arXiv:2304.01717v1 [stat.ML] 4 Apr 2023

刻画相关特征对可解释人工智能方法的贡献

AhmedSalih[a, b], Ilaria Boscolo Galazzo[c], Zahra Raisi-Estabragh[a, d], Steffen E. Petersen[a, d, e, f], Gloria Menegaz *[c], Petia Radeva *[g, h]

[a] 威廉·哈维研究所、*NIHR Barts* 生物医学研究中心、伦敦玛丽女王大学，*Charterhouse Square*，伦敦，*EC1M 6BQ*，英国伦敦[b]计算机科学系、理学院、扎胡大学，*Zakho Way Road*，杜胡克，*42002*，伊拉克库尔德斯坦地区[c]创新医学工程系、维罗纳大学，*Via S. Francesco*，*22*，维罗纳，*37129*，意大利维罗纳[d]*Barts* 心脏中心、圣巴塞洛缪医院、*Barts Health* 国家医疗服务体系信托，*West Smithfield*，伦敦，*EC1A 7BE*，英国伦敦[e]健康数据研究，伦敦，英国伦敦[f]艾伦·图灵研究所，伦敦，英国伦敦[g]数学`与计算机科学系、巴塞罗那大学，加泰罗尼亚大道，巴塞罗那，*08007*，西班牙巴塞罗那[h]计算机视觉中心，*Cerdanyola del Vall´es*，巴塞罗那，西班牙巴塞罗那[i]

*：*Gloria* 和 *Petia* 共同担任首席研究员，

摘要

可解释人工智能（XAI）提供了有助于理解机器学习模型如何运行并得到特定结果的工具。它有助于提高模型的可解释性，使模型更加值得信赖且更加透明。在此背景下，人们提出了许多可解释人工智能方法，其中 SHAP 和 LIME 最为流行。然而，所提方法假设机器学习模型中使用的预测变量彼此独立，而这一点通常并不一定成立。这一假设对可解释人工智能结果的鲁棒性提出了质疑，例如信息量较大的预测变量列表。为此，我们提出了一种简单但实用的代理指标，用于修改任何可解释人工智能特征排序方法的结果，从而将预测变量之间的依赖关系纳入考量。所提方法的优势在于其模型无关性，同时能够在存在共线性的情况下，以简单的方式计算每个预测变量对模型的影响。

关键词：

提交至爱思唯尔的预印本 *2023*年*4*月*5*日

1. 引言

可解释人工智能（可解释人工智能）旨在揭开机器学习模型的奥秘，尤其是深度神经网络和传统神经网络等先进而复杂的模型。可解释人工智能有助于理解这些模型如何得出特定结果、图像中的每个特征或区域如何影响模型输出，以及模型对所得决策的确定程度 [1]。这些概念和原则对最终用户不可或缺，因为它们能增强用户对模型结果的信任，并提高其透明度 [2]。目前已有许多可解释人工智能方法被提出，用于将机器学习模型分解为更易于解释的模型。

SHAP（SHapley Additive exPlanations）是一种基于博弈论提出的模型无关可解释人工智能方法 [3]。它将模型中的每个特征视为一名参与者，并将结果视为收益。随后，它根据博弈论／联盟计算 SHAP 分数，从而计算每个特征对模型结果的边际效应。局部可解释模型无关解释（LIME）[4] 是另一种可解释人工智能方法，用于解释模型在单个实例上的局部表现。它展示了每个特征对特定对象的模型结果有何贡献，以及模型在多大程度上具有确定性。部分依赖图是另一种可解释人工智能方法，用于展示某个特征对模型结果的边际效应 [5]。它展示了这种关系是单调的、线性的，还是更为复杂。研究还考虑了经典方法，通过报告模型中特征的权重（系数）来揭示特征对输出的影响 [6]。

可解释人工智能方法面临的主要问题之一，是所使用预测变量之间存在依赖关系。当模型提供信息性最强的预测变量时，这会对可解释人工智能的结果产生显著影响。例如，SHAP 在计算每个预测变量的得分（影响）时，假设特征彼此独立 [7]。因此，如果两个特征相关，其中一个会获得较高得分，而另一个得分较低，因为由于与另一个预测变量存在依赖关系，它并未提升模型性能。同样，LIME 和经典机器学习模型使用模型中每个特征的权重来揭示其影响。该权重表示在保持所有

其他预测变量在模型中保持不变时，一个特征对结果的依赖关系。这种方法并不精确，因为模型中的一组预测变量（共线变量）可能会同时发生变化。在实际应用中，模型中所使用的预测变量彼此独立的情况非常罕见。因此，可解释人工智能方法的结果（例如信息性预测变量列表）应考虑这种依赖关系，以提供更稳健的列表。本文提出一种新的代理指标/方法，该方法获取任意特征排序可解释人工智能方法的结果，并根据预测变量之间的共线性对其进行调整，从而生成模型中信息性最强特征的新排序列表。

2. 研究现状

近年来，多项研究开始探讨预测变量之间的依赖关系及其影响。Kjersti Aas 等人 [7]提出了四种方法，以处理 Kernel SHAP [3]中的共线性问题。Kernel SHAP 的主要贡献在于，特征子集 ς 的条件分布取决于S中的其余特征，其中S 是全部特征的集合。在 Kernel SHAP 的原始形式中，作者隐含地假设特征彼此独立，因此可以用预测变量的边际分布替代条件分布 ς。Kjersti Aas 等人 [7] 通过四种基于特征分布的方法，对所有特征的分布进行估计，从而改进了原始 Kernel SHAP。这些方法涵盖高斯分布、高斯 copula 分布、经验条件分布，以及经验方法与高斯分布和/或高斯 copula 方法的组合。尽管该方法解决了一个重要问题，但其计算成本较高，影响了此类方法的实用性。此外，该方法旨在解释单个预测（局部解释），而非全局解释；全局解释展示模型中所有实例的每个特征对整体结果的贡献。此外，所选分布可能会影响方法得出的结果，这意味着最终用户应了解所用数据的分布，以便选择恰当的分布；但一般而言，这一前提并不成立。

另一种用于处理预测变量之间依赖关系的方法被提出，称为 Shapley 队列细化 [8]。该方法首先创建一个相似性队列，将靠近目标实例（需要解释的实例）的实例（数据点）结合起来，然后对新的队列应用 Shapley 方法。该方法将对象分为与目标对象相似或不相似两类。

相似性通过比较每个对象与目标对象的预测变量值来衡量。如果预测变量是二元变量，则很容易比较并识别相似对象。如果预测变量是连续变量，他们会使用基于距离的方法来识别与目标对象相似的对象。所选的相似性度量将对识别模型中每个预测变量的重要性产生重大影响。该方法具有许多优点，包括使用真实数据（预测变量），并且当两个预测变量高度相关时，会为它们给出相似的得分。然而，该方法严重依赖所选的相似性度量，而相似性度量会影响该方法的结果。

3. 材料与方法

3.1. 所提出的方法

评估任何可解释人工智能方法的一个重要方面，是该可解释人工智能方法或代理指标的简洁性。对于可解释人工智能或评估标准而言，简洁至关重要，因为大多数终端用户并非可解释人工智能方法方面的专家，而是来自不同的领域或应用场景。任何领域的终端用户都需要理解可解释人工智能的工作原理，从而增强对可解释人工智能结果的信任。事实上，这正是提出可解释人工智能时的目的：增强对机器学习模型的信任。在此，我们旨在提出一种简单的方法，以简单的方式修改任何可解释人工智能方法的结果，从而能够处理特征之间的共线性，同时易于理解。

所提出的方法是对先前提出的用于评估可解释人工智能方法稳定性的代理指标的扩展 [9]。在先前的研究中，研究者提出了一项名为归一化变动率（NMR）的新指标，用于衡量从模型中迭代移除排名最高的预测变量时，任意可解释人工智能方法生成的信息性预测变量列表会受到何种影响。NMR值介于0和1之间：0表示信息性预测变量列表具有稳健性，而接近1则表示该列表不稳定。这是由预测变量之间的依赖关系造成的，也是该方法的主要弱点。NMR值表明，应用主成分分析（PCA）后，可解释人工智能方法生成的结果（排名靠前的预测变量列表）更加稳定、稳健。在该研究中，作者仅提出了一项新指标，用于诊断可解释人工智能结果因预测变量之间的共线性而产生的不稳定性。在本文中，我们更进一步，提出了一种修改可解释人工智能结果的新方法

从而克服这一局限，并将预测变量之间的依赖关系纳入考量。

与 [9], 类似，我们迭代地移除排名靠前的预测变量，并重新训练模型，然后在测试数据上应用可解释人工智能方法，以生成信息性预测变量列表。这里我们引入一个名为修正信息位置（MIP）的新指标。我们持续移除预测变量、重新训练模型并应用所选的可解释人工智能方法，直到模型中剩下 2 个预测变量。每次迭代（移除排名最高的预测变量）时，我们将每个预测变量的索引（表示其重要性）除以当前模型中的预测变量数量进行归一化。最后，我们将每次迭代中每个预测变量索引与该次迭代中的预测变量数量相除所得的结果求和，从而计算每个预测变量的得分。最后，按照得分升序对预测变量排序。得分较低的预测变量信息量更大，因为它们比其他预测变量更快进入排名靠前的列表；得分较高的预测变量信息量较低，因为它们在大多数迭代中都存在于模型中。方程 1 展示了如何计算模型中每个预测变量的 MIP。

$$\mathrm{MIP} = \sum_{j=1}^{NF-2} x_{p,j} \tag{1}$$

其中，j 是迭代次数，NF 是预测变量的数量，x_p 是将索引 i 除以每次迭代中的预测变量数量 p 所得的结果。算法 1 展示了所提方法的伪代码。我们还使用特征的得分计算标准差（SD）。SD 越高，列表越稳健，NMR 越小。在最佳情况下，SD 将达到最大值，而 NMR 将达到最小值（接近零）。此外，如果模型中的特征彼此独立（通常并非如此），可解释人工智能（例如 SHAP）的结果将与所提方法的结果非常相似。

算法 1 用于计算 NMR 和 MIP 值的伪锥。蓝色字体标识相对于先前方法新增的步骤。确保：NF = 特征总数 特征数量 要求：训练模型 要求：测试模型 要求：应用 SHAP，并按特征重要性降序排列 while $NF \neq 2$ do 将每个预测变量的位置（索引）除以特征数量：xp= ip/NF，其中 i 是预测变量 p 的索引 移除信息量最大的预测变量 应用 SHAP，并按特征重要性降序排列 检查与上一步列表相比有多少个预测变量改变了位置 根据索引变化计算预测变量的移动量，并将其求和（M）计算最大可能移动量： $MPM = 2\sum_{i=1}^{i+2}(NF - i)$ for$i < NF$ 计算移动率： $MR = M/MPM$end while$NMR = \sum(MR)/NF_{total} - 2$要求：计算 P 中 p 的修正信息位置（MIP），其中 P 是预测变量集合 do$MIP = \sum_{j=1}^{NF-2} x_{p,j}$if p存在，且 j是迭代次数 end for

3.2. 数据

为说明所提方法，我们使用了英国生物样本库的数据（申请编号 2964）。所用数据包含从心脏磁共振成像中提取的九项心脏表型，包括左心室（LV）舒张末期和收缩末期容积（分别为 LVEDV 和 LVESV）、左心室每搏量（LVSV）、左心室射血分数（LVEF）、左心室质量（LVM）、右心室（RV）舒张末期和收缩末期容积（分别为 RVEDV 和 RVESV）、右心室每搏量（RVSV）以及右心室射血分数（RVEF），共涉及 2,000 名研究对象（其中女性 1,032 名）。这九项表型被用作预测变量，将研究对象分为男性和女性。我们有意选择这些特征，因为生理性别会导致其数值出现差异 [10]。

3.3. 实现

在 [9], 中，研究表明，可解释人工智能的结果取决于模型，这意味着不同模型处理预测变量之间共线性的方式可能有所不同，进而使可解释人工智能方法生成不同的排序。因此，我们选择了一项结果为二元变量的简单任务，并评估了五种分类器中特征排序的稳定性。所考虑的分类器包括：支持向量分类器（SVC）、轻量级梯度提升机（LGBM）、决策树（DT）、随机森林（RF）和逻辑回归（LR）。数据被划分为训练数据（用于训练模型）和未见数据（用于测试模型，占 20%）。为确定五种模型各自的最优参数，我们在训练数据上进行了超参数调优，采用 10 折交叉验证，并为每个模型设置了一组参数（例如，决策树的最大深度、支持向量分类器中的 C 等）。我们以准确率作为选择各模型最优参数的标准。随后，使用参数最优的模型在测试数据上进行测试。最后，采用 SHAP 作为可解释人工智能方法，生成信息性预测变量列表，以揭示哪些预测变量对模型结果的影响更大。

3.4. 验证

为验证所提方法，我们对原始预测变量应用主成分分析（PCA），生成相互独立的成分。我们自动选择成分数量，使其能够解释 95% 的方差，最终得到 4 个主成分（PCs）。随后，我们将这些成分输入五个模型，方法与输入原始特征时相同。最后，应用 SHAP 根据贡献度对主成分进行排序。

3.5. 生物学合理性评估

为了在临床上评估所提方法，我们使用了表 1 [11] 中报告的九项测量指标的绝对性别差异值。我们根据性别差异的绝对值按降序排列特征，得分越高，表明该特征对于区分男性和女性越重要。我们将此列表称为基准，以区别于 SHAP 列表和所提方法的列表。随后，我们计算了基准、SHAP 以及所提方法中九个特征排序之间的 Kendall τ-b 秩相关和 Pearson相关，以揭示 SHAP 列表或所提方法列表是否与基准顺序更为一致。

是否与基准顺序更为一致。

4. 结果

所提方法使用 Python 3.9 实现，并可在线访问：`https://github.com/amaa11/MIP`。为展示所用特征之间的共线性程度，图1展示了这些特征之间的相关性。结果表明，大多数所用特征之间都存在较强的正相关或负相关。若未能充分处理共线性问题，这一问题势必会影响可解释人工智能方法的结果。

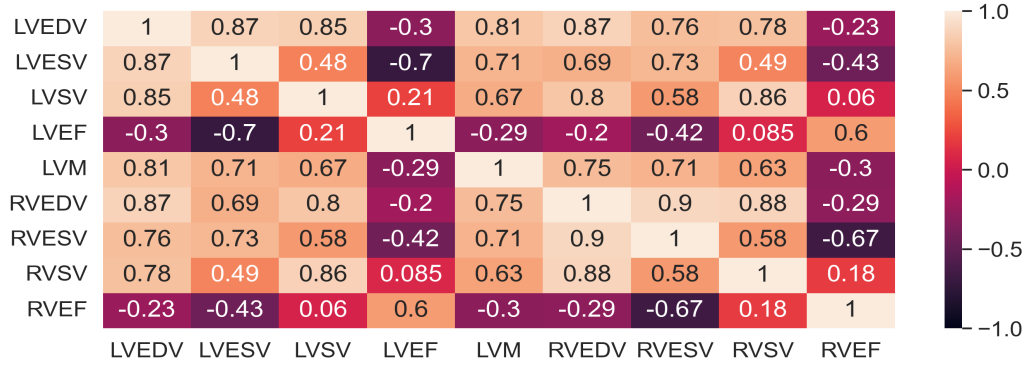

图1：所用特征的相关矩阵。

表1列出了 SHAP 针对每个模型生成的信息性预测变量，并展示了基于所提方法修正后的列表。为便于比较每个模型中各预测变量的位置，表中将特征的顺序编码为数字。表中结果显示，SHAP 的输出依赖于模型：尽管五个模型的准确率非常接近（≈ 0.84），每个模型生成的列表却各不相同。这是因为所用模型处理共线性的方式各不相同，因此 SHAP 会产生不同的结果。这引出了一个问题：应该选择其中哪一个列表？不过，每个模型的 NMR 值各不相同，其中最小值对应的列表更为稳健。

表1：SHAP 为所使用的每个模型生成的信息性预测变量排序列表。预测变量编码为 LVM（1）、RVEDV（2）、RVESV（3）、RVEF（4）、LVEDV（5）、LVSV（6）、LVESV（7）、LVEF（8）、RVSV（9）。

SHAP				
LBGM	**DT**	**LR**	**RF**	**SVC**
LVM	LVM	LVM	LVM	LVM
右心室舒张末期容积	右心室收缩末期容积	右心室舒张末期容积	右心室收缩末期容积	右心室舒张末期容积
RVESV	RVEDV	RVEF	RVEDV	LVEDV
RVEF	LVSV	LVEDV	LVEDV	RVEF
LVEDV	LVESV	RVSV	左心室收缩末期容积	LVEF
LVSV	LVEF	左心室收缩末期容积	RVSV	RVSV
RVSV	RVSV	LVEF	RVEF	左心室收缩末期容积
左心室收缩末期容积	RVEF	右心室收缩末期容积	LVEF	右心室收缩末期容积
LVEF	左心室舒张末期容积	LVSV	LVSV	LVSV
LBGM	**DT**	**LR**	**RF**	**SVC**
1	1	1	1	1
2	3	2	3	2
3	2	4	2	5
4	6	5	5	4
5	7	9	7	8
6	8	7	9	9
9	9	8	4	7
7	4	3	8	3
8	5	6	6	6
修改后的SHAP				
LBGM	**DT**	**LR**	**RF**	**SVC**
LVM	LVM	LVM	LVM	LVM
右心室舒张末期容积	右心室舒张末期容积	右心室舒张末期容积	右心室舒张末期容积	右心室舒张末期容积
右心室收缩末期容积	右心室收缩末期容积	RVSV	右心室收缩末期容积	RVSV
RVSV	左心室舒张末期容积	右心室收缩末期容积	左心室舒张末期容积	右心室收缩末期容积
左心室舒张末期容积	左心室收缩末期容积	左心室舒张末期容积	左心室收缩末期容积	左心室舒张末期容积
左心室收缩末期容积	RVSV	RVEF	RVSV	RVEF
RVEF	LVSV	LVSV	LVSV	LVSV
LVSV	RVEF	左心室收缩末期容积	RVEF	左心室收缩末期容积
LVEF	LVEF	LVEF	LVEF	LVEF
LBGM	**DT**	**LR**	**RF**	**SVC**
1	1	1	1	1
2	2	2	2	2
3	3	9	3	9
9	5	3	5	3
5	7	5	7	5
7	9	4	9	4
4	6	6	6	6
6	4	7	4	7
8	8	8	8	8

表2展示了一个考虑共线性后如何重新排列列表的示例。具体步骤是：在每次迭代中，移除排名靠前的特征，然后将每个预测变量的索引（由SHAP生成）除以该次迭代中模型的预测变量数量。以第一次迭代（j）为例，当模型中有9个预测变量（NF）时，LVSV的索引（i）为4，将其除以预测变量数量9，所得得分为0.44。随后，为计算同一预测变量的得分，我们将各次迭代中索引除以特征数量所得的结果相加，具体如方程1（0.44+ 0.75+ 0.71+ 0.8+ 1+ 0.5+ 0.33）所述。其他预测变量的得分也采用相同方式计算。表格中的最后两列（橙色背景）按升序展示了每个预测变量的得分。得分越小，说明该预测变量提供的信息越多；反之，得分越高，说明该预测变量的信息量越少。得分表示在每次移除排名最靠前的预测变量后，各预测变量以多快的速度（显著性）到达列表顶部。之后，计算每个预测变量得分的标准差。通过这一指标，可以将产生最大方差的方法识别为最稳定的方法。SD越高，意味着各特征正按顺序接近列表顶部。这也意味着，一组特征以相同步数接近列表顶部的可能性更小（共线性情况下则相反）。这也与此前提出的代理指标（NMR）一致。产生最低NMR值的模型最稳定。

表 2：展示如何计算每个预测变量得分的示例。最后一行表示每次迭代中模型所包含的预测变量数量。

SHAP结果							
LVM	0.11						
右心室收缩末期容积	0.22	右心室舒张末期容					
右心室舒张末期容积	0.33	右心室收缩末期容积	0.25 右心				
LVSV	0.44	左心室收缩末期容积	0.38 左心				
左心室收缩末期容积	0.56	RVEF	0.50	左心室收缩末期容积	0.43	RVEF	0.3
LVEF	0.67	左心室舒张末期容积	0.63	RVEF	0.57 左心		0.5
RVSV	0.78	LVSV	0.75	LVSV	0.71	RVSV	0.7
RVEF	0.89	RVSV	0.88	RVSV	0.86	LVSV	0.8
左心室舒张末期容积	1.0	LVEF	1.0	LVEF	1.0	LVEF	1.0
9		8		7		6	
SHAP结果							
RVSV	0.2						
左心室收缩末期容积	0.4	左心室收缩末期容积	0.25				
LVEF	0.6	LVSV	0.5	LVSV	0.33		
RVEF	0.8	LVEF	0.75	LVEF	0.67	RVEF	0.5
LVSV	1	RVEF	1	RVEF	1.00	LVEF	1
5		4		3		2	

SHAP结果	修正SHAP	得分
LVM	LVM	0.11
右心室收缩末期容积	右心室舒张末期容积	0.46
右心室舒张末期容积	右心室收缩末期容积	0.47
LVSV	左心室舒张末期容积	2.12
左心室收缩末期容积	左心室收缩末期容积	2.52
LVEF	RVSV	3.42
RVSV	LVSV	4.53
RVEF	RVEF	5.56
左心室舒张末期容积	LVEF	6.69

表3展示了每个模型中预测变量得分的标准差。随机森林模型的SD最大值为2.71。这表明，与其他模型相比，修改后的SHAP列表在随机森林模型中更加稳定。这也与每个模型的NMR值一致，其最小值为

随机森林模型。

表3：每个模型的 NMR 和标准差（SD）。

SD		NMR	
模型	**修改后 SHAP**	**模型SHAP**	
DT	2.36	DT	0.43
LBGM	2.41	LBGM	0.49
LR	2.11	LR	0.57
RF	**2.71**	RF	**0.14**
SVC	2.16	SVC	0.58

为揭示预测变量之间的依赖关系所产生的影响，表4列出了 SHAP 针对所使用的五种模型生成的各组成部分的排序。结果显示，四种模型（随机森林、LBGM、支持向量分类器和逻辑回归）得出的结果完全相同，而决策树略有不同。此外，表中显示，所提方法在这四种模型中的结果与 SHAP 结果完全相同，而在决策树模型中则略有不匹配。这些结果证实了这一假设：当预测变量相互独立时，所提方法的结果将与 SHAP 的结果相似。

表 4：SHAP 为每个使用主成分的模型生成的信息性预测变量排序列表。

DT		RF, LBGM, 支持向量分类器和逻辑回归	
提出的 SHAP	**提出**		**SHAP**
PC1	PC1	PC1	PC1
PC4	PC4	PC4	PC4
PC2	PC3	PC2	PC2
PC3	PC2	PC3	PC3

表5中说明了九个特征的代码及其排序。基准列表与 SHAP 之间的 Kendall τ-b 秩相关值为 0.38（p 值=0.1），而基准列表与所提方法之间的相关性为 0.61（p 值 $= 0.02$）。此外，基准列表与 SHAP 列表之间的皮尔逊相关为 0.58（p 值$= 0.09$），而基准顺序与所提方法之间的相关值为 0.70（p 值$= 0.03$）。

表5：基于基准、SHAP 和所提方法得到的信息性预测变量排序列表。预测变量编码如下：LVM（1）、RVESV（2）、LVESV（3）、LVEDV（4）、LVSV（5）、RVEDV（6）、RVSV（7）、RVEF（8）、LVEF（9）。

实际顺序			编码顺序		
基准 SHAP		所提方法	基准 SHAP	所提方法	
RVESV	LVM	LVM	2	1	1
LVM	右心室收缩末期容积	右心室舒张末期容积	1	2	6
左心室收缩末期容积	右心室舒张末期容积	右心室收缩末期容积	3	6	2
左心室舒张末期容积	LVSV	左心室舒张末期容积	4	5	4
LVSV	左心室收缩末期容积	左心室收缩末期容积	5	3	3
右心室舒张末期容积	LVEF	RVSV	6	9	7
RVSV	RVSV	LVSV	7	7	5
RVEF	RVEF	RVEF	8	8	8
LVEF	左心室舒张末期容积	LVEF	9	4	9

5. 讨论

在实际应用中，预测变量往往存在高度共线性。必须考虑并妥善处理这一问题，尤其是在医学和生物学领域，因为错误识别信息性变量可能导致严重后果。可解释人工智能方法的另一个相关方面是简洁性，它有助于提升可理解性。大多数最终用户既不是机器学习模型方面的专家，也不是可解释人工智能方法方面的专家，而是非专业用户。因此，必须以简单易懂的方式介绍用于解释模型的方法和途径，以便用户理解并正确实施。我们提出了一种简单的方法，通过修改可解释人工智能的结果来考虑预测变量之间的依赖关系，从而生成更稳健的列表。所提方法基于这样一个事实：如果模型中的两个预测变量存在共线性，并且它们对结果的影响程度相近，那么它们在信息性预测变量列表中的重要性也应相近。为实现这一目标，我们会移除模型中排名最高的预测变量，重新训练并测试模型，然后检查各预测变量的位置。所提方法与模型无关，这意味着它可应用于任何在结果中未考虑预测变量之间共线性的可解释人工智能方法。

与以往提出的方法相比，我们所提方法

具有若干优势。首先，所提方法简单易懂，因为它无需像 [7] 和 [8]那样生成新的分布或队列来消除依赖关系的影响。其次，在 [7] 中，所选分布可能会影响 SHAP结果。同样，在 [8] 中，所选相似度矩阵会影响新生成的队列。换言之，这两种方法都依赖用户，需要用户选择适合数据的分布或相似度矩阵。然而，最终用户可能不熟悉这类问题，因而可能选择错误的选项，影响结果的质量。所提方法不依赖用户，用户无需担心结果质量。第三，这两种所提方法提供的是局部解释，也就是说，它们可以解释模型对单个对象或一组对象的预测，但结果无法推广到所有实例。只要有一份信息性预测变量列表，所提方法既可用于局部解释，也可用于全局解释。在局部解释中，用户可以遵循第 2 节所示的相同步骤，生成一份新的信息性预测变量列表，并将共线性纳入考量。在所提方法中，我们采用了基于功能的评估。作为此类评估的代理指标，所提代理指标既无偏，也不会带来高昂成本，而基于应用的评估则并非如此 [12]。

在临床验证方面，相关性分析结果表明，与 SHAP 列表相比，所提方法得出的特征排序与依据生物学先验确定的参考（基准）排序更加一致。此外，Zahra Raisi-Estabragh 等人 [13]发表了所用四项特征——左心室质量、左心室收缩末期容积、左心室舒张末期容积和左心室射血分数——在男性与女性之间的平均差异。他们的结果表明，左心室质量的平均差异最大，其次是左心室舒张末期容积和左心室收缩末期容积，而左心室射血分数的差异非常小，因此排在列表末尾。这些结果与所提方法相符，排序完全一致。这证实，相较于 SHAP，所提方法与基准排序更加一致。

所提方法的主要局限在于它依赖于可解释人工智能模型。MIP 无法识别排名靠前的特征，而是依赖可解释人工智能方法（例如 SHAP）来突出这些特征，然后调整其顺序以考虑共线性。其次，如果使用的预测变量彼此独立，那么其结果将与其他可解释人工智能方法（或分类器）的结果没有区别。这一点既可以视为所提方法的弱点，也可以视为其优势。若将其视为弱点，那么在预测变量彼此独立时，

用户可能会放弃实施该方法。另一方面，即使预测变量彼此独立，也可以将其视为一种验证，并将其与可解释人工智能方法的结果进行比较。

6. 结论

总而言之，我们提出了一种简单但实用的度量方法，在将可解释人工智能方法应用于任何机器学习模型时，能够考虑共线性。该度量方法有助于用户更好地使用可解释人工智能方法，并生成更有意义的解释。它可应用于任何未考虑预测变量之间共线性的可解释人工智能方法。考虑到共线性，它既可用于单个对象的局部分析，也可用于全局分析，以展示特征对所有实例的边际贡献。所提方法针对现实应用中的一个主要问题进行处理，并给出了更稳健的结果。

7. 资助

AS获得英国心脏基金会项目资助（PG/21/10619）。IBG和GM感谢Fondazione CariVerona的支持（2018年卓越科学研究资助计划，EDIPO项目，参考编号2018.0855.2019）。ZRE感谢英国国家健康与护理研究院（NIHR）综合学术培训计划的支持，该计划支持其临床学术讲师职位；她还获得了英国心脏基金会临床研究培训奖学金（编号FS/17/81 /33318）的支持。SEP感谢英国国家健康与护理研究院（NIHR）Barts生物医学研究中心的支持，并获得了欧盟地平线2020研究与创新计划的资助，资助协议编号为825903（euCanSHare项目）。本文获得伦敦医学影像与人工智能价值医疗中心（AI4VBH）的支持。该中心的资金来自政府“工业战略挑战基金”中的“数据助力早期诊断与精准医疗”方向，由英国创新署代表英国研究与创新署（UKRI）负责管理和实施。文中观点仅代表作者本人，不一定代表AI4VBH联盟成员、英国国家医疗服务体系（NHS）、英国创新署或UKRI的观点。

8. 披露

SEP为位于加拿大阿尔伯塔省卡尔加里的Cardiovascular Imaging Inc提供咨询服务。其余作者均无需披露的利益关系。

参考文献

[1]David Gunning、Mark Stefik、Jaesik Choi、Timothy Miller、Simone Stumpf 和杨光忠。XAI——可解释人工智能。*Science Robotics*，4(37):eaay7120，2019年。[2]Liliana Szabo、Zahra Raisi-Estabragh、Ahmed Salih、Celeste Mc-Cracken、Esmeralda Ruiz Pujadas、Polyxeni Gkontra、Mate Kiss、Pal Maurovich-Horvath、Hajnalka Vago、Bela Merkely 等。心血管成像中可信且负责任的人工智能临床医生指南。*Frontiers in Cardiovascular Medicine，2022*年，第*9*卷，2022年。[3] Scott M Lundberg 和 Su-In Lee。解释模型预测的统一方法。*Advances in neural information processing systems*，30，2017年。[4] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。“为什么应该相信你？”解释任意分类器的预测。载于第*22*届*ACM SIGKDD*国际知识发现与数据挖掘会议论文集，第1135–1144页，2016年。[5] Jerome H Friedman。贪心函数逼近：梯度提升机器。*Annals of Statistics*，第1189–1232页，2001年。[6] Vaishak Belle 和 Ioannis Papantonis。可解释机器学习的原则与实践。*Frontiers in Big Data*，第39页，2021年。[7] Kjersti Aas、Martin Jullum 和 Anders Løland。当特征具有依赖关系时解释个体预测：更准确的 Shapley 值近似。*Artificial Intelligence*，298:103502，2021年。[8]Masayoshi Mase、Art B Owen 和 Benjamin Seiler。通过 Shapley 队列细化解释黑箱决策。*arXiv*预印本 *arXiv:1911.00467*，2019年。

[9]Ahmed Salih、Ilaria Boscolo Galazzo、Federica Cruciani、Lorenza Brusini 和 Petia Radeva。研究基于磁共振成像的痴呆分类中的可解释人工智能：可解释方法的新稳定性判据。载于*2022*年*IEEE*国际图像处理会议（*ICIP*），第4003–4007页。IEEE，2022年。

[10] Steffen E Petersen、Nay Aung、Mihir M Sanghvi、Filip Zemrak、Ken- neth Fung、Jose Miguel Paiva、Jane M Francis、Mohammed Y Khanji、Elena Lukaschuk、Aaron M Lee 等。基于英国生物样本库人群队列中高加索人的心血管磁共振（CMR）评估心脏结构与功能的参考范围。《心血管磁共振杂志》，19(1):1–19，2017年。

[11] Sarah R St Pierre、Mathias Peirlinck 和 Ellen Kuhl。生理性别很重要：女性与男性心脏的全面比较。《生理学前沿》，13:303，2022年。

[12] Jianlong Zhou、Amir H Gandomi、Fang Chen 和 Andreas Holzinger。评估机器学习解释的质量：方法与指标综述。《电子学》，10(5):593，2021年。[13] Zahra Raisi-Estabragh、Asmaa AM Kenawy、Nay Aung、Jackie Cooper、Patricia B Munroe、Nicholas C Harvey、Steffen E Petersen 和 Mo- hammed Y Khanji。左心室心脏磁共振正常参考范围的差异：系统综述与荟萃分析。《欧洲心脏杂志：心血管成像》，22(5):494–504，2021年。