arXiv:2602.10261v1 [cs.LG] 10 Feb 2026

基于核的胸部X光图像学习：预测COVID-19患者的ICU升级

Qiyuan Shi* Jian Kang† Yi Li‡ 密歇根大学生物统计学系，美国密歇根州安娜堡

摘要

核方法因能够捕捉复杂的非线性数据模式，已广泛应用于机器学习中的分类和预测任务。然而，单核方法具有固有局限，因为它们依赖单一类型的核函数（例如高斯核），可能不足以充分表征现实世界数据的异质性或多面性。多核学习（MKL）通过基于较简单的核构建复合核，并整合异质来源的信息，克服了这些局限。尽管取得了这些进展，传统的MKL方法主要针对连续结局而设计。我们将MKL扩展至适用于结局变量属于指数族的情形，从而涵盖更广泛的数据类型，并将所提出的方法称为集成多重加性核回归广义线性模型（GLIMARK）。通过实证研究，我们展示了GLIMARK能够有效恢复或近似真实的数据生成机制。我们将其应用于COVID-19胸部X光数据集，预测ICU升级这一二元结局，并提取具有临床意义的特征，凸显了该方法在真实场景中的实用价值。

关键词： 胸部X光片；特征提取；广义线性模型；多核学习。

*电子邮件：qsh@umich.edu

†电子邮件：jiankang@umich.edu

‡电子邮件：yili@umich.edu

1 引言

医学影像（如胸部X光片（CXR）和CT扫描）广泛用于疾病诊断、预后评估和结局预测。在COVID-19患者中，这类影像的应用尤为普遍 [Benmalek 等，2021，Cheng 等，2022]。然而，尽管一些研究表明，胸部X光评分和基线CT扫描均与 ICU 入院相关，但在预测患者病情恶化并转入重症监护病房（ICU）方面，对这些影像的研究仍较少。疫情期间，ICU面临严重的资源短缺，因此，及时且准确地做出升级护理决策，对于优化护理和有效分配资源至关重要 [Meille 等，2023年，Arabi 等，2022]。与CT相比，胸部X光片具有关键优势：检查速度更快、便于移动且更易获取，因此在这类高压情境下，是一种适用于实时决策的实用工具 [Jacobi 等，2020]。

Michigan Medicine 是美国中西部的一家大型医疗服务机构，维护着涵盖五百多万名患者的多模态数据集，包括电子健康记录数据和胸部X光影像数据。胸部X光影像数据集包含近两百万张图像，来自2017年至2023年间入住 Michigan Medicine 的178,071名患者。将胸部X光图像与电子健康记录数据集整合后，便可获取重症监护病房状态及其他患者级信息，从而开展更全面的分析。一个自然的问题是，我们能否以及如何利用胸部X光图像，以及患者入院时记录的其他临床数据，来预测重症监护病房的使用情况。研究结果有望为早期干预提供参考，并从长远来看改善患者预后。

对胸部X光图像进行量化有多种方法，例如评分系统和基于特征的方法。评分系统虽然相对简单，但由于存在多种框架，且各自采用不同的标准和解读方式，因此缺乏一致性 [Wasilewski 等人，2020]。基于特征的方法包括影像组学和深度学习技术，提供了一种更加数据驱动的替代方案。特别是影像组学特征，通常通过从医学影像中提取可量化且可重复的特征而手工设计；这种方法备受青睐，并已成功应用于许多 COVID-19 影像研究 [Sun 等人，2023]。

影像组学特征建模的一项关键准备工作是进行恰当的特征映射。这对于揭示复杂模式并捕捉数据中的非线性关系至关重要。传统的单核方法可能无法捕捉数据的全部复杂性，而多核学习 (MKL) [G¨onen 和 Alpaydın，2011]通过组合候选核来更好地表征多样化模式，提供了一种更灵活的方法。近期

MKL 领域的进展聚焦于多个方面，包括核优化、计算效率以及在不同领域中的应用。例如，Liu 等人 [2019]解决了缺失数据问题，Shen 等人 [2021] 实现了在线学习，而Han 等人 [2018] 则控制了模型复杂度。

然而，大多数 MKL 扩展主要关注平方损失以及支持向量机（SVM）等机器学习方法，而非概率模型，因此限制了其在面向多种数据类型的更广泛统计工具集中的适用性。MKL 的另一项挑战是降维（DR）。降维中的常见方法已整合到 MKL 中，例如 MKL-DR [Lin 等人， 2010] 但这些方法仍在很大程度上依赖投影或嵌入，从而损害了可解释性。已有研究开始关注对原始特征的解释 [Briscik 等人， 2023]。为解决可解释性问题，我们聚焦于多重加性回归核（MARK） [Bennett 等人， 2002], 其基于列生成（CG）[Demiriz 等人， 2002]。其一项关键优势在于，它不再寻求最优核，而是着眼于特征-核组合（即列），从而保留原始特征的可解释性。此外，选择过程主要涉及矩阵乘法；随着现代 GPU 的快速发展，这类计算任务能够得到高效处理。

尽管如此，原始 MARK 方法仍存在明显局限。首先，该方法仅依赖平方损失函数，因而隐含地假设结局为高斯型，这使其在二元结局情形下（如本研究）是否适用存在疑问。其次，尽管该方法的计算规模随样本量近乎线性增长，但面对大量特征-核组合时仍难以应对。这一局限尤为重要，因为与多核学习研究中通常使用的数据集不同 [Wolberg 等人， 1993, Harrison Jr 和 Rubinfeld， 1978], 影像组学通常涉及感兴趣区域（ROIs）——图像中特定的解剖或病理区域，例如肺叶、肿瘤或纤维化区域——而不同区域可能需要采用不同的表示方式。这可能带来沉重的计算负担。给定样本量为 N， P 个特征分区，以及 Q 种核，完整的核矩阵$\boldsymbol{K}$包含 NPQ 列。以规模为例，波士顿房价数据集有 $D = 13$ 个特征，最多可产生 $2^{13} = 8192$ 种可能的子集。然而，这些组合中的大多数并不代表有意义的分组，因此对结构化建模的用处有限。相比之下，我们的影像组学数据集有 $D = 2046$，由此产生数量惊人的可能分区（2^{2046}），其中许多对应解剖学或临床相关分组，因此更适合用于核分配。

为解决分析我们数据集中所呈现的胸部X光图像时遇到的这些局限

我们提出了一种改进的 MARK，称为集成多重加性核回归广义线性模型（GLI-MARK），在以下方面有所改进：1）我们在广义线性模型（GLM）框架中引入基于似然的损失函数，以适应多种结果，包括二元、泊松和高斯结果；2）我们提出了一个用于建模胸部 X 线影像组学特征的层级框架。所提出的 GLIMARK 能够在多个尺度上探索特征模式，兼顾模型复杂度与模式学习能力。此外，采用 $\mathcal{H}_k$-范数进行修改，可以依据表示定理给出更严谨的解释。我们的分析聚焦于 ICU 升级这一二元结果。我们应用该方法分析数据，发现纳入多核能够进一步发现重要的影像组学特征。具体而言，邻域灰度差异矩阵（NGTDM）和二维形状特征被确定为排名靠前的特征类别，而传统回归模型和随机森林（RF）模型均未能检测到这些特征。此外，“表示定理”患者突显了可供进一步放射学分析的关键视觉特征，并可作为预测未来患者结局的基准。

第2节回顾多核学习（MKL）中的相关工作。第3节介绍我们提出的GLIMARK。第4节开展模拟研究，以说明如果结局由某些真实函数形式生成，GLIMARK能够恢复或近似真实情况。第5节将该方法应用于来自Michigan Medicine医院系统的真实胸部X线数据。第6节讨论我们方法的优势与局限，以及其在侧位胸部X光片和视觉变换器中的潜力。

2 相关工作与研究动机

2.1 非高斯结局数据下的表示定理结果

给定由 N 个相互独立个体组成的训练样本，对于 $i = 1$、.....、N，令 $\boldsymbol{z}_i$ 表示预测变量（例如胸部X线影像数据），y_i 表示结果（例如住院患者是否病情加重并转入重症监护病房）。因此，给定 $\boldsymbol{z}_i$ 时，y_i 的条件分布可由再生核希尔伯特空间（RKHS）$\mathcal{H}_k$ 中诱导出的函数参数 $f(\cdot)$ 指定，其形式如下：

$$f(\cdot) = \sum_{j=1}^{N} \alpha_j k(\boldsymbol{z}_j, \cdot) + b,$$

其中， $k(\cdot,\ \cdot)$ 是与 $\mathcal{H}_k$ 相关联的再生核。该表示由表示定理 [（Wahba 和 Wang） 得出， 2019], 该定理指出，对于再生核希尔伯特空间中的许多正则化学习问题，包括我们这里的（2），其解都可以表示为训练点处核函数取值的有限线性组合。这类表示广泛应用于多种场景，包括核岭回归（KRR） [（Lin 等）、 2020],核 Cox 模型 [（Rong 等）以及深度学习 [（Unser）， 2019]。

2.2 核构造

多核学习的一个主要目标是从预定义的核中构造一个正定核。构造方式可以是线性的、非线性的，甚至依赖于数据 [G¨onen 和 Alpaydın， 2011]。例如，两个有效核的和仍然是有效核，这促使我们使用锥组合来构造新的核函数。Lanckriet 等人 [2004]通过组合线性核、多项式核和径向基函数（RBF）核，成功提升了支持向量机的性能，其决策函数为： $f(\boldsymbol{z})$ $= \sum_{j=1}^{N} \alpha_j k(\boldsymbol{z}_j,\ \boldsymbol{z}) + b,$

其中 $k(\boldsymbol{z}_j,\ \boldsymbol{z}) = \sum_{q=1}^{Q} \mu_q k_q(\boldsymbol{z}_j,\ \boldsymbol{z})$。这里， $\boldsymbol{z}_j \in \mathbb{R}^D$ 为训练点， $\boldsymbol{z} \in \mathbb{R}^D$ 是用于进行预测的任意点， Q 是核的数量 f 核，以及 $\mu_q \geq 0,\ q = 1,\ \ldots\ldots,\ Q$ 是候选核的权重。

这种组合核的构造能够有效处理各种数据源。假设预测变量由 P 个组成部分或分区构成，每个部分或分区都能捕捉数据的一个不同方面。例如，在胸部 X 线影像组学中，特征提取自两个感兴趣区域，每个区域均经过十一种滤波器处理，并归入七个特征类别。由此共得到 2046 个特征，这些特征根据感兴趣区域、滤波器和特征类别的组合自然划分为 154 个分区： $P = 2 \times 11 \times 7 = 154$。

通过从数据中学习核权重 μ_p ，我们可以构建灵活的数据驱动核表示： $k(\boldsymbol{z}_j, \boldsymbol{z}) = \sum_{p=1}^{P} \mu_p k_{[p]}(\boldsymbol{z}_j^p, \boldsymbol{z}^p)$，其中 $\boldsymbol{z} = \{\boldsymbol{z}^p\}_{p=1}^{P}$ 和 $\boldsymbol{z}_j = \{\boldsymbol{z}_j^p\}_{p=1}^{P}$。当 P 较大时，与其构建 P 个新核（这可能不切实际），不如在所有分区中使用一组通用的 Q 个核，这是一种更有效的方法： $k(\boldsymbol{z}_j, \boldsymbol{z}) = \sum_{q=1}^{Q} \sum_{p=1}^{P} \mu_{p,q} k_q(\boldsymbol{z}_j^p, \boldsymbol{z}^p)$。这种允许核适应异构数据源的表述与多视图学习相一致 [Yan 等人， 2021]。在图像分析中，分区可能对应于不同的感兴趣区域（例如肺部与周边区域）、不同的方面（如纹理和形状）、不同的成像视角（例如胸部X光片的后前位（PA）与侧位（LAT）），甚至不同的成像模态（例如胸部X光片与计算机断层扫描）。受多视图学习范式的启发，我们将核-分区组合称为一个“视图”。

（图 1）。

3 基于核的广义线性模型与多重加性回归（GLI-MARK）

3.1 方法

为处理我们数据示例中遇到的非高斯结局数据（例如二元 ICU 使用），我们将 MKL 扩展到广义线性模型（GLM）框架 [Bennett 等人，2002，Zhu 和 Hastie，2005]。假设结局 y_i 服从指数族中的某个分布，其概率密度函数为：

$$p(y|\theta) = \exp\left(y\theta - b_{glm}(\theta) + c_{glm}(y)\right), \tag{1}$$

为简化起见，我们省略离散参数，与 Fei 和 Li [2021],的处理相同，因为我们关注的是均值参数。除仅取决于观测数据的常数项外，单个观测的对数似然为：$\ell(\theta_i|y_i) = y_i\theta_i - b_{glm}(\theta_i) \quad i = 1, ..., N$，其中 $b_{glm}(\cdot)$ 是一个将 y_i 的均值与 θ_i相联系的函数。在广义线性模型的典范连接函数下，决策函数 f 通过以下关系与结果的期望相联系：

$$\theta_i = f_i = g_{glm}(\mu_i), \quad \text{where } f_i = f(\boldsymbol{z_i}) \text{ and } \mu_i = \mathbb{E}[Y_i|\boldsymbol{z_i}].$$

下文中，我们交替使用 θ_i 和 f_i。因此，我们提出采用以下正则化经验风险函数：

$$H = -\frac{1}{N}\sum_{i=1}^{N}\ell(f_i|y_i) + \lambda P(\|f\|_{\mathcal{H}_k}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i f_i - b_{glm}(f_i)] + \lambda P(\|f\|_{\mathcal{H}_k}), \tag{2}$$

其中，$\lambda > 0$ 是用于控制模型拟合与复杂度之间权衡的正则化参数，$P(\cdot)$ 是单调递增函数，而H 是需要最小化的目标函数。这里，f 的再生核希尔伯特空间范数记为 $\|f\|_{\mathcal{H}_k}$。当训练样本量为 N时，多核学习的一种自然扩展是构造如下决策函数：

$$f(\boldsymbol{z}) = \sum_{j=1}^{N}\alpha_j\sum_{q=1}^{Q}\sum_{p=1}^{P}\mu_{p,q}k_q(\boldsymbol{z}_j^p, \boldsymbol{z}^p) + b. \tag{3}$$

但这会带来若干问题。首先，为了解决可识别性问题，需要增加额外约束，例如 $\sum_{p=1}^{P}\sum_{q=1}^{Q}\mu_{p,q}=1$，然而，这可能会使优化变得复杂，尤其是在再生核希尔伯特空间范数$\|f\|_{\mathcal{H}_k}$同时取决于 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\mu}$时。此外，核方法本质上基于相似性，这意味着要对一个新点$\boldsymbol{z}$ 进行预测，需要计算 $\boldsymbol{z}$ 与所有训练点 $\boldsymbol{z}_j$ 之间的相似度，针对$j=1$、.....、 N个 PQ 视图逐一计算。计算开销很大，尤其是在涉及多核时。核多重加性回归（MARK） [Bennett 等人提出的 2002] 方法通过取消对系数的所有约束，并对 $\boldsymbol{\alpha}$进行重新参数化，有效地将多核学习问题简化为标准回归模型：

$$f(\boldsymbol{z})=\sum_{j=1}^{N}\sum_{q=1}^{Q}\sum_{p=1}^{P}\alpha_{j,p,q}k_q(\boldsymbol{z}_j^p,\boldsymbol{z}^p)+b. \tag{4}$$

为处理过参数化问题，MARK 仅选择有限组$\alpha_{j,p,q}$或其对应的患者-视图组合。

由于可用于非高斯结局数据的基于 MARK 的方法很少，我们将其扩展以适用于属于指数族的结局数据（1）。遵循 MARK 和列生成（CG）的思想 [L¨ubbecke 和 Desrosiers, 2005], 我们设计了一种算法，用于选择由设计矩阵的列表示的患者-视图组合。具体而言，我们构造了一个对应于模型（4）的设计矩阵，该矩阵连接了所有 P 分区和 Q核的核矩阵：

$$\boldsymbol{K}=\begin{bmatrix}\boldsymbol{K}_{1,1} & \dots & \boldsymbol{K}_{1,Q} & \boldsymbol{K}_{2,1} & \dots & \boldsymbol{K}_{2,Q} & \dots & \boldsymbol{K}_{P,1} & \dots & \boldsymbol{K}_{P,Q}\end{bmatrix}, \tag{5}$$

其中，每个块 $\boldsymbol{K}_{p,q}$ 是一个 $N\times N$ 矩阵，且 $\boldsymbol{K}_{p,q}$ 的每个元素由 $k_q(\boldsymbol{z_i}^p,\boldsymbol{z_j}^p)$ $i,j=1,.....,N$给出，其中 $k_q(\cdot,\cdot)$ 的定义见第 2.2节。因此，（5）中的每一列都是一个患者-视图组合。随后，我们将模型（4）（对于任意 $\boldsymbol{z}_i\in\mathcal{X}$）表示为：

$$f_i=f(\boldsymbol{z}_i)=\sum_{j=1}^{J}\alpha_j[\boldsymbol{K}]_{i,j}+b, \tag{6}$$

其中， $J=NPQ$和$[\boldsymbol{K}]_{i,j}$是 $\boldsymbol{K}$ 的第 (i,j) 个条目。

我们将该算法称为集成多重加性核回归广义线性模型（GLIMARK），下文将对此进行详细介绍。GLIMARK 通过重复以下过程，迭代构建稀疏预测模型

：在每次迭代中进行选择和更新：选取 $\boldsymbol{\alpha} \in \mathbb{R}^J$ 的一个分量（或等价地，选取 $\boldsymbol{K}$中相应的一列），随后更新所有已选分量的估计值，以降低式（2）的值，直到达到预先设定的步数，或改进幅度小于某个阈值，例如 $tol = 10^{-6}$。这里，$\boldsymbol{\alpha}^{(t)} \in \mathbb{R}^J$ 和 $b^{(t)}$ 分别表示在第 $\boldsymbol{\alpha}$ 次迭代时 b 和 t 的估计值。这个过程本质上是一种前向选择方法：它从零模型 $\boldsymbol{\alpha}^{(0)} = \mathbf{0}$开始，在每次迭代中，根据剩余候选变量中$\alpha_j$ 对式（2）的影响，依次添加一列：

$$\begin{aligned}\frac{\partial H(\boldsymbol{\alpha}, b)}{\partial \alpha_j} = &- \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \frac{\partial b_{glm}(f_i(\boldsymbol{\alpha}, b))}{\partial f_i}\right][\boldsymbol{K}]_{i,j} \\ &+ \lambda \frac{\partial P(\|f(\boldsymbol{\alpha}, b)\|_{\mathcal{H}_k})}{\partial \alpha_j}\end{aligned} \tag{7}$$

所有偏导数均在 $(\boldsymbol{\alpha},\ b) = (\boldsymbol{\alpha}^{(t)},\ b^{(t)})$ 处求值。

选择结果用索引集 $S^{(t)}$ 表示，其中 $S^{(t)} = (j^{*(1)},\ j^{*(2)} \ldots,\ j^{*(t)})$ 且 $S^{(0)} = \emptyset$。每一步，我们从 $\boldsymbol{\alpha}$ 中选择一个尚未选取的分量，其索引为 $j^{*(t+1)}$，并更新 $S^{(t+1)}$，其中

$$j^{*(t+1)} = \arg\max_{j: j \notin S^{(t)}} \left|\frac{\partial H(\boldsymbol{\alpha}, b)}{\partial \alpha_j}\right| \tag{8}$$

偏导数的求值结果为 $(\boldsymbol{\alpha},\ b) = (\boldsymbol{\alpha}^{(t)},\ b^{(t)})$。本质上，偏导数的绝对值是前向选择的判据。

估计通过基于梯度下降的方法实现，例如 Adam [Kingma, 2014] 这些方法通过最小化式（2）实现：

$$f_i = \sum_{j \in S^{(t+1)}} \alpha_j [\boldsymbol{K}]_{i,j} + b. \tag{9}$$

因此，对于所有 $j \in S^{(t+1)}$，均使用 Adam 更新 $\alpha_j^{(t+1)}$ 和 $b^{(t+1)}$，并且 $\alpha_j^{(t+1)} = 0, \quad \forall j \notin S^{(t+1)}$。

所提出的 GLIMARK 已通过算法1进行总结，并在网络附录A中进行了详细说明。该算法支持多种响应数据类型。例如，当选择 $P(\cdot)$平方函数时，算法中的目标函数 H 可针对二项分布和泊松分布指定为：

$$H(\boldsymbol{\alpha}, b) = \begin{cases} -\frac{1}{N}\sum_{i=1}^{N}\left[y_i f_i - \log(1+\exp(f_i))\right] + \lambda\|f\|_{\mathcal{H}_k}^2, & \text{(Binomial)} \\ -\frac{1}{N}\sum_{i=1}^{N}\left[y_i f_i - \exp(f_i)\right] + \lambda\|f\|_{\mathcal{H}_k}^2, & \text{(Poisson)} \end{cases}$$

算法1GLIMARK

Let $b^{(0)} = g_{glm}(p_+)$, $p_+ = \frac{1}{N}\sum_{i=1}^{N} y_i$, $tol = 10^{-6}$
Let $S^{(0)} := \emptyset$
Let $\boldsymbol{\alpha}^{(0)} = \mathbf{0}$
for $t := 0$ to T_{max} **do**
    Select $j^{*(t+1)}$ according to (8)
    if $\left| \frac{\partial H}{\partial \alpha_{j^{*(t+1)}}} \right|_{(\boldsymbol{\alpha}, b) = (\boldsymbol{\alpha}^{(t)},\, b^{(t)})} \leq tol$ **then**
        return $\boldsymbol{\alpha}^{(t)}, b^{(t)}$ as the final estimates, i.e., $\hat{\boldsymbol{\alpha}}$ and $\hat{b}$
    end if
    Update $S^{(t+1)}$ as $S^{(t+1)} = S^{(t)} \cup \{j^{*(t+1)}\}$
    Minimize Eq. (2) with (9) using Adam, give updates for $\alpha_j^{(t+1)}, \quad \forall j \in S^{(t+1)}$ and $b^{(t+1)}$. Let $\alpha_j^{(t+1)} = 0, \quad \forall j \notin S^{(t+1)}$
end for
return $\boldsymbol{\alpha}^{(t+1)}, b^{(t+1)}$ as the final estimates, i.e., $\hat{\boldsymbol{\alpha}}$ and $\hat{b}$

其梯度为:

$$\frac{\partial H(\boldsymbol{\alpha}, b)}{\partial \alpha_j} = \begin{cases} -\frac{1}{N}\sum_{i=1}^{N}\left[y_i - \frac{\exp(f_i)}{1+\exp(f_i)}\right][\boldsymbol{K}]_{i,j} + \lambda \frac{\partial \|f\|^2_{\mathcal{H}_k}}{\partial \alpha_j}, & \text{(Binomial)} \\ -\frac{1}{N}\sum_{i=1}^{N}\left[y_i - \exp(f_i)\right][\boldsymbol{K}]_{i,j} + \lambda \frac{\partial \|f\|^2_{\mathcal{H}_k}}{\partial \alpha_j}. & \text{(Poisson)} \end{cases}$$

3.2 层次结构、解读与实现

支持向量机方法将决策函数视为输入与支持向量之间相似度度量的加权和，这与侧重于解释方差的传统回归方法形成对比。表示定理通过表明，最优决策函数虽定义在无限多个变量上，却可以表示为“表示元”的线性组合，从而衔接了这两种方法。受此启发，多核学习在多个特征空间中度量相似度，每个特征空间由一个视图定义，即核与特征子集的组合，用以捕捉特定视角（图 1）。在 GLIMARK 中，决策函数成为新观测值与选定患者—视图组合之间相似度的加权和，这些组合对应于拼接矩阵 $\boldsymbol{K}$的列。视图呈现层次结构，对单个属性或组应用核，类似于 MARK 框架中的弱学习器和强学习器。

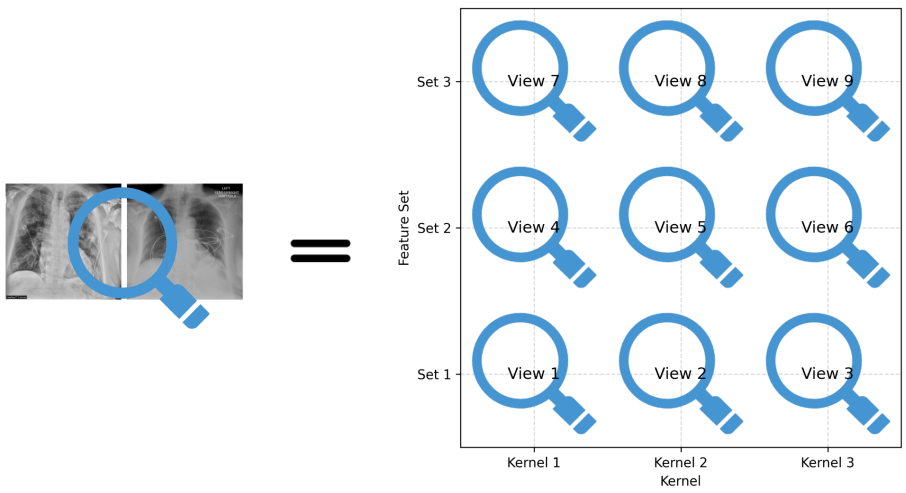
由于核在不同尺度上运作，且正则化通过 $\|\boldsymbol{\alpha}\|_2$或 $\|f\|_{\mathcal{H}_k}$实现，因此双层 GLIMARK 类似于混合组稀疏套索方法，例如 [Friedman 等人的研究 2010]：最小化$_{\boldsymbol{\alpha}} - \frac{1}{N}\sum_{i=1}^{N} \ell(f_i|y_i) + \lambda_1 \sum_{\ell=1}^{L} \|\boldsymbol{\alpha}_\ell\|_2 + \lambda_2 \|\boldsymbol{\alpha}\|_1$. 该框架促使不同尺度的特征相互竞争，使模型能够自动选择合适的尺度。与传统方法不同，双层 GLIMARK 对每个变量使用两次：一次在组层级，一次单独使用，这可能会降低某些尺度下的系数幅度。

在诸如核支持向量机分类这样的简单情况下，$\boldsymbol{\alpha}$ (对偶)与特征权重 $\boldsymbol{w}$ (原始)之间的关系由 [Bishop 2006]给出： $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \phi(\boldsymbol{z}_i)$，其中 $f(\boldsymbol{z}) = \boldsymbol{w} \cdot \phi(\boldsymbol{z}) + b$ 位于原始问题中。特征映射 $\phi(\cdot)$可以从核理论上或通过经验获得 [Scholkopfetal.、1999、Minh et al.、 2006] 。对于 GLIMARK，这些关系并不那么直接。为了评估变量检测性能，我们提出按组灵活地对 $|\hat{\alpha}_j|$ 求和，得到 I_{Group}。例如，$I_{\text{First-Order L RBF}}$ 是在所有个体上进行边际汇总，展示了肺部区域中来自一阶特征和 RBF 核的视图总体重要性，而 $I_{\text{First-Order}}$ 表示一阶特征的总体重要性。或者， $I_{\text{PatientID}}$ 对每个个体在各视图上求和，并识别出重要的表示元。较高的值表示跨特征空间的广泛比较，或某一空间中的显著权重，从而突出显示可供进一步分析的关键训练点。

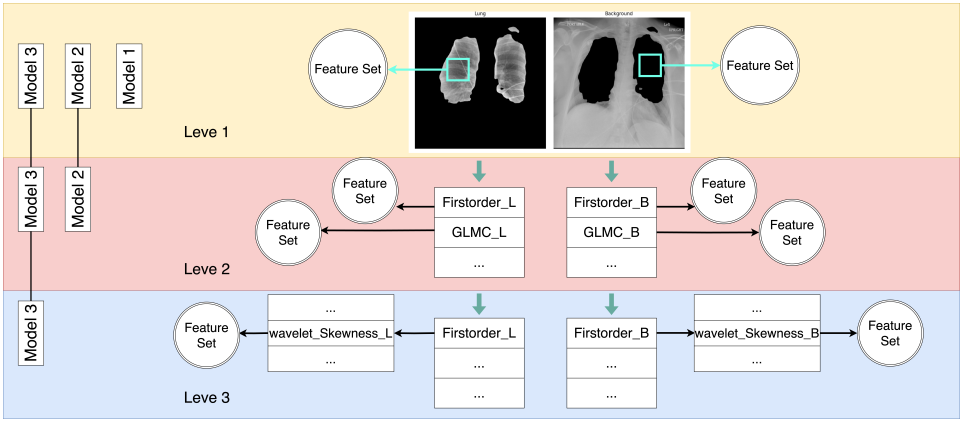
对于分类性能，GLIMARK 可以预测二元结果的概率，从而生成 ROC 曲线。我们使用约登指数 [Fluss 等人， 2005], 定义为敏感度加特异度再减去 1，以确定分类的最佳阈值，不过也可以采用 0.5 作为截断值。在层级数据的交叉验证中，需要调整两个超参数： λ，用于正则化模型复杂度；以及 T，用于确定达到最佳性能所需的项数。基于约登指数的准确率、0.5 截断值、曲线下面积或 F 分数均可用于指导这些参数的微调。

4 模拟研究

这项模拟研究有两个主要目的。第一个目的是确定其检测结果与预测变量之间潜在线性和非线性关系以及可能存在的交互作用的能力。第二个目的是评估该模型相较于其他常见方法（如支持向量机和随机森林）的分类性能，并探究参数 λ 和 T 的最佳组合，这两个参数均用于控制



(a) 两幅图像的多视图比较。



(b) 层次模型结构。

图1

表 1：模拟设置。变换包括逐元素立方（$z_{i,g}^3$）、平方（$z_{i,g}^2$）、指数（$z_{i,g}^{exp}$）和个体变量项（$z_{i,g1}$，表示第1组中的第一个变量）。架构表示核是应用于组（G）、个体（I）还是两者（G + I）。

情景	非线性关系	架构
A	$f(\boldsymbol{z}_i) = \sum_{g=1}^{6} \langle \boldsymbol{z}_{i,g}, \boldsymbol{\beta}_g \rangle + 1$	G
B	$f(\boldsymbol{z}_i) = \langle \boldsymbol{z}_{i,1}^3, \boldsymbol{\beta}_1 \rangle + \langle \boldsymbol{z}_{i,2}^2, \boldsymbol{\beta}_2 \rangle + \langle \boldsymbol{z}_{i,1}, \boldsymbol{\beta}_1 \rangle + 1$	G
C	$f(\boldsymbol{z}_i) = 3 \sum_{g=1}^{6} \langle \boldsymbol{z}_{i,g}^{exp}, \boldsymbol{\beta}_g \rangle + \langle \boldsymbol{z}_{i,4}, \boldsymbol{\beta}_4 \rangle + 1$	G
D	$f(\boldsymbol{z}_i) = \langle \boldsymbol{z}_{i,1}, \boldsymbol{\beta}_1 \rangle + \langle \boldsymbol{z}_{i,2}, \boldsymbol{\beta}_2 \rangle + \sum_{j<k} \gamma_{jk} z_{i,1j} z_{i,1k} + 1$	G
E	$f(\boldsymbol{z}_i) = \langle \boldsymbol{z}_{i,1}, \boldsymbol{\beta}_1 \rangle + \sum_{j<k} \gamma_{jk} z_{i,1j} z_{i,1k} + 1$	G + I
F	$f(\boldsymbol{z}_i) = 3\,\text{正弦}\left(2\sum_{g=1}^{10} \lvert z_{i,g1} \rvert + 1\right) + \exp\left(0.0001 \sum_{g=1}^{10} z_{i,g2} - 2\right) + 100 \langle \boldsymbol{z}_{i,3}^2 \boldsymbol{\beta}_3 \rangle + \langle \boldsymbol{z}_{i,4}, \boldsymbol{\beta}_4 \rangle + 1$	G + I
G	$f(\boldsymbol{z}_i) = \frac{1}{20}[10\,\text{正弦}\left(50 \sum_{g=1}^{10} \lvert z_{i,g1} \rvert + 1\right) + 25\text{指数函数}\left(0.1 \sum_{g=1}^{10} z_{i,g2} - 2\right) + 3 \langle \boldsymbol{z}_{i,3}^2, \boldsymbol{\beta}_3 \rangle + 5 \langle \boldsymbol{z}_{i,4}, \boldsymbol{\beta}_4 \rangle + 1]$	G + I

模型的复杂度。

对于我们的首个目标，我们采用了与 Friedman 等人 [2010], 类似的设置，其中包含 10 组变量（G1 至 G10），每组包含 10 个由标准高斯分布生成的变量。组内及组间均不存在相关性。G1 至 G6 中的有效变量数量分别为（10、8、6、4、2、1），而 G7 至 G10 中的变量均为无关变量。对于观测 $\boldsymbol{z}_i$，$\boldsymbol{z}_{i,\ g}$ 表示组 g中的有效变量，$\boldsymbol{\beta_g}$ 表示其对应的系数，系数从 ± 1中随机选取。情景（表 1）包括：模型 A、B 和 C 涉及线性或非线性关系，模型 D 和 E 则考察交互效应。

对于每种情景，我们进行 100 次独立模拟，样本量为 400。实施 GLIMARK 时，我们固定 $\lambda = 0.001$ 和 $T = 400$。结果见图 2。

情景 A（线性模型）： 线性 GLIMARK 能正确识别有效组，其 I_{Group} 贡献与有效变量的数量成正比（G1-G6：19% 至 7%）。非有效组中出现少量噪声（5%）。添加 RBF 核不会改变检测结果，只会引入

噪声（例如，G1-RBF 中为 3%）。

情景 B（三次项和平方项）： 线性 GLIMARK 无法捕捉 G2，将其 I_{Group} 估计为 7%，与噪声相近。加入二次多项式核后，检测效果有所改善，G2 的贡献增加到 9%，这表明多项式核有助于捕捉相关的非线性结构。

情景 C（线性项和指数项）： 线性 GLI- MARK 仅检测到 G4 的线性效应，却无法捕捉 G1-G6 的指数贡献。它还引入了 9% 的噪声。额外加入的 RBF 核成功识别出前两组中的指数效应，但未能识别 G4 中的线性效应。这凸显了估计值对核函数选择和真实效应大小的敏感性。

情景 D（G1 中的成对交互作用）： 真实模型包含 G1 内部的成对交互作用，但可能的组合中只有 50% 处于激活状态（$\gamma_{jk} \in \{1.5, 0, -1.5\}$）。线性 GLIMARK 对 G1 和 G2 的贡献估计相近（19% 对 18%），无法将交互作用与主效应区分开来。混合 GLIMARK（线性 $+$ 二次多项式）能更好地分解这些效应，将 20% 归于 G1-多项式，将 6% 归于 G2-线性。这与多项式核的特征映射相符，该映射无需显式指定即可捕捉双向交互作用。

情景 E（简化交互情形）： 我们仅考虑 G1，其中包含五个有效变量。共有 10 个可能的双向交互作用，其中 70% 被随机选为活跃交互作用。交互作用系数γ_{jk} 被赋予 $\{1.5, 0, -1.5\}$中的值，其中非零值对应于活跃交互作用。我们测试了两个混合 GLIMARK 模型：一个对所有交互作用使用二次多项式核、对个体变量使用线性核；另一个将多项式核替换为 RBF 核。两个模型均成功检测出活跃交互作用。

作为我们第一个目标的总结，当变量与其背后的函数形式密切相关时，多核 GLIMARK 模型能够突出显示这些变量，并将其归因于正确的核。但该模型也会受到不确定性的影响，并且对设定错误较为敏感。

针对我们的第二个目标，我们采用了与第一个目标相同的变量设置，对应于情景 F 和 G（表1）。该设置模拟了同时包含组效应和个体效应的复杂非线性函数形式。我们对 λ 和 T 进行了网格搜索。对于每个 λ 值，我们模拟了 100 对训练-测试数据集对，并运行了一个三核 GLIMARK，同时改变T 值。所使用的核包括线性核、二次多项式核和 RBF 核，其中西格玛值为 10。

对于二元数据，我们根据准确率评估模型性能

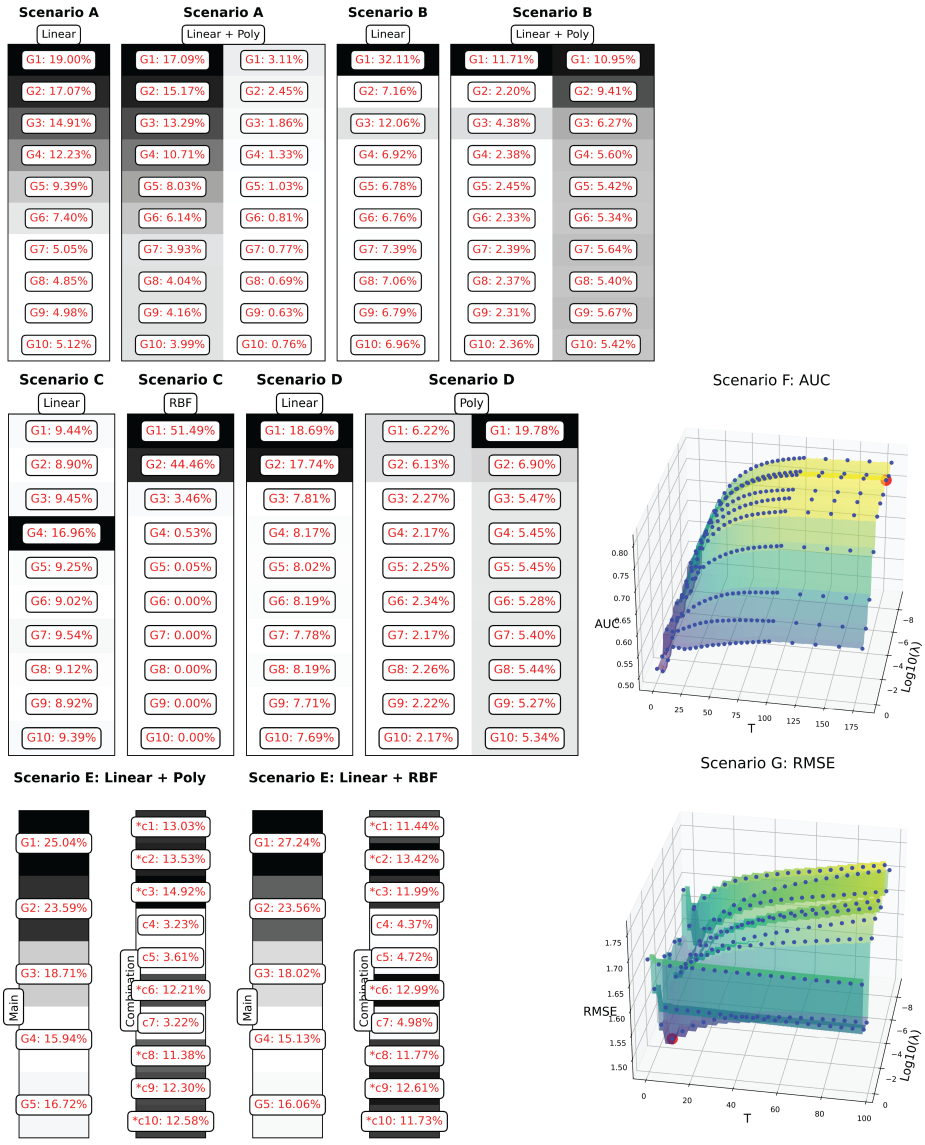


图 2：模拟结果。红色文本表示每个视图的贡献百分比。G# 表示一个组，而C#指一种可能的组合。星号表示真实模型中真正活跃的组合。情景 A–E 表示 GLI-MARK 的特征检测性能，而情景 F–G 表示其随两个超参数 λ 和 T 变化时的整体性能。

基于最优 Youden's J 统计量和曲线下面积（曲线下面积）。对于计数数据，我们使用均方根误差（均方根误差）。作为比较，我们还在相同的数据集上测试了线性支持向量机、径向基函数核支持向量机，以及包含 100 棵树的随机森林模型。

场景F（二元数据）： 总体而言，GLIMARK 的曲线下面积随着 T 增大而增大，随着 λ 增大而减小。最佳曲线下面积出现在 $T = 180$ 和 $\lambda = 10^{-7}$。这表明，在此情形下，二元数据受益于相对较弱的正则化和更多的表示元。使用最优截断值时，相应的准确率为 0.78。相比之下，线性支持向量机模型、径向基函数支持向量机模型和随机森林模型的准确率分别为 0.65、0.75 和 0.76。

场景G（计数数据）： 我们还在场景G中模拟了泊松数据，并发现了不同的模式。当 $T = 6$ 时，均方根误差达到最小值；超过这一点继续增加表示元会导致性能下降。同样，随着 λ 减小，均方根误差会增大，这表明与二元数据不同，必须采用更强的正则化来防止过拟合。GLIMARK 的最低均方根误差为 1.48，而随机森林模型为 1.53。

5 应用于PrecisionHealthCOVID-19CXR数据

5.1 患者群体与放射组学特征

本研究纳入的患者来自Precision Health数据库，并符合以下标准：(1) 确诊或疑似COVID-19；(2) 在Michigan Medicine至少接受过一次AP或后前位胸部X光片检查；(3) 就诊时间介于2020年3月至2022年3月之间。二元结局定义为：患者在首次COVID-19确诊后30天内是否被转入Michigan Medicine的任一重症监护病房。纳入的2,396名患者中，54%为男性。种族构成为：69%为白人，20%为非裔美国人，11%来自其他种族群体。年龄中位数为60岁（IQR：44–77）。共有756名患者（31.5%）需要ICU升级。详细的患者人口统计学信息和合并症信息见补充材料。对于每位患者，仅考虑首次COVID-19确诊日期当天获得的胸部X光片；如果当天有多张胸部X光片可用，则仅纳入其中一张进行分析。

图像采用医学数字成像与通信（DI-COM）格式。每幅图像平均包含约 2757×3195个像素。像素归一化会将原始数值映射到 0 到 255 的范围内，并且

直方图均衡化可增强图像对比度。分割使用常用的网络架构 U-Net 完成 [Ronneberger 等人， 2015]。借助真实标注的胸部X光图像及相应掩膜，U-Net 可以实现良好的分割。分割掩膜是一种二值图像，其中每个像素都对应原始图像中的相同位置，并被标记以指示其属于肺部区域还是背景；两者均视为感兴趣区域的一部分。这些掩膜通常由临床医生生成。我们使用了两个公开的带肺部掩膜的胸部X光数据集：蒙哥马利县（MC）数据集，以及印第安纳大学数据集中包含 55 幅图像且带有人工标注的子集 [Candemir 等人，2013，Xue 等人， 2023]。网络图 3 展示了图像处理流程。

我们使用 PyRadiomics 软件包 [Van Griethuysen 等人， 2017] 提取特征。工作流程包括一个滤波步骤，随后是特征提取步骤。首先，我们对每张图像应用十一种滤波器之一以增强图像。然后，根据固定公式，提取属于七个类别之一的特征，这些特征对应肺分割或背景分割。网络表格 3 列出了所有滤波器和类别。例如，一阶统计量类别包含十九个特征，灰度游程长度矩阵类别包含十六个特征。结合所有滤波器和类别，每张图像针对每种分割可产生 1032 个特征，总计 2064 个特征。网络图 1–2 展示了总结影像组学特征与 ICU 升级之间单变量关联的探索性图表。对于 GLIMARK 模型，我们将层次结构呈现为一个三级框架。第一级包含基于分割的所有特征。第二级包含基于类别的所有特征。第三级包含个体特征（图 1）。

5.2 模型层级与核选择

为确定模型的层次结构，我们按照图1所示，在嵌套模型阶段中进行逐步探索性分析。首先，我们固定 $\lambda = 0.001$ 和 $T = 400$，与模拟研究中的做法相同。接着，在分割层级运行单层GLIMARK模型，采用偏移量为 1 的二次多项式核，以及一系列 σ 取值范围较广的 RBF 核。随后，我们为 RBF 核选择最优的 σ ，并将其与多项式核组合，最终确定为模型 1。在此基础上，我们纳入特征组层级数据，并提出具有不同 σ 取值的新 RBF 核，再次与多项式核组合，随后选取最佳 RBF 核并最终确定为模型 2。最后，我们扩展模型 2 中排名靠前的组级特征，并将这些组中的个体特征作为第三级数据纳入，同时采用相同的多项式

核和一个具有较小 σ 的 RBF 核（模型 3）。这样做的目的是在允许模型探索不同层级模式的同时，控制模型在 P 和 Q 方面的复杂度。此特征探索步骤应用于整个数据集。确定层次结构后，我们对全部三个模型进行五折交叉验证，以评估并报告平均性能。

5.3 结果与结论

5.3.1 探索性分析结果

嵌套探索性分析的详细结果见网络附录 B。简而言之，模型 1 显示，径向基函数核和多项式核在两个片段中均有显著贡献。到了模型 2，特征类别项占据主导地位，大幅降低了分割层级的贡献。模型 3 在单个特征层级纳入了四个特征类别，并将其视为最终模型进行交叉验证。

5.3.2 模型比较与特征重要性

我们比较了 GLIMARK 模型与随机森林及采用 ℓ_1 和 ℓ_2 正则化的逻辑回归模型在五折交叉验证微调中的表现。随机森林和逻辑回归均使用了所有特征。

图3显示，随着 T 的增加，三个 GLIMARK 模型的表现均有所提升。当 GLIMARK 的表示元约为 200 个时，模型 3 的曲线下面积开始高于随机森林。模型 3 的曲线下面积最高约为 0.83。GLIMARK 模型的约登 J 指数和 F 分数通常也高于随机森林。值得注意的是，在所有指标上，项数少于 400 的 GLIMARK 模型表现与使用全部 2046 个预测变量进行微调的逻辑回归模型相当。这表明 GLIMARK 能有效平衡简约性与预测能力。

图4 展示了不同模型中特征重要性的分布。模型 3 的表现与采用 Lasso 正则化的逻辑回归相近，均突出了较少但更重要的特征类别。GLI-MARK 进一步聚焦于三个主要类别：一阶、邻域灰度差异矩阵和二维形状。这些发现与现有文献一致。研究表明，一阶影像组学特征能够预测 COVID-19 患者的 ICU 入院和气管插管情况 [Varghese 等人， 2021], 这与相关研究结果一致：肺部区域像素强度升高与 COVID-19 相关的肺部受累相对应 [Al-Zyoud 等人， 2023]。二维形状 L 特征用于测量肺部区域的形状，

研究表明，重症 COVID-19 患者的肺表面存在变形 [Hiremath 等人， 2024]。以往研究尚未广泛报道邻域灰度差异矩阵特征在新冠肺炎预后中的作用，但我们的分析发现，这些特征也是重要的预测变量。

5.3.3 表示元

鉴于整体性能，我们认为模型 3 是最佳模型。然而，在本节中，我们重新审视这三个模型，以突出其各自独特且互补的发现。GLIMARK 模型的可解释性源于其与现有数据点的相似性，即正 α 系数预计与 ICU升级病例相关。为评估这一点，我们在图 $\hat{\boldsymbol{\alpha}}$中比较了三个模型的系数符号与训练点真实标签，图 5。在使用 400 个表示元时，模型 1 的一致率最高（93%），其次是模型 2（87%）。模型 3 的一致率仅为 60%。一种解释是，在基于单个属性比较两个点时，余弦相似度和欧氏相似度都只能提供有限的可解释性。

另一个有趣的发现是，最具代表性的样本会随数据尺度的不同而变化。图 5 显示，在模型 1 中，患者1417在两个片段中都最具影响力。然而，当模型 2 按类别进一步对特征进行分类时，患者1417不再位列前十的最具代表性的样本之中；相反，具有邻域灰度差矩阵L的患者1718成为最具影响力的样本。同样，从模型 2 转换到模型 3 会改变最具影响力的患者，这凸显了个体特征选择的影响。

由于核选择和特征聚合方法并非唯一，我们对胸部X光图像的分析属于探索性研究，而非定论，重点关注图像特征与 ICU结局之间的潜在关系。为此，我们在网络图4-6中展示了每个模型中最具代表性的样本所对应的胸部X光图像，旨在帮助专业放射科医生及其他感兴趣的研究人员解读与ICU升级相关的视觉特征，为未来研究提供参考。

5.4 结论

GLIMARK 使用约 100 个表示元即可达到与随机森林和逻辑回归相当的性能，同时显著降低模型复杂度。它识别出一阶、邻域灰度差异矩阵和二维形状是与结局相关的最重要特征类别，而其他模型

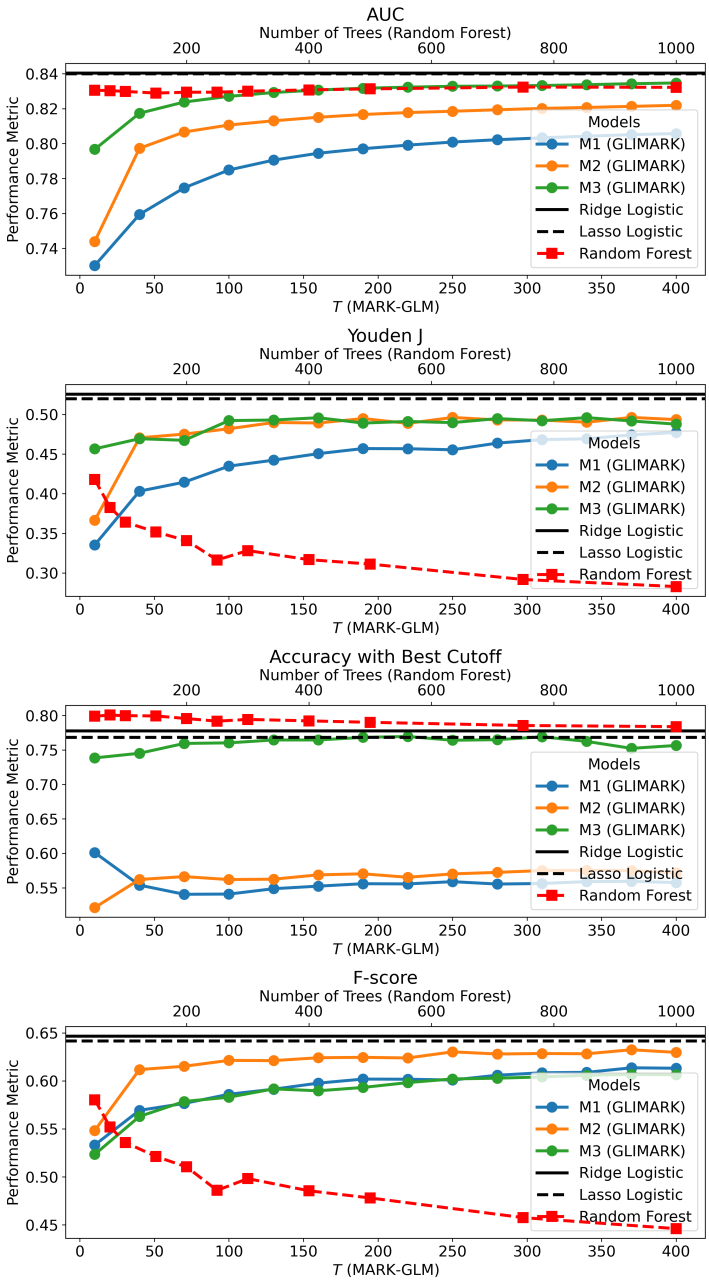


图 3：性能比较。对于随机森林，每个点表示固定 $n_{\text{estimators}}$下的平均交叉验证性能。对于 GLIMARK，由于性能取决于 λ 和 T，因此每个点表示在给定 T时，针对不同 λ 值所获得的最高平均交叉验证性能。19

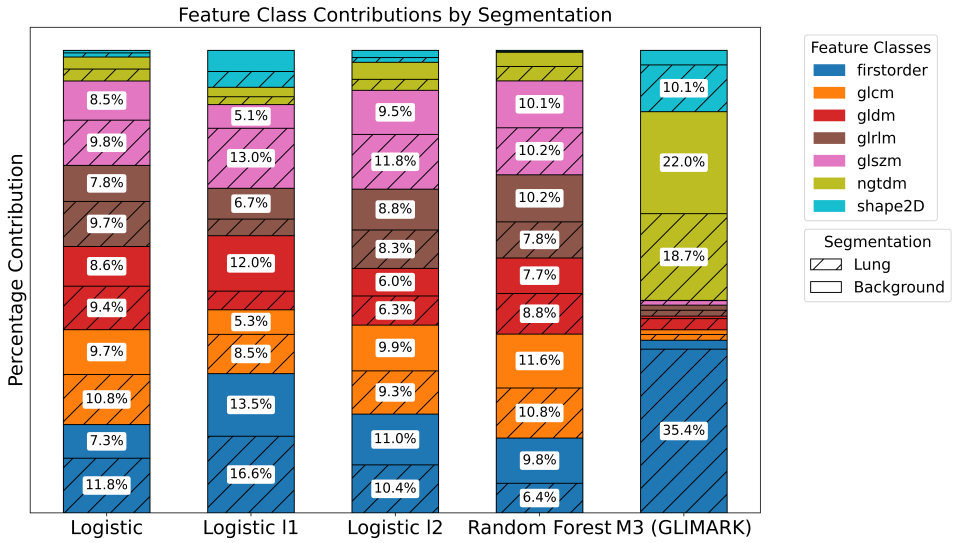
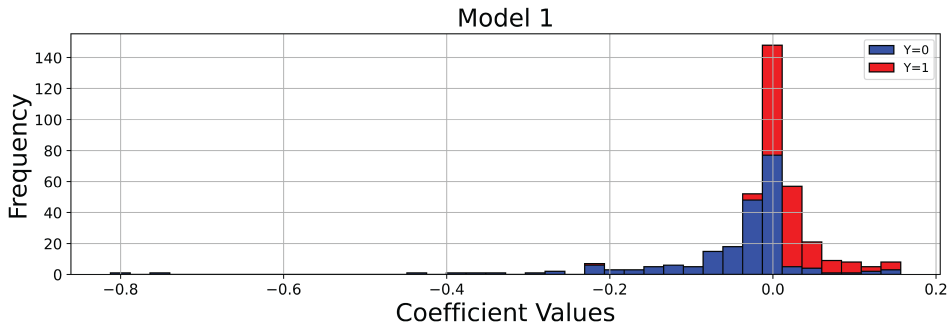
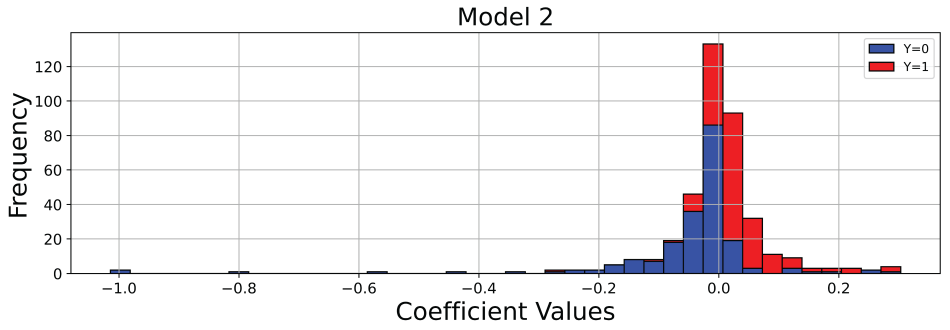


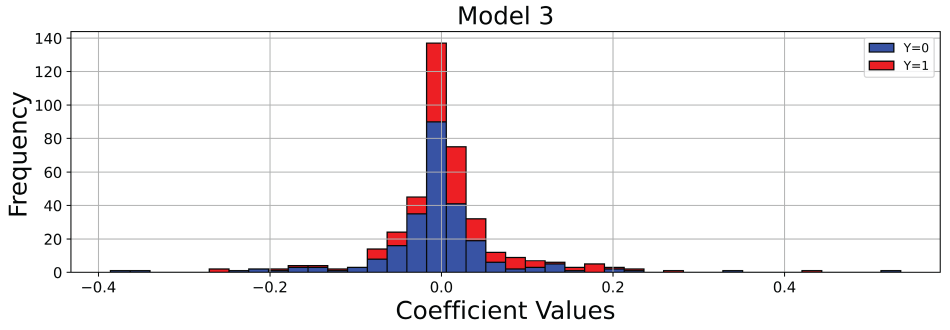
图 4：特征重要性。对于逻辑回归模型，各类别的贡献百分比按该类别内系数值绝对值之和计算。对于随机森林，各类别的重要性基于该类别内基尼重要性之和。除基础逻辑回归模型外，所有模型均采用各自的最佳微调配置。对于 GLIMARK 模型，第三级估计值会合并回其对应的第二级特征类别；我们省略了第一级的少量贡献，因为第一级不提供特征类别估计值。



Coefficient	Patient	Feature	Kernel	Y
-0.81	individual1417	L	RBF: 325.0	0.0
-0.74	individual1417	L	RBF: 550.0	0.0
0.66	individual1567	B	RBF: 1000.0	1.0



Coefficient	Patient	Feature	Kernel	Y
-1.01	individual1718	ngtdm_L	RBF: 10.0	0.0
-1.0	individual72	shape2D_B	Poly: 2	0.0
-0.79	individual1267	glrlm_B	RBF: 150.0	0.0



Coefficient	Patient	Feature	Kernel	Y
0.54	individual2218	original_shape2...	Poly: 2	0.0
0.44	individual783	wavelet-HL_firs...	Poly: 2	1.0
-0.39	individual1698	wavelet-HL_firs...	Poly: 2	0.0

图5：系数贡献和最具代表性的样本。每个模型仅显示影响力最大的三个点。

这表明特征重要性的分布更加均匀。这些发现与放射学文献相一致。此外，GLIMARK 通过突出显示与每项特征相关的最具代表性的患者，增强了模型的可解释性。

6 讨论

GLIMARK 等分层列生成方法的一项优势，在于能够灵活地探索不同尺度的特征。采用高层级特征组的模型会探索特征之间可能存在的交互作用，而低层级模型则隐式地选择特征。它通过一个包含所有特征的单一核来提高模型的复杂度。与核主成分分析不同，它不会映射特征，因此保留了一定程度的可解释性，尤其是在使用手工特征时。

还有一些值得探索的研究方向。首先，尽管我们针对胸部X线数据采用了多视图学习策略，但并未充分发挥 GLIMARK 处理异构数据的潜力。基于器官的分割虽然很常见，却无法进一步探索更小尺度上的局部模式。近年来，随着视觉变换器（ViTs）的普及 [Dosovit- skiy 等人提出了 2020],许多聚焦于较小图像批次的方法。我们数据集中的高质量胸部X光图像，使 Transformer 模型能够生成聚焦于更精细部位（如气道、心脏或膈肌）的注意力图。借助视觉变换器，也可以为侧位图像识别感兴趣区域。因此，可以将更系统的多视图策略构建为一个网格：其中一条轴表示胸部X线视角及其注意力感兴趣区域（例如后前位/前后位与侧位），另一条轴表示影像组学滤波器，第三条轴则表示影像组学特征类别。

其次，我们用于聚合特征和应用核的架构并不全面。领域专用核已得到广泛研究[Shervashidze 等人，2011，Lodhi 等人， 2002]。然而，对于胸部X光图像，相关核的研究有限，值得进一步探索。此外，手工设计的特征并非提取特征的唯一途径。例如，ResNet-50 等深度学习特征也已得到广泛应用。未来的工作可以着重于结合影像组学特征和深度神经网络特征。

致谢

本研究获得美国国家癌症研究所的资助，资助编号为 R01CA269398 和 R01CA249096。我们使用了人工智能辅助工具（例如 Chat-

GPT）来辅助部分代码开发；这些工具未用于撰写或编辑手稿文本。感谢 Stephen Salerno 博士和 Yuming Sun 博士的指导，并感谢他们分享用于整理密歇根DataDirect数据集的代码。

数据可用性

支持本研究结论的数据可从密歇根DataDirect获取。这些数据的使用受到限制，本研究是在获得许可的情况下使用这些数据的。如需获取数据，可在取得必要许可后联系密歇根DataDirect（PHDataHelp@umich.edu）提出申请。

参考文献

Walid Al-Zyoud、Dana Erekat 和 Rama Saraiji。基于阈值的分割在COVID-19胸部X光片图像分析中的应用。*Heliyon*，9(3)，2023年。Yaseen M Arabi、Sheila Nainan Myatra 和 Suzana M Lobo。COVID-19疫情期间重症监护病房需求激增：综述。*Current Opinion in CriticalCare*，28(6):638–644，2022年。Elmehdi Benmalek、Jamal Elmhamdi 和 Abdelilah Jilbab。比较计算机断层扫描与胸部X光成像在COVID-19诊断中的应用。*BiomedicalEngineering Advances*，1:100003，2021。Kristin P Bennett、Michinari Momma 和 Mark J Embrechts。MARK：一种用于异构核模型的提升算法。载于第八届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第24–31页，2002。Christopher M. Bishop。模式识别与机器学习，第7章，第7.1节，第326–328页。施普林格，纽约，美国纽约州，第一版，2006。Mitja Briscik、Marie-Agn`es Dillies 和 S´ebastien D´ejean。改进核主成分分析中变量的可解释性。*BMC Bioinformatics*，24(1):282，2023年。Sema Candemir 等。利用带非刚性配准的解剖图谱对胸部X线片进行肺分割。*IEEE Transactions on MedicalImaging*，33(2):577–590，2013。

Jianhong Cheng 等。基于纵向胸部X线片和临床数据的深度学习重症监护室死亡率预测。*European Radiology*，32(7):4446–4456，2022年。Ayhan Demiriz、Kristin P Bennett 和 John Shawe-Taylor。通过列生成实现线性规划提升。*Machine Learning*，46:225–254，2002。Alexey Dosovitskiy 等。一张图像胜过16×16个词：用于大规模图像识别的Transformer。*arXiv*预印本 *arXiv:2010.11929*，2020。Zhe Fei 和 Yi Li。高维广义线性模型的估计与推断：一种分裂与平滑方法。*Journal of MachineLearning Research*，22(58):1–32，2021。Ronen Fluss、David Faraggi 和 Benjamin Reiser。尤登指数及其相关截断点的估计。*Biometrical Journal: Journal ofMathematical Methods in Biosciences*，47(4):458–472，2005年。Jerome Friedman、Trevor Hastie 和 Robert Tibshirani。关于组套索和稀疏组套索的说明。*arXiv*预印本 *arXiv:1001.0736*，2010。Mehmet G¨onen 和 Ethem Alpaydın。多核学习算法。*The Journal of Machine Learning Research*，12:2211–2268，2011。Yina Han、Yixin Yang、Xuelong Li、Qingyu Liu 和 Yuanliang Ma。通过(r, p)范数实现矩阵正则化多核学习。*IEEE Transactionson Neural Networks and Learning Systems*，29(10):4997–5007，2018。David Harrison Jr 和 Daniel L Rubinfeld。享乐性住房价格与清洁空气需求。*Journal of Environmental Economics and Manage-ment*，5(1):81–102，1978。Amogh Hiremath、Vidya Sankar Viswanathan、Kaustav Bera、Rakesh Shi- radkar、Lei Yuan、Keith Armitage、Robert Gilkeson、Mengyao Ji、Pingfu Fu、Amit Gupta、Cheng Lu 和 Anant Madabhushi。深度学习揭示轻症和重症COVID-19患者基线胸部CT中的肺部形状差异：一项多中心回顾性研究。*Computers in Bi-ology and Medicine*，177:108643，2024。ISSN 0010-4825。doi：https: //doi.org/10.1016/j.compbiomed.2024.108643。

Adam Jacobi、Michael Chung、Adam Bernheim 和 Corey Eber。2019冠状病毒病（COVID-19）中的便携式胸部X光片：图解综述。《临床影像学》，64:35–42，2020年。Diederik P Kingma。Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014年。Gert RG Lanckriet、Nello Cristianini、Peter Bartlett、Laurent El Ghaoui 和 Michael I Jordan。通过半定规划学习核矩阵。*Journal of Machine Learning Research*，5(1月):27–72，2004年。Shao-Bo Lin、Di Wang 和 Ding-Xuan Zhou。带通信的分布式核岭回归。*Journal of Machine Learning Research*，21(93):1–38，2020年。Yen-Yu Lin、Tyng-Luh Liu 和 Chiou-Shann Fuh。用于降维的多核学习。《*IEEE*模式分析与机器智能汇刊》，33(6):1147–1160，2010年。Xinwang Liu 等人。缺失多核学习算法。《*IEEE*模式分析与机器智能汇刊》，42(6):1303–1316，2019年。Huma Lodhi、Craig Saunders、John Shawe-Taylor、Nello Cristianini 和 Chris Watkins。使用字符串核进行文本分类。*Journal of MachineLearning Research*，2(2月):419–444，2002年。Marco E L¨ ubbecke 和 Jacques Desrosiers。列生成的若干专题。《运筹学》，53(6):1007–1023，2005年。Giacomo Meille、Sandra L. Decker、Pamela L. Owens 和 Thomas M. Selden。COVID-19入院率以及美国医院住院患者和重症监护病房占用率的变化。《*JAMA*健康论坛》，4(12):e234206，2023年。doi: 10.1001/jamahealthforum.2023.4206。Ha Quang Minh、Partha Niyogi 和 Yuan Yao。Mercer定理、特征映射与平滑。载于国际计算学习理论会议，第154–168页。施普林格，2006年。Yaohua Rong、Sihai Dave Zhao、Xia Zheng 和 Yi Li。核Cox部分线性回归：为癌症患者生存构建预测模型。《医学统计学》，43(1):1–15，2024年。

Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-Net：用于生物医学图像分割的卷积网络。载于医学图像计算与计算机辅助介入 – *MICCAI 2015*，第18届国际会议，慕尼黑，德国，2015年10月5日至9日，会议论文集，第三部分，第234–241页。施普林格，2015年。B. Scholkopf 等人。基于核的方法中的输入空间与特征空间。*IEEE*神经网络汇刊，10(5):1000–1017，1999年。doi: 10. 1109/72.788641。Yanning Shen、Saeed Karimi-Bidhendi 和 Hamid Jafarkhani。分布式量化在线多核学习。*IEEE*信号处理汇刊，69:5496–5511，2021年。Nino Shervashidze、Pascal Schweitzer、Erik Jan Van Leeuwen、Kurt Mehlhorn 和 Karsten M Borgwardt。Weisfeiler-Lehman图核。*Journal of Machine Learning Research*，12(9)，2011年。Yuming Sun 等人。利用机器学习评估影像组学特征对院内COVID-19死亡率的预后效用。《科学报告》，13(1):7318，2023年。Michael Unser。深度神经网络的表示定理。*Journal ofMachine Learning Research*，20(110):1–30，2019年。Joost JM Van Griethuysen 等人。用于解读放射影像表型的计算影像组学系统。癌症研究，77(21):e104–e107，2017年。Bino Abel Varghese 等人。利用胸部X线片上的影像组学预测COVID-19临床结局。《英国放射学杂志》，94 (1126):20210221，2021年。Grace Wahba 和 Yuedong Wang。表示定理。*Wiley StatsRef:* 在线统计学参考资料，第1–11页，2019年。Piotr Wasilewski、Bartosz Mruk、Samuel Mazur、Gabriela P´o-ltorak-Szymczak、Katarzyna Sklinda 和 Jerzy Walecki。放射影像中的COVID-19严重程度评分系统——综述。《波兰放射学杂志》，85(1):361–368，2020年。William Wolberg、Olvi Mangasarian、Nick Street 和 W. Street。威斯康星乳腺癌（诊断）。UCI机器学习库，1993年。DOI: https://doi.org/10.24432/C5DW2B。

Zhiyun Xue、FengYang、Sivaramakrishnan Rajaraman、Ghada Zamzmi 和 Sameer Antani。CXR肺区检测中域偏移的跨数据集分析。*Diagnostics*, 13(6):1068，2023年。

Xiaoqiang Yan、Shizhe Hu、Yiqiao Mao、Yangdong Ye 和 Hui Yu。深度多视图学习方法：综述。*Neurocomputing*，448:106–129，2021。

Ji Zhu 和 Trevor Hastie。核逻辑回归与导入向量机。计算与图形统计学杂志，14(1):185– 205，2005年。

补充材料

其他表格、图和技术细节详见随附的补充材料。