获得高保真、高聚焦的解释结果，MOPSO-LIME 能够为膝骨关节炎辅助诊断提供更加稳定、可读且可信的视觉解释支持。

4　结论

本研究针对 KOA X 光影像诊断中深度模型“高准确 - 难解释”的现实矛盾，提出了一种基于多目标粒子群优化的解释增强框架。在固定诊断网络结构的前提下，本文将 LIME 的关键解释参数转化为可优化变量，并从“局部拟合优度（保真性）- 解释稀疏性（聚焦性）”两个维度对解释过程进行联合建模与搜索，从而实现解释质量的自动平衡与优化。实验结果表明，在不影响原始模型诊断性能的前提下，所提出的 MOPSO-LIME 方法能够显著提升解释结果与真实病灶区域的一致性，减少冗余高亮区域，并增强解释输出在不同样本与模型结构下的稳定性与可读性。相关结果验证了通过优化视角重构解释生成过程，在无需修改诊断模型结构的情况下提升解释可信度与工程可用性的可行性。

从方法论角度看，本文工作并非提出新的解释机制，而是提供了一种将解释质量需求形式化、可量化并可调节的工程化路径。通过构建“解释质量量化 - 多目标优化搜索 - 定量与可视化联合验证”的闭环框架，本文为医学影像可解释性研究提供了一种兼顾可复现性与临床可读性的实践范式，也为后续解释方法的系统评估与改进提供了参考思路。

未来工作将围绕以下方向拓展研究：

（1）将所提出的优化框架推广至多模态医学数据及自监督学习模型；

（2）探索引入医学先验知识与病理结构约束，以增强解释结果的结构一致性与临床语义关联；

（3）结合临床医生的交互反馈机制，构建人机协同的解释验证与评估流程。

参考文献

[1] CHEN H, GOMEZ C, HUANG C M, et al. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review[J]. NPJ Digital Medicine, 2022, 5 (1): 156.

[2] 程伟彬, 李观明. 医学人工智能：从技术性能到临床效用[J]. 广东医学, 2025, 46 (11): 1601 - 1605.

[3] NASARIAN E, ALIZADEHSANI R, ACHARYA U R, et al. Designing interpretable ML system to enhance trust in healthcare: a systematic review to proposed responsible clinician-AI-collaboration framework[J]. Information Fusion, 2024, 108: 102412.

[4] TIULPIN A, THEVENOT J, RAHTU E, et al. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach[J]. Scientific Reports, 2018, 8 (1): 1727.

[5] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 618 - 626.

[6] CHATTOPADHAY A, SARKAR A, HOWLADER P, et al. Grad-CAM ++: generalized gradient-based visual explanations for deep convolutional networks[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 839 - 847.

[7] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?" Explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 1135 - 1144.

[8] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]//Advances in Neural Information Processing Systems, 2017.

[9] SUNDARARAJAN M, TALY A, YAN Q. Axiomatic attribution for deep networks[C]//International Conference on Machine Learning. PMLR, 2017: 3319 - 3328.

[10] ADEBAYO J, GILMER J, MUELLY M, et al. Sanity checks for saliency maps[J]. Advances in Neural Information Processing systems, 2018, 31.

[11] ARUN N, GAW N, SINGH P, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging[J]. Radiology: Artificial Intelligence, 2021, 3 (6): e200267.

[12] NAUTA M, TRIENES J, PATHAK S, et al. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI[J]. ACM Computing Surveys, 2023, 55 (13s): 1 - 42.

[13] HEDSTRÖM A, WEBER L, KRAKOWCZYK D, et al. Quantus: an explainable ai toolkit for responsible evaluation of neural network explanations and beyond[J]. Journal of Machine Learning Research, 2023, 24 (34): 1 - 11.

[14] PATRÍCIO C, NEVES J C, TEIXEIRA L F. Explainable deep learning methods in medical image classification: a survey[J]. ACM Computing Surveys, 2023, 56 (4): 1 - 41.

[15] KENNEDY J, EBERHART R. Particle swarm optimization[C]// Proceedings of ICNN'95-International Conference on Neural Networks. IEEE, 1995, 4: 1942 - 1948.

[16] 高宇. 面向医学图像分类的可解释性与公平性增强方法研究 [D]. 北京: 北京科技大学, 2025.

[17] KIM B, WATTENBERG M, GILMER J, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (tcav) [C]//International Conference on Machine Learning. PMLR, 2018: 2668 - 2677.

[18] PATRÍCIO C, NEVES J C, TEIXEIRA L F. Coherent concept-based explanations in medical image and its application to skin lesion diagnosis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 3799 - 3808.

[19] ZHANG Z, XIE Y, XING F, et al. Mdnet: a semantically and visually interpretable medical image diagnosis network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6428 - 6436.

[20] MISERA L, MÜLLER-FRANZES G, TRUHN D, et al. Weakly supervised deep learning in radiology[J]. Radiology, 2024, 312 (1): e232085.

(收稿日期: 2026 - 01 - 15)

作者简介:

李海芳 (1986 -), 女, 硕士, 讲师, 主要研究方向: 可解释人工智能、时间序列预测。

唐超 (1996 -), 男, 硕士研究生, 主要研究方向: 可解释人工智能。

张强 (1971 -), 通信作者, 男, 博士, 教授, 主要研究方向: 生物计算、人工智能。E-mail: zhangq@dlut.edu.cn。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部