arXiv:2602.21876v1 [stat.AP] 25 Feb 2026

用于预测供体肾脏弃用的机器学习模型比较评估

Peer Schliephacke[1†], Hannah Schult[1†], Leon Mizera[1†], Judith W¨urfel[1†], Gunter Grieser[2], AxelRahmel[3], Carl-LudwigFischer-Fr¨ohlich[3], Antje Jahn-Eimermacher[1*]1*达姆施塔特应用科学大学数学与自然科学系，Sch¨offerstraße 3，德国黑森州达姆施塔特，64295。[2]达姆施塔特应用科学大学计算机科学系，Sch¨offerstraße3，德国黑森州达姆施塔特，64295。[3]德国器官移植基金会，Deutschherrnufer 52，德国黑森州法兰克福，60594。

*通讯作者。电子邮箱：antje.jahn@h-da.de；
†这些作者对本研究的贡献相同。

摘要

肾移植可以改善终末期肾衰竭患者的预期寿命和生活质量。如果能够降低被弃用而未移植的肾脏比例，就能让更多患者受益于肾移植。在这一背景下，机器学习（ML）可以通过及早识别弃用风险较高的供体器官来辅助决策，例如及时采取干预措施，以提高器官利用率，如实施挽救性分配。尽管已有多种机器学习模型投入应用，但由于数据集各不相同，且特征工程和评估策略存在差异，其结果难以比较。本研究旨在对用于预测供体肾脏弃用的机器学习模型进行系统且可复现的比较。我们使用德国4,080名已故供体（死亡判定依据为神经学标准）的数据，训练了五种常用机器学习模型：逻辑回归、决策树、随机森林、梯度提升和深度学习，以及一个集成模型。我们建立了统一的基准测试框架，其中包括标准化的特征工程与特征选择，以及贝叶斯超参数优化。我们从判别能力（MCC、

曲线下面积、F1分数）、校准（布里尔分数）和可解释性（SHAP）方面评估模型性能。集成模型的判别性能最高（MCC=0.76、曲线下面积=0.87、F1=0.90），而逻辑回归、随机森林和深度学习等单一模型表现相近，且优于决策树。Platt缩放提升了基于树和神经网络的模型的校准效果。SHAP在各模型中均将供体年龄和肾脏标志物识别为主要预测因子，这符合临床常识。本研究表明，统一的数据预处理、特征选择和评估策略，对预测成功的影响可能大于机器学习算法的选择。

关键词： 器官移植、供体肾脏弃用、机器学习、校准、可解释人工智能、特征选择

1 引言

供体器官全球短缺仍是现代医学面临的重大挑战。尽管移植需求持续上升，器官捐献却有所波动 [1] ，且往往无法满足临床需求。与此同时，供体器官的质量有所下降，供体年龄不断增加，每位供体可使用的器官数量减少 [2]。这些情况使分配与接受决策既高度复杂又极为紧迫，因为保存时限与缺血时间相关风险施加了严格的时间约束。因此，相当比例已获取的供体肾脏最终被弃用。因此，在此背景下一个尤为相关的研究问题是尽早识别具有高弃用风险的供体器官，例如以便实施提高器官利用率的干预措施，如及时开展补救分配，或应用机器灌注以改善器官质量。

机器学习（ML）模型可以利用大量可用数据支持临床决策，包括活检报告、实验室检验值和自由文本形式的临床记录。因此，各种机器学习模型已应用于器官移植领域，且往往取得了令人鼓舞的结果 [3–8]。其中，肾移植因可用病例相对较多而成为经常研究的领域。常用的机器学习模型包括决策树（DT）和回归模型，以及随机森林（RF）、梯度提升和人工神经网络等更灵活的方法。这些方法通常较难解释，但可能具有更高的预测性能。这些模型也被用于预测肾脏弃用风险这一具体任务，但结果各不相同：逻辑回归（LR）仍是该领域最常用的方法之一，可识别与器官弃用相关的关键预测因子，例如活检结果和较高的肾脏供体档案指数评分 [5, 9–13]。一些研究报告称，采用梯度提升或随机森林等更先进的机器学习模型可提高预测性能 [7, 14]。另一些研究则发现，机器学习方法的表现与逻辑回归相近 [15]。然而，由于各研究所用的数据集差异很大，因此难以直接比较。例如，一项研究分析了包含数百个预测因子的大型全国性数据集，而另一项研究依赖的是规模小得多的单中心数据，

样本量和有限的一组预测因子。此外，特征选择的作用 [16] 常常被忽视，这可能会使具有内置特征选择机制的模型（如随机森林）占据优势。因此，难以系统地判断哪些模型最适合预测肾脏弃用风险，进而支持器官移植中的临床决策。本研究旨在弥补这一空白：在同一个肾移植候选者数据集上采用统一的预处理和转换流程，对不同的机器学习方法进行系统比较，以预测肾脏弃用风险。

本研究有两项主要贡献。第一，系统评估了多种先进的机器学习模型，并采用模型校准和贝叶斯超参数优化等先进技术，以确保比较客观。第二，除了比较预测性能外，还通过使用 SHAP 比较模型的可解释性，从而识别关键预测因子，并支持临床可解释性。

由于保密限制，本研究使用的专有数据无法共享，但我们全面开放了所有配套脚本，涵盖数据处理、特征工程与选择、模型优化、训练、评估、校准和可视化，以提高透明度和可复现性。

本文结构如下。第 2 节介绍数据、机器学习模型和特征工程策略，并概述模型评估方法。第 3 节呈现模型比较结果，包括判别性能、模型校准和可解释性。最后，第 4 节总结全文，并指出未来研究方向。

2 方法

2.1 数据

本研究使用了2020年4月至2024年8月期间德国已故成年供者提供用于移植的所有肾脏数据。该数据集由德国器官获取组织（DSO）提供，包含 $n = 4080$ 名成年供者的信息（所有死亡均依据神经学标准判定）。预测肾脏是否会被弃用时，仅考虑与供体相关的特征，有意排除受者特征。这样一来，仅凭器官提供时可获得的信息，便能识别出弃用风险较高的供体肾脏，而此时尚未将其分配给特定受者。

德国器官获取组织（DSO）以独立研究数据库的形式提供了供体数据，用于二次分析。所有身份标识均已删除，数据均已匿名化。该数据集包含DSO根据国家移植法（TPG）必须向国家移植登记处提供的数据。对这些匿名化数据进行的二次分析符合移植法的规定。

若某器官曾被提供用于肾移植，但最终因任何原因未被移植，则定义为已废弃。为确保观测值相互独立并避免个体内相关性，弃用状态按供体层面而非器官层面判定。在大多数情况下（4,080 例中的 3,716 例），同一供体的两枚肾脏要么均

已移植，要么均已废弃，因此供体层级分类较为简单。在 364 例中，仅一枚肾脏已移植，另一枚则已废弃。这些病例在供体层面归类为“已移植”。总之，所有供体中有 929 例（ 22.8% ）被归类为已废弃，3151 例（ 77.2% ）被归类为已移植（即未弃用）。需要注意的是，少数供体的肾移植状态存在不一致。由于数据经过匿名化处理，这些不一致无法得到解决，因此分析时采用了数据中报告的状态。

提供多个器官的供体不会被排除，因为器官组合不会影响肾脏弃用。不过，这些器官组合可能会影响移植后功能，而这并非本研究的重点。

数据包括对身体各部位检查所得的结果、已开具药物的信息、病毒学、病理学和微生物学领域的实验室测量值，以及住院情况和医学诊断的详细信息，其中包括 ICD-10 编码（国际疾病和相关健康问题统计分类） [17]。总而言之，p=1128 个变量可供使用，并在特征工程步骤中进一步处理（见第2.2节）。无法对研究队列的所有这些变量进行全面描述。因此，我们提取了后来被证明对预测最重要的特征，并在表1中进行了总结。

数据集根据供体 ID 进行划分，将 80% 的供体 ID 分配至训练集，20% 分配至测试集，以确保训练数据与评估数据相互独立。特征工程分别在两个数据集中进行，阈值仅根据训练集计算，并应用于测试集。此外，从训练数据中固定留出 10% 作为专用验证集，用于最终模型训练和校准，并确保其与用于最终评估的留出测试集严格分离。

2.2 特征工程

本节介绍特征变换，包括时间序列处理、冗余降维、分类变量编码和缺失值填补。附录中还介绍了针对特定特征的一些其他变换B。

为避免选择偏倚，变量在未经医学顾问预先筛选的情况下进行了选择、提取和特征工程。仅在存在不确定性的情况下，主要研究者才会咨询医学专家，以确保数据标准化恰当（例如，将供体尿量按供体体重和标准时间间隔进行调整，如 $ml/kg/h$）。医学顾问还进行了二次验证，以确保数据反映逝者器官捐献程序中的真实情境。

2.2.1 时间序列特征提取

许多实验室测量值以时间序列的形式出现。49个变量被识别为时间序列。另有40个变量未被视作时间序列，因为超过50%的供体对这些变量仅有一次记录值。对于这些变量，我们选择了最近一次记录值或首次记录值。我们将时间

表 1 图中报告的最重要特征的描述性统计 6。未纳入独热编码生成的各个虚拟变量，因为它们无法涵盖原始分类变量的全部信息。

特征	所有供体 $N = 4080$	已移植 $N = 3151$	已废弃 $N = 929$
原始报告的变量：			
年龄（年），中位数 [四分位距]	58.0 [46.0，69.0]	56.0 [44.0，66.0]	68.0 [56.0, 78.0]
利尿量（$kg/24h/kg$（体重）），中位数 [四分位距]	36.88 [24.21, 53.33]	38.55 [26.52，57.1]	30.59 [15.06，42.86]
给予肝素类药物，n（%）	3873（94.9）	3128（99.3）	745（80.2）
从时间序列中提取的变量：			
MDRD估算肾小球滤过率（$ml/min/1.73m^2$），中位数 [四分位距]			
- 最小值	69.69 [44.92，95.12]	76.45 [53.84，100.42]	39.48 [19.81，66.8]
肌酐（g/l），中位数 [四分位距]			
- 最小值	0.78 [0.6, 1.0]	0.72 [0.58, 0.9]	1.02 [0.77，1.5]
- 最大值	97.2 [74.22, 141.61]	88.4 [70.7, 123.8]	154.7 [97.2, 291.7]
- 截距	81.58 [61.93, 116.02]	75.92 [59.47，101.44]	118.88 [84.0，197.31]
- 最后一个	79.6 [61.9, 114.9]	75.1 [59.2, 99.0]	118.5 [79.6, 193.6]
- 标准差	11.22 [5.79, 22.55]	10.3 [5.36，18.8]	19.59 [7.47, 59.77]
亚硝酸盐，中位数 [四分位距]			
- 计数	1.0 [1.0, 1.33]	1.0 [1.0，1.32]	1.0 [1.0，1.33]
尿素（$mmol/l$），中位数 [四分位距]			
- 最小值	4.3 [3.0，5.9]	4.0 [2.8，5.3]	5.9 [4.2, 8.0]
- 最大值	7.0 [5.0, 10.0]	6.3 [4.7, 9.0]	9.9 [6.8, 16.3]
血气 吸入氧浓度 (1/1)，中位数 [四分位距]			
- 第一	0.51 [0.4，0.7]	0.51 [0.4，0.79]	0.54 [0.4，0.68]
血气 pH、中位数 [四分位距]			
- 计数	2.0 [1.0，3.0]	2.0 [1.0，3.0]	2.0 [1.0，2.34]
钾（$mmol/l$），中位数 [四分位距]			
- 最大值	4.5 [4.2，4.9]	4.5 [4.2，4.9]	4.7 [4.4，5.2]
国际标准化比值（1/1），中位数 [四分位距]			
- 第一	1.1 [1.01，1.25]	1.1 [1.0，1.23]	1.14 [1.05，1.3]
亚硝酸盐阴性，n（%）			
- 第一	3499（85.8）	2830（89.8）	669（72.0）
葡萄糖阴性，n（%）			
- 第一	2928（71.8）	2392（75.9）	536 (57.7)
蛋白质定性阴性，n (%)			
- 第一	2069 (50.7)	1772 (56.2)	297 (32.0)
蛋白质定性阴性，n (%)			
- 第一	2082 (51.0)	1790 (56.8)	292 (31.4)

序列变量划分为以下亚型，以最大化每个变量所提取的信息：

- 类型1：仅包含表示阳性或阴性结果的分类值的时间序列变量。

- 类型 2：数值型时间序列变量，每位捐献者的平均观测数低于 2。

- 类型 3：数值型时间序列变量，每位捐献者的平均观测数为 2 或以上。

对于每个时间序列变量，我们为每位供体提取了以下特征：首次记录值、最后记录值、条目数，以及首条与末条记录之间的时间跨度（以小时计）。

对于具有数值的时间序列变量（类型 2 和类型 3），我们还提取了标准差、最小值和最大值。

对于类型 3 的时间序列变量，我们进一步提取了描述其随时间变化的线性趋势的特征：针对每位供体进行线性回归，以自首次记录以来经过的小时数作为自变量，以变量值作为因变量进行建模。因此，每位供体 d 的线性回归模型为：

$$y_{d,i} = \beta_{d,0} + \beta_{d,1} \cdot t_i + \varepsilon_{d,i} \tag{1}$$

其中，$y_{d,\,i}$ 表示变量在时间点 t_i 的取值。估计得到的系数 $\beta_{d,\,0}$ 和 $\beta_{d,\,1}$ 被用作额外特征。

以往研究使用最后记录值 [18] 或最小值 [18, 19]。我们额外提取的特征可能捕捉到随时间发生的变化：条目数可能表明供体的变量值随时间发生了变化。同样，第一条与最后一条记录之间的时间跨度也能反映供体状况的时间动态。标准差也能揭示变量值的变异程度。估计得到的线性回归系数包含有关数值总体上随时间增加还是减少，以及变化幅度的信息。总之，我们从时间序列数据中提取了共 364 个特征。

2.2.2 供体用药数据的特征提取

原始数据中有 1772 个不同的药物名称。这些名称的结构是：第一个词描述药物，其余词提供剂量等附加信息。仅考虑每条药物记录的第一个词后，药物名称的数量减少至 1103 个。随后，选取出现频率最高的 40 个药物名称作为最终特征集。这一选择涵盖了数据集中约 80% 的药物给药记录。我们将这些药物名称转换为二元特征，用于指示供体是否使用过该药物。

2.2.3 分类变量编码

对于取值超过两个的分类变量，我们采用独热编码。因此，对于包含 ICD-10 编码的变量，我们将训练集中供体相对频率低于 1% 的编码替换为缺失值。

2.2.4 缺失值填补

我们的数据中约有 14% 的值缺失，其中 628 个变量包含不完整数据。除非另有说明，我们在训练集和测试集中均按照以下相同方式进行插补：

- 250根据专家建议，可利用逻辑依赖关系对部分变量进行插补。这些变量包括某些诊断相关变量，其中缺失值可解释为阴性（即未发现诊断）。
- 对于某些分类变量，缺失值予以保留并视为单独类别。这些变量的缺失值可能会影响是否弃用的决策（例如“尿糖”）。
- 对于两个变量（“心肺复苏持续时间”和“体外膜肺氧合”），自由文本变量提供了额外信息，并用于插补。

- 30 描述生理参数的连续变量通过从正态分布中随机抽取样本进行插补，该分布的均值和标准差分别取自各变量，并仅保留落在分布中央 95% 范围内的值。该分布针对训练集和测试集分别计算。
- 对于训练集中缺失值比例较高（>70%）的 20 个变量，我们删除了其具体取值，并将这些变量二分为“缺失”和“未缺失”。在这些情况下，基于稀疏数据进行插补被认为不可靠，因此未进行插补。
- 训练集中其余 328 个存在缺失值的变量，均采用迭代插补算法进行插补 [20] 该算法使用岭回归预测训练集和测试集中的缺失值。如果测试集中有些变量存在缺失值，而训练集中这些变量没有缺失值，则使用该变量的均值填补。选择这种方法是为了考虑变量之间的复杂关系。

最后，我们使用基于直方图的梯度提升分类树评估了这一插补策略的效果。HistGradientBoostingClassifier 具有内置的缺失值插补方法 [21]，因此可以将基于该方法的结果与采用我们的插补策略预先插补后的结果进行比较。我们使用网格搜索和五折交叉验证，分别在最终插补后的特征集和未经插补的原始特征集的训练集上训练 HistGradientBoostingClassifier。我们比较了模型在测试集上的表现（F1分数和马修斯相关系数）。采用我们的插补策略所得到的分数优于 HistGradientBoostingClassifier 的内置策略，表明该插补策略是有效的。

2.3 建模方法

以下部分概述了我们的建模方法，包括应用于肾移植数据集的机器学习算法、特征选择技术和超参数优化方法。

2.3.1 机器学习算法

训练以下机器学习算法，以预测供体器官被弃用或移植的概率。

决策树

决策树（DT） [22]是用于分类与回归任务的监督学习模型。它们根据特征阈值递归地将数据划分为多个子集，形成树状结构，其中每个内部节点代表一项决策，每个叶节点给出最终预测。决策树以可解释性强而著称，因为每次划分都基于由基尼不纯度等不纯度度量得出的简单规则；但它们容易过拟合 [22], ，尤其是在树很深时。

逻辑回归

逻辑回归（LR） [23] 是一种属于线性模型族的机器学习算法，常用于二分类任务。逻辑回归通过对输入特征的线性组合应用 Sigmoid 函数（逻辑函数）来建模结果的概率。回归模型通常使用最小二乘法来求取每个特征对应的参数，使其相对于目标变量的平方误差和最小 [23]。它的另一项优势是可解释性强，因为可以直接从模型系数推断特征重要性。

随机森林

随机森林（RF） [24]是一种集成学习方法，通过构建多个决策树并汇总其输出，提高预测准确率并减少过拟合。该森林模型采用装袋法，让每棵树在数据和特征的随机子集上进行训练，从而比单棵决策树更不容易过拟合。另一个优点是可以汇总所有树的特征重要性度量，从而增强模型的可解释性。

极端梯度提升

极端梯度提升（XGBoost） [25] 是一种基于梯度提升 [26], 的集成学习算法，适用于分类和回归任务。该模型依次构建决策树，每棵新树都会纠正前面各棵树的错误，并通过梯度下降优化性能。XGBoost 在结构化数据 [27]上表现尤为出色，甚至可以胜过深度学习算法 [28]。另一个优点是提供特征重要性度量，即使模型复杂，也能增强其可解释性。

多层感知机

多层感知机（MLP）（也称为深度学习） [29] 是一种用于分类与回归任务的人工神经网络。它由多个互连神经元层组成，包括一个输入层、一个或多个隐藏层以及一个输出层；每个神经元都会先对输入进行加权求和，再应用非线性激活函数。多层感知机能够捕捉复杂的非线性关系，但需要仔细调整超参数，并使用大型数据集才能实现良好的泛化。虽然它更常用于图像或文本等非结构化数据，但也可用于结构化表格数据。

集成模型

实现了一种称为数据集成模型（差分进化，DE）的集成模型，用于结合多个预训练分类器的预测结果。本研究中的差分进化模型整合了逻辑回归（LR）、XGBoost（XGB）、随机森林（RF）和多层感知机（MLP）的输出。每个模型都会为两个类别分别生成一个概率，集成模型则通过计算各模型预测概率的均值来汇总这些输出。最终分类结果由平均概率最高的类别决定。决策树被

排除在集成模型之外，因为其性能相对较低，后文将对此进行说明。

2.3.2 特征选择

为应对数据集的高维特性，我们采用特征选择方法，找出最有前景的特征子集。这可以减少冗余、提高模型可解释性，并通过聚焦于与预测最相关的变量来防止过拟合 [16]。由于不同的机器学习模型以不同方式捕捉关系，我们针对每个模型分别进行特征选择，但始终采用相同的方法。这样可确保每种算法使用的特征集都适合其特定的归纳偏置和复杂度。为捕捉特征之间的非线性和多变量交互作用，我们采用一种基于遗传算法的优化方法 [30] ，利用其搜索大型搜索空间的能力 [31]。这种方法能够发现较简单的单变量滤波器（例如相关性或卡方检验）可能会遗漏的复杂关系 [32, 33]。

如图所示，针对每种机器学习模型分别应用这一通用方法1 ，具体说明如下。

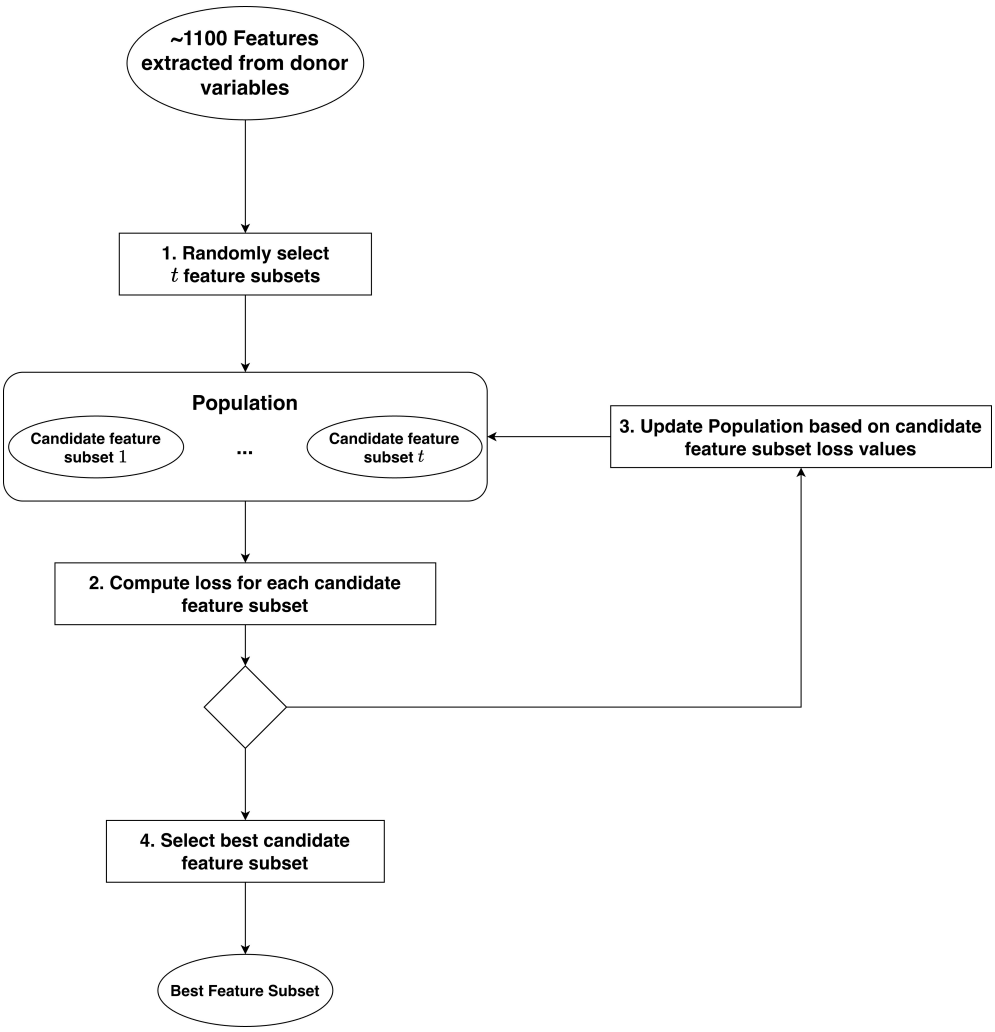

图 1 特征选择过程示意图。从完整特征空间开始，通过训练和验证监督模型创建并评估多个候选特征子集。根据预测性能为每个子集分配损失值，并选择损失最小的子集作为最终特征子集。

我们采用非支配排序遗传算法 II（NSGA-II） [34] 进行单目标特征选择优化，因为与其他进化策略 [35]相比，它能够原生处理二元和分类型参数。我们使用 Optuna [36]实现的 NSGA-II 算法。

1. 第一，将一个由50个随机选择的特征子集组成的种群选为候选。此处特征集表示为长度为 ~ 1100 的布尔向量（数量

特征集中的特征数量），其中每个元素代表一个特征，并根据该特征是否包含在子集中而设置为1或0。

2. 接下来，对于每个特征子集，我们仅使用相应的特征来训练正在研究的机器学习算法，并根据损失指标评估其性能。下面将更详细地介绍这些性能指标的计算过程：

- 我们通过嵌套超参数优化来构建模型架构：针对每个候选特征子集，我们执行随机超参数搜索 [37]，进行 10 次试验，以避免默认模型超参数针对特定特征子集发生过拟合。在这 10 次试验中，我们均使用 3 折交叉验证评估生成的模型配置，并训练和验证机器学习算法。

- 通过 3 折交叉验证得到的验证损失使用归一化的马修斯相关系数（MCC，另请参见第 2.4.1 节）[38], 计算。该指标是一种平衡指标，可将真/假正例和真/假负例合并为单一相关性得分；对于类别不平衡场景，MCC 比其他常见指标能提供更有参考价值的评估 [39, 40]。
- 某个子集的总体损失定义为内层搜索产生的 10×3 验证损失的均值，再加上针对特征数量的显式惩罚项，即 $\lambda \cdot n_{\text{特征}}$（其中 $\lambda = 0.0005$），以抑制选择庞大且冗余的特征子集。

3. 最后，选出表现最佳的特征子集，并以此作为进化操作（选择、交叉和变异）的基础，从而产生下一代特征子集种群。对这些子集重复上述过程。4. 经过1000次迭代后，该过程终止，并分别为每种机器学习算法得出表现最佳的特征子集作为结果。

我们的方法沿用了相关领域的既有工作，但不同之处在于采用按模型的多目标进化搜索以及嵌套的随机架构评估策略，二者共同旨在为每个分类器生成稳健、紧凑的特征集，而非单一的全局子集 [41, 42]。

此外，插补后冗余或取值恒定的93个特征已被剔除。

2.3.3 模型超参数优化

特征选择产生了五个经过优化的特征子集，分别针对每个机器学习模型进行定制。随后，为了确定每种算法的最佳架构，我们针对相应的优化特征空间进行专门的超参数优化。

我们使用带有树结构帕岑估计器（TPE）采样器的 Optuna [43]作为贝叶斯超参数策略 [44]，为每种机器学习算法确定最佳超参数配置。优化目标是 5 折交叉验证归一化 MCC 的均值。对于每种算法，我们均进行了 300 次试验，相应的超参数搜索空间列于表 C1-C5，选定的（最佳）架构则见表 C6-C10。

2.4 模型评估

以下各节将介绍模型评估的具体步骤。我们首先报告判别性能的评估方法，包括模型之间的统计比较；随后介绍对预测概率进行校准分析的方法；最后总结如何利用基于 SHAP 的归因获得可解释性结果。

2.4.1 判别性能

为最终评估模型性能，我们使用 30 个不同的随机种子重新训练每种选定的超参数配置，以量化性能变异性，这符合推荐的做法 [45–47]。

因此，随机种子控制了学习算法中固有的随机性来源，有助于提高可复现性，并能够评估随机初始化效应导致的性能变异性。对于 DT，当多个候选项带来相同的不纯度降低时，随机性可能会影响特征阈值的选择。在 LR 中，使用 SAGA 求解器 [48], 时，迭代优化过程中对数据子集进行采样会产生随机性。在 RF 等基于集成模型的方法中，随机种子决定训练样本的自助采样，以及每次分裂时的随机特征选择，直接影响树多样性，进而影响整体模型方差 [49]。类似地，在 XGBoost 中，随机种子决定构建单棵树时对行和列进行子采样的随机因素，从而影响偏差与方差之间的权衡。最后，对于 MLP，随机种子决定网络权重和偏置的初始化，以及训练过程中的小批次；这两者都会显著影响收敛和最终性能 [50–52]。使用多个随机种子重新训练每个模型，可以捕捉可达到性能的分布，并量化每种算法固有的稳定性或变异性，而不局限于单次运行的性能估计。

重新训练采用原始训练-验证集划分：通过随机划分，将 90% 的数据用于训练，10% 用于验证。特征标准化采用 Z 分数缩放，缩放参数（相应特征的均值和标准差）仅根据训练集计算，随后应用于验证集和测试集，以防止信息泄漏。模型性能在留出测试集上进行评估，该测试集未用于训练或验证。

我们针对每次种子运行报告三项互补的性能指标：F1分数（见公式 2 [53]）、ROC 曲线下面积（曲线下面积） [54]，以及马修斯相关系数（见公式3 和 4 [39, 40]）。在公式 2-4中，TP、TN、FP 和 FN 分别表示真正例、真负例、假正例和假负例的计数。在供体肾脏接受的语境中，真正例（负例）对应于被正确预测为已移植（已废弃）的肾脏，假正例和假负例则相反。各项指标在所有种子运行中收集，并以分布而非单点估计的形式报告，以便可视化每种算法的变异性。

F1分数（范围为 $0-1$）是精确率（亦称阳性预测值）与召回率（亦称灵敏度）的调和平均数，侧重于正类表现，但未能计入真负例，因此在类别不平衡下可能具有误导性 [39]。曲线下面积衡量分类器为随机选取的已移植器官赋予比随机选取的废弃器官更高的成功移植概率的能力（范围为 0–1，其中 0.5 表示随机表现），但对于小型或不平衡的数据集可能过于乐观 [55, 56]。基于这些原因，我们强调马修斯相关系数，因为它能为不平衡二分类提供更具信息量的评估，涵盖混淆矩阵的全部四个组成部分，并在类别规模不同时仍给出均衡的度量 [39, 40]。

为确定模型之间观察到的差异是否具有统计显著性，我们首先进行方差分析（ANOVA）[57]，随后使用 Tukey诚实显著差异（HSD）事后检验 [58] 对模型进行两两比较。

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3}$$

$$MCC_{Normed} = \frac{MCC + 1}{2} \tag{4}$$

2.4.2 模型校准

模型校准在临床领域至关重要，因为经过良好校准的概率能够传达可靠的预测不确定性，并有助于做出更安全的决策 [59–61]。我们使用布里尔分数 [62, 63]，量化校准程度，其定义为预测概率与真实二元结果之间的均方误差（范围为 0–1，数值越低越好）[64, 65]。

事后校准会调整训练后模型的预测概率，使其更好地匹配真实结果频率，从而提高概率估计的可靠性。

具体而言，后校准方法会转换原始预测分数，例如分类器未经校准的输出（如决策函数或 sigmoid前对数几率），将其转化为经过良好校准的概率估计。我们比较两种标准的后校准方法。Platt缩放 [66]拟合一个参数化的 Sigmoid函数（一个双参数逻辑模型，通常以逻辑回归的形式实现），将这些分数映射为概率。其参数在验证数据上进行优化，然后应用于测试集预测。相比之下，保序回归 [67]，提供一种非参数单调映射，能够对原始分数与真实概率之间更复杂的偏差进行建模。在我们的设置中，每个校准器都针对每个模型单独使用验证集预测进行训练（校准模型采用默认超参数），应用于测试集，并使用布里尔分数进行评估，同时绘制可靠性图以直观呈现残余失校准。

请注意，由于每个基模型都是用不同的特征子集训练的，我们在后校准分析中排除了差分进化模型。因此，事后

校准方法无法应用于组合预测。此外，校准结果并未对最终重新训练的 30 个不同随机种子进行汇总，因为目前没有简单的方法可以合并校准图。因此这里只展示了最佳测试运行对应的结果。虽然这可能会导致校准估计略微过于乐观，但主要目标（模型比较）不会受到影响，因为这一点适用于所有模型。

2.4.3 可解释性

为评估可解释性，我们使用 SHAP [68] 作为一种模型无关方法，可统一应用于不同的模型家族。SHAP 基于合作博弈论中的 Shapley值 [69]。对于每个观测值，我们计算各特征对预测风险与边际（平均）种群风险之差的贡献。因此，所有特征的贡献之和等于模型输出相对于基线（期望）预测的偏差。可加性使 SHAP 适用于局部（单个实例）解释和汇总后的全局解释。在此，我们通过汇总所有观测值中各特征的SHAP值，着重进行用于模型比较的全局解释。由于全局特征重要性关注的是幅度而非正负，因此使用绝对SHAP值进行汇总。每个特征均以其平均绝对SHAP值概括，并附上相应的标准差，以反映不同样本和运行之间的变异性。此外，为了进行详细的模型分析，我们还通过蜂群图考察个体可解释性结果，以捕捉实例级行为。

SHAP已广泛应用于表格数据和临床机器学习研究，用于突出预测驱动因素并支持临床医生解释 [70–72] ，尤其有助于提升医学预测任务的透明度 [61]。

在模型比较并非主要关注点的单项研究中，可以使用模型专属解释器作为合适的工具来解释结果。因此，我们在SHAP结果的基础上补充了模型特定可解释性方法（例如决策树特征重要性、线性模型系数）；这些方法可能受到过拟合的影响，且并不适用于所有所用模型家族 [73–75]。

我们使用同一个解释器对象（置换）实现SHAP，因为SHAP是一种与模型无关的可解释性框架。训练数据集用作背景数据，模型的输出概率则充当SHAP与算法之间的链接函数。

与DE模型的后校准类似，由于异质特征集上的特征归因无法比较，因此无法应用SHAP。此外，我们仅报告表现最佳的测试运行的SHAP值和图表，因为汇总多个随机种子的特征归因不会产生有意义或可解释的结果。

3 结果

本节首先展示并讨论所有模型的判别性能。随后，我们分析其概率校准，最后评估模型的可解释性。

3.1 判别性能

表2 汇总了最终重新训练和评估运行的测试性能得分。差分进化在所有性能指标上均优于其他模型，在曲线下面积和F1分数方面的提升尤其显著。相比之下，差分进化、多层感知机和逻辑回归的归一化 MCC 差异相对较小。

表 2 第2.4.1 节定义的六种评估模型（DT、LR、RF、XGB、MLP 和 DE）的性能评分。数值表示测试集上的 F1分数、曲线下面积和归一化 MCC 均值。最佳性能评分以粗体突出显示。

模型	F1	AUC	归一化 MCC
DT	0.7336	0.6892	0.6219
LR	0.8192	0.8564	0.7515
RF	0.8136	0.8420	0.7334
XGB	0.7985	0.8219	0.7119
MLP	0.8279	0.8499	0.7550
DE	**0.9042**	**0.8721**	**0.7588**

图 2（左）展示了不同随机种子下各模型基于归一化测试集 MCC 的模型性能分布：DT 表现最差，而 XGB 排名倒数第二，但强于 DT。DE 的得分显著最高；根据 Tukey-HSD 事后检验，RF、MLP 和 LR 的表现与集成模型相当，差异不显著。图 2 （右）展示了经过调整的两两 Tukey p 值，颜色表示统计显著性。

附录图 D1 展示了测试F1分数的相应性能图，而附录图 D2 展示了测试 AUC 分数的性能图。表现最佳模型的排名保持一致。集成模型的F1分数显著高于所有其他模型。对于测试曲线下面积，集成模型同样排名第一，仅逻辑回归在统计上并非显著更差。决策树模型在全部三项分数上的表现都低得多。

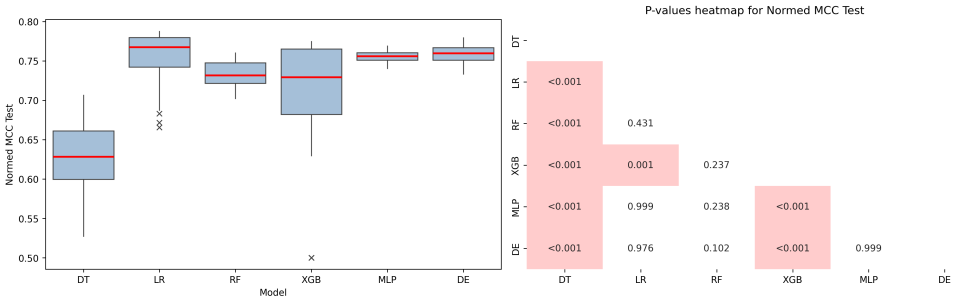

图2 左侧：六种不同机器学习模型的判别性能评估。每个面板均报告多种方法的标准化测试集马修斯相关系数（纵轴，数值越高越好）：XDT、逻辑回归、随机森林、XGBoost、多层感知机和差分进化。箱线图展示了各方法的性能分布。**右侧：**显示模型两两统计显著性的 Tukey 校正 p 值下三角热图。

在我们的实验中，决策树表现最差，这与既往关于临床表格数据的研究结果一致 [76, 77]。单棵树往往缺乏处理复杂模式所需的能力，并且容易过拟合 [78]。

对于每项测试指标，集成模型始终表现最佳，这与集成模型能够降低方差、提高评分准确性的证据一致 [4, 79, 80]。在我们的设定中，这一优势更加明显，因为每个基模型都通过特征选择，在针对其模型族定制的数据空间上进行训练。这种多样化降低了误差相关性，并增强了组合预测。

集成模型在 F1分数和曲线下面积上的提升最大，而在马修斯相关系数上的差距较小。这表明，在我们的研究情境中，马修斯相关系数是一项要求更高、信息量更大的指标。随机森林、多层感知机和逻辑回归取得了有竞争力的马修斯相关系数分数，这说明此前的特征选择和超参数优化是有效的。

3.2 模型校准

表 3 报告了 Brier 分数，对各模型采用的三种方法进行概率校准比较：不进行后校准、使用 Platt 缩放进行后校准，以及使用保序缩放进行后校准。数值越低，表示概率校准效果越好。在此数据集中，对 DT、RF 和 MLP 而言，Platt 缩放得到的 Brier 分数最低；而 LR 和 XGBoost 未经校准的版本表现最佳。值得注意的是，LR 不采用任何后校准方法时，整体 Brier 分数最佳。

相应的校准曲线 [81] 见图 3和 4，图中展示了未经后校准以及使用 Platt 缩放进行后校准的模型。图例报告了 Brier 分数（越低越好）。虚线对角线表示完美校准。此处未展示保序缩放的校准曲线，因为该方法总体表现较差（见附录图 E3）。横轴表示预测概率，纵轴表示每个概率分箱内的经验概率（观测阳性比例）。图 4 显示，DT、RF 和 MLP 的校准效果有所改善，其可靠性曲线更接近理想对角线。

表3 不同算法的校准性能，以布里尔分数衡量。

算法	未进行后校准	进行后校准	
		Platt缩放	保序缩放
DT	0.1453	**0.1412**	0.1419
LR	**0.1023**	0.1036	0.1082
RF	0.1291	**0.1171**	0.1248
XGB	**0.1137**	0.1154	0.1204
MLP	0.1344	**0.1190**	0.1203

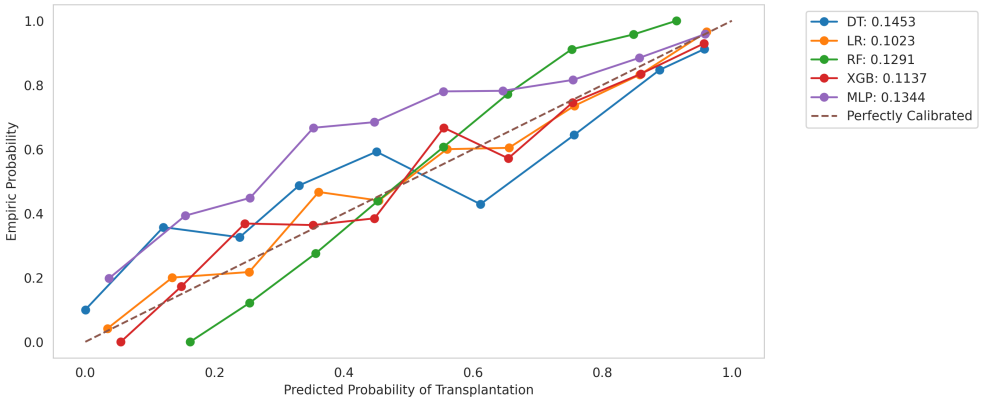

图 3 决策树、随机森林、逻辑回归、XGBoost 和多层感知机未经后校准的校准曲线。

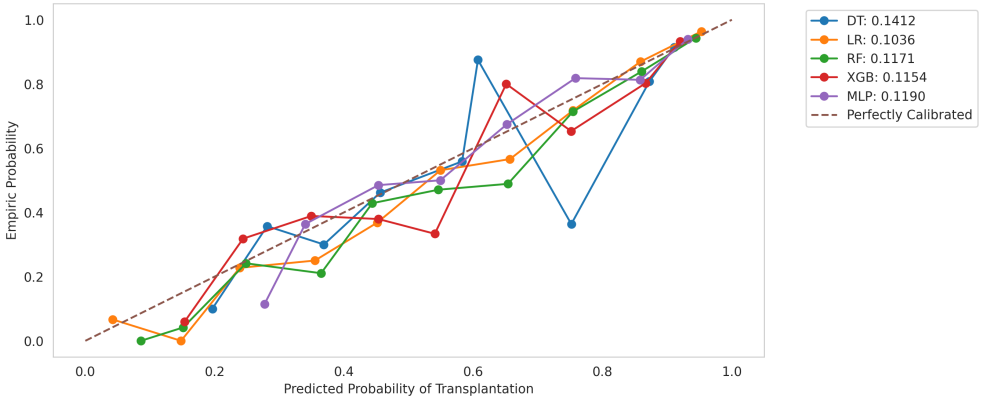

图 4 决策树、随机森林、逻辑回归、XGBoost 和多层感知机经 Platt 后校准的校准曲线。

在各模型中，Platt缩放比保序缩放能产生更可靠的概率校准。然而，是否需要后校准取决于模型：逻辑回归和XGBoost在未经校准时已具有良好校准，收益很小或没有收益，而决策树、随机森林和多层感知机从Platt缩放中获益更多。这一观察与既有认识一致，因为逻辑回归模型通常默认校准良好，其损失通过典则logit连接函数进行优化，从而使预测概率与真实观测值对齐 [82]。若未经校准，决策树和随机森林的Brier分数预计较差，因为它们可能产生过度自信的概率估计，尤其是在训练数据有限或不平衡时，从而导致预测概率与实际结果不一致 [83]。同样，多层感知机预计校准较差，最可能是由于参数数量庞大导致过拟合，因此预计校准技术会带来收益 [84]。

3.3 可解释性

详细的 SHAP 蜂群图 [68] 展示了根据背景数据预测结果得出的最重要特征。这些可视化图展示了每个样本的 SHAP 值（反映某项特征对预测结果的影响方向和幅度）、整个队列中贡献值的分布，以及贡献值如何随原始特征值变化（以颜色表示），从而凸显潜在的交互作用和离群值。图5 以逻辑回归（LR）为例说明了这一点（其他算法类型见附录F ）。

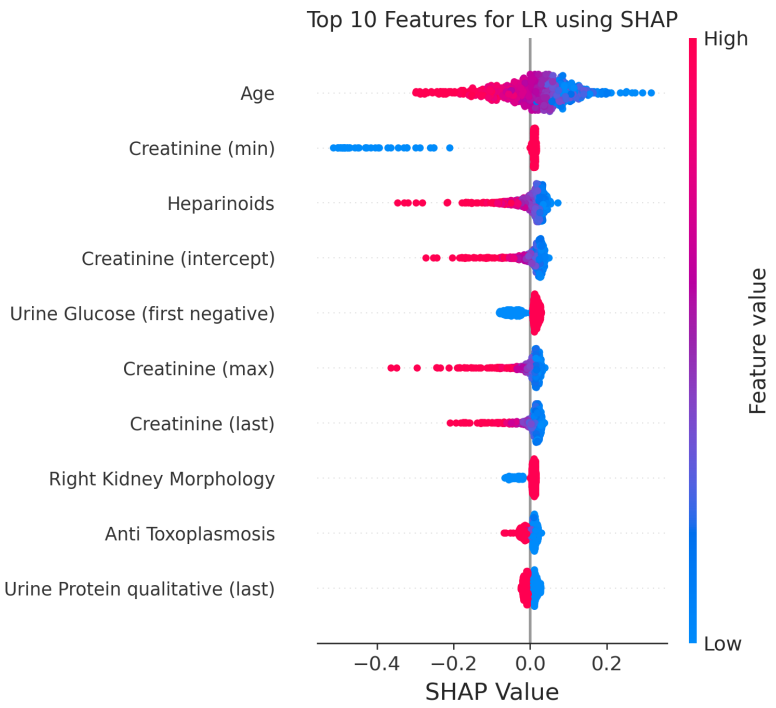

图 5 SHAP值蜂群图，展示各特征对活体亲属供体移植预测的分布及影响。每个点代表一个个体样本（测试集），其位置由SHAP值决定，颜色则表示该特征的实际值，从而展示特征影响的大小和方向。重要特征包括年龄、肾功能指标的时间序列聚合以及用药信息。

为比较六种不同模型在可解释性方面的结果，我们将图5 所示的 SHAP 值取均值，汇总为单一数值。图6 报告了每种模型中影响最大的十项特征的聚合重要性。值得注意的是，多个特征始终位列贡献最大的特征之中，其 SHAP 值幅度通常也最大，例如年龄、肝素类药物（类肝素药物），以及与肾功能相关的时间序列信息，如肌酐、尿素值和估算肾小球滤过率。不同机器学习模型始终判定为肾脏接受预测最重要的特征，与领域预期及既往研究一致。这些特征包括年龄、估算肾小球滤过率、肌酐和尿素等肾功能指标、pH 值及其他血气值，以及汇总的尿液信息 [4, 61, 85, 86]。对类肝素药物表观上较高的重要性应谨慎解读，因为在大多数情况下，该药物会作为标准捐献程序的一部分，在器官切取前立即通过静脉给药。因此，其重要性较高，部分可能反映了它与成功获取器官之间的关联，而非真正的预后效应。

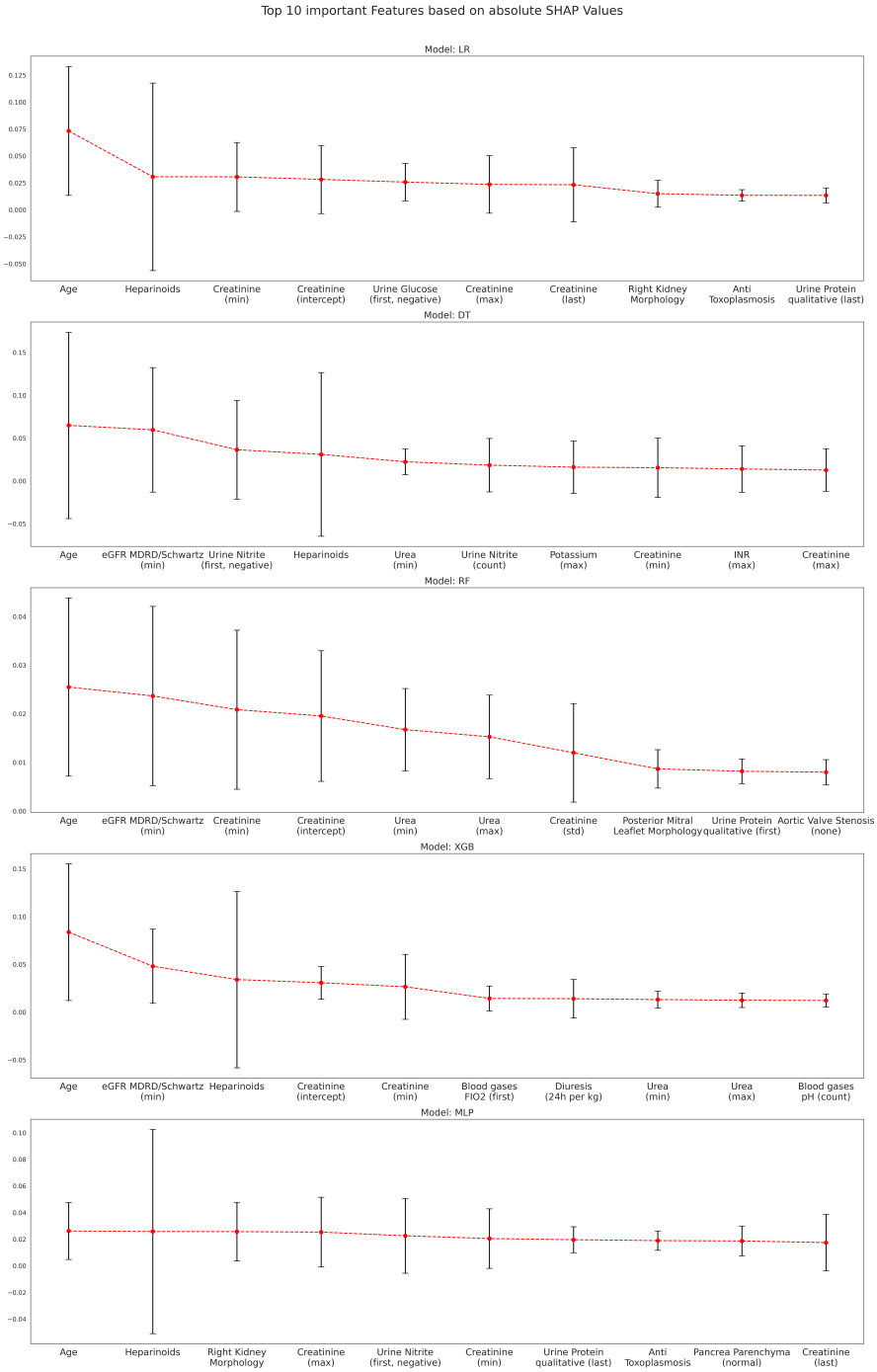

图 6 五种模型（DT、LR、RF、XGB、MLP）中按 SHAP 平均绝对值（含标准差）排名的前 10 个特征。年龄（年龄）和肾功能指标（估算肾小球滤过率、肌酐、尿素）对模型的影响最大，其他临床变量的贡献较小，且因模型而异。诸如“最小值”“数值计数”“最大值”“截距”“首个值为负”“最后一个值”“标准差”“首个定性值”等后缀，对应于实验室变量时间序列的聚合特征。20

尽管模型特定可解释性图对模型比较没有帮助，我们仍在附录中展示了这些图。附录 F。不同模型的 SHAP 全局特征重要性与模型特定可解释性大体重叠（见附录图 F4–F10）。然而，重叠程度因算法而异，两种方法得出的特征排名也有所不同，这表明共同重要特征的范围和一致性取决于底层模型。

4 讨论

我们的结果基于一套简洁且可迁移的基准测试流程。首先，我们对供体数据的数据处理和特征构建进行标准化：采用领域感知插补，从时间序列中提取特征，并对供体用药数据进行编码。其次，我们使用 NSGA-II 对每个模型进行监督式特征选择，同时优化预测性能（归一化 MCC）和简约性，并结合嵌套超参数评估以减少过拟合，随后在定制的搜索空间中进行贝叶斯 TPE 调优。第三，我们强调稳健且可比的评估：通过固定随机种子的重新训练，并报告 MCC、F1分数和曲线下面积的分布；同时采用重复测量 ANOVA 和图基 HSD 检验，确保模型比较公平。第四，在需要时，我们对概率进行校准，使用布里尔分数和可靠性图，在基于验证集预测训练的 Platt 映射和保序映射之间进行选择，然后将其应用于留出测试集。最后，我们使用 SHAP 统一不同模型家族的解释方法，提供带符号的逐样本归因和汇总后的全局重要性，同时指出其假设，并辅以模型特定诊断。

尽管深度学习在许多领域已展现出优于经典机器学习的性能 [87, 88], ，但在本研究中这一优势并不明显，逻辑回归和随机森林的表现与多层感知机相当。这一发现很可能反映了我们数据集的特点：规模适中且为表格结构，从而限制了深度学习模型充分发挥其表征能力 [28, 89]。在此类情境中，逻辑回归等较简单的模型在保持可解释性和较低计算需求的同时，仍能提供具有竞争力的预测准确率，因而对临床应用具有吸引力。

我们的结果还凸显了指标选择的重要性。F1分数和曲线下面积等常用报告指标可能会高估模型性能，因为它们要么忽略真负例，要么未能兼顾阳性预测值和阴性预测值。 [39]。因此，我们建议将马修斯相关系数作为类别不平衡的临床数据集的主要评价指标，因为它纳入了混淆矩阵的全部四个组成部分，并能提供更为均衡的判别能力度量[40]。

与先前报告认为保序回归是更优的后校准方法相反， [83]，我们的分析表明，Platt缩放取得了更好的校准效果，尤其是对于基于树的模型和神经网络模型。Brier分数的差异相对较小，表明这些模型本身已得到良好校准，这很可能归因于严格的特征选择和超参数优化。因此，我们建议有选择地应用校准，并以模型类型为指导、借助可靠性图加以验证；对于小型数据集，优先采用Platt缩放。

最后，模型可解释性仍是临床采用中的关键考量。比较模型无关方法与模型特定方法，可见二者具有互补的优势与局限。模型特定的特征重要性反映了算法偏置（例如，回归系数或基于不纯度的评分），而SHAP可在各模型家族之间提供统一的重要性尺度 [90, 91]。然而，SHAP归因值可能受特征相关性、背景数据和独立性假设的影响，且对绝对SHAP值的聚合会忽略方向性。此外，SHAP值能够反映关联，但无法区分潜在的因果机制，类肝素药物被赋予较高重要性便是一例。尽管存在这些限制，SHAP在识别一致且临床上合理的预测因子，以及在供者层面实现局部可解释性方面显示出价值，这相对于全局特征排名是一项优势 [68, 74]。尽管如此，SHAP分析应结合模型诊断与专家临床判断进行解读。

5 结论

鉴于机器学习模型种类繁多，且临床数据集存在异质性，不同机器学习模型已被用于预测供体弃用风险，但对于哪种模型表现最佳，尚无定论 [7, 14, 15]。我们基于同一数据集进行系统比较，得出了更明确的见解：结合不同模型家族的集成模型判别性能最高，但在预测推导、校准评估和可解释性方面存在局限。对于没有这些缺点的单个模型，多层感知机的表现优于基于树的集成模型（如随机森林、XGBoost）。有些出乎意料的是，简单的线性加性模型（逻辑回归）表现同样出色。这支持了可解释性良好的逻辑回归模型继续得到广泛应用 [5, 9–13], ，也与Sageshima等人 [15] 的研究结果一致；他们同样报告称，逻辑回归与其他机器学习模型的表现相近。重要的是，可解释性并非线性模型所独有，也可以通过SHAP值从更复杂的模型中提取。我们发现，SHAP在不同模型家族中都能稳定识别已知的临床驱动因素（如年龄、肾功能指标），从而支持一致的解读。正如预期，单棵决策树的表现明显较差，这可能是因为其捕捉数据复杂性的能力有限。

我们的结果基于一套简洁且可迁移的基准测试流程，该流程优先关注方法，而非单一数值结果。总而言之，我们的结果表明，与供体弃用预测的成功与否相比，方法学严谨性和数据预处理比具体选择哪种模型类型更为关键。

声明

资助。 本研究项目由德国联邦教育与研究部资助（项目编号 13FH019KX1）。本文所呈现的研究结果由作者负责。出版费用由达姆施塔特应用科学大学资助。

利益冲突/竞争性利益。 作者声明，在本研究、署名和/或本文发表方面不存在任何潜在利益冲突。

伦理审批与参与同意。 数据由德国器官获取组织（DSO）提供，作为用于二次分析的独立研究数据库。所有身份标识均已移除，且仅使用匿名化数据。根据《移植法》（《移植法》，《联邦法律公报》第一部分第2206页）第§14条和§15条，允许分析这些匿名化患者数据。根据《移植法》（《移植法》，《联邦法律公报》第一部分第2206页）第§14条，无需取得知情同意，因为研究仅使用已故器官供体的回顾性匿名化数据。根据黑森州医学会《职业行为准则》第§15条，无需取得医学伦理审批，因为研究仅使用已故器官供体的回顾性匿名化数据。本研究遵循《赫尔辛基宣言》开展。

发表同意。 不适用

数据可用性。 本文所述分析基于德国器官获取组织（DSO）提供的回顾性匿名化数据。由于数据保护和机构限制，这些数据无法公开。

材料可用性。不适用

代码可用性。所采用分析方法的实现代码可在以下网址获取：
https://github.com/PeerHan/donor-kidney-discard-prediction。

作者贡献。Axel Rahmel、Antje Jahn 和 Gunter Grieser 构思并设计了这项研究，并制定了分析计划；Peer Schliephacke、Hannah Schult、Leon Mizera 和 Judith W¨urfel 设计了特征工程并完成了所有分析，Carl-Ludwig Fischer-Fr¨ohlich 提供医学指导。Peer Schliephacke 起草了论文初稿。所有作者均审阅、编辑并批准了论文终稿。

致谢。我们衷心感谢 Thomas La Rocca（DSO）、Lukas Klein、Adrian F¨uller、Anika Fuchs、Danilo Z¨ahle、David Heiß、FriedrichHeitzer、Katharina Litzinger、Kevin Saliu、Leonard Holdau 和 Tristan Funk（h da）对本工作的支持与贡献。

参考文献

[1]Organtransplantation, D.S.：《年度报告：德国器官捐献与移植（2023年）。https://dso.de/SiteCollectionDocuments/DSO-Jahresbericht%202023.pdf 访问日期：2025年3月19日

[2]Yoon, J., Alaa, A.M., Cadeiras, M., Schaar, M.v.d.：《器官移植中的个性化供体-受体匹配》。arXiv。arXiv:1611.03934 [cs]（2016年）。https://doi.org/10.48550/arXiv.1611.03934 。http://arxiv.org/abs/1611.03934访问日期：2025年3月20日

[3]Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., Khovanova, N.：用于抗体不相容肾移植结局预测的决策树和随机森林模型。《生物医学信号处理与控制》**52**，456–462（2019年）

[4]Yoo, D., Divard, G., Raynaud, M., Cohen, A., Mone, T.D., Rosenthal, J.T., Bentall, A.J., Stegall, M.D., Naesens, M., Zhang, H., Wang, C., Gueguen, J., Kamar, N., Bouquegneau, A., Batal, I., Coley, S.M., Gill, J.S., Oppenheimer, F., De Sousa-Amorim, E., Kuypers, D.R.J., Durrbach, A., Seron, D., Rabant, M., Van Huyen, J.-P.D., Campbell, P., Shojai, S., Mengel, M., Bestard, O., Basic-Jukic, N., Juri´c, I., Boor, P., Cornell, L.D., Alexander, M.P., Toby Coates, P., Legendre, C., Reese, P.P., Lefaucheur, C., Aubert, O., Loupy, A.: 面向肾移植患者的机器学习驱动的虚拟活检系统。《自然·通讯》15(1), 554 (2024年) https://doi.org/10.1038/s41467-023-44595-z 。出版者：自然出版集团。访问日期：2025年2月16日[5] McKenney, C., Torabi, J., Todd, R., Akhtar, M.Z., Tedla, F. M., Shapiro, R., Florman, S.S., Holzner, M.L., van Leeuwen, L.L.: 被浪费的潜力：解读供肾短缺、利用不足与不断上升的弃用率三重困境。《移植学》5(2), 51–64 (2024年) https://doi.org/10.3390/transplantology5020006[6] Gotlieb, N., Azhie, A., Sharma, D., Spann, A., Suo, N.-J., Tran, J., Orchanian- Cheff, A., Wang, B., Goldenberg, A., Chass´e, M., 等: 机器学习应用在实体器官移植中的前景。《NPJ数字医学》5(1), 89 (2022年)[7]Pettit, R.W., Marlatt, B.B., Miles, T.J., Uzgoren, S., Corr, S.J., Shetty, A., Havelka, J., Rana, A.: 机器学习在预测腹部器官移植中供体弃用方面的效用。《临床移植》37(5), 14951 (2023年) https://doi.org/10.1111/ctr.14951[8] Cucchetti, A., Vivarelli, M., Heaton, N.D., Phillips, S., Piscaglia, F., Bolondi, L., La Barba, G., Foxton, M.R., Rela, M., O' Grady, J., Pinna, A.D.: 人工神经网络在预测终末期肝病患者死亡率方面优于MELD评分。《肠道》56(2), 253–258 (2007年) https://doi.org/10.1136/gut.2005.084434[9] Zhou, S., Massie, A.B., Holscher, C.M., Waldram, M.M., Ishaque, T., Thomas, A.G., Segev, D.L.: 肾脏弃用预测模型的前瞻性验证。《移植》103(4), 764–771 (2018年) https://doi.org/10.1097/tp.0000000000002362[10] Massie, A.B., Desai, N.M., Montgomery, R.A., Singer, A.L., Segev, D.L.: 提高难以分配的逝者供肾的分配效率：预测弃用或延迟概率。《美国移植杂志》10(7), 1613–1620 (2010年) https://doi.org/10.1111/j.1600-6143.2010.03163.x[11]Marrero, W.J., Naik, A.S., Friedewald, J.J., Xu, Y., Hutton, D.W., Lavieri,

M.S., Parikh, N.D.: 美国逝者供体肾脏弃用的预测因子。《移植》**101**(7), 1690–1697 (2016年) https://doi.org/10.1097/tp.0000000000001238

[12] Cohen, J.B., Shults, J., Goldberg, D.S., Abt, P.L., Sawinski, D.L., Reese, P.P.: 肾脏同种异体移植器官供体提议：拒绝的预测因子以及延迟器官接受对同种异体移植物存活的影响。《美国移植杂志》**18**(2), 391– 401(2017年) https://doi.org/10.1111/ajt.14449

[13] Narvaez, J.R.F., Nie, J., Noyes, K., Leeman, M., Kayler, L.K.: 难以匹配的肾脏供体提议：弃用的供体层面与系统层面预测因素。《美国移植杂志》 **18**(11), 2708–2718 (2018年) https://doi.org/10.1111/ajt.14712

[14]Barah, M., Mehrotra, S.: 使用机器学习预测肾脏弃用。《移植》**105**(9), 2054–2071 (2021年) https://doi.org/10.1097/tp.0000000000003620[15] Sageshima, J., Than, P., Goussous, N., Mineyev, N., Perez, R.: 利用结构化供体特征和非结构化供体叙述预测高风险供体的肾脏弃用和未恢复使用。《JAMA外科》**159**(1), 60 (2023年) https://doi.org/10.1001/jamasurg.2023.4679[16] B¨uy¨ukkececi, M., Okur, M.C.: 机器学习中特征选择及其稳定性的综合综述。《加齐大学科学学报》**36**(4), 1506–1520 (2023年)[17] WHO, W.H.O.: ICD-10: 国际疾病和相关健康问题统计分类第10次修订版, 第1卷, 第5版。世界卫生组织（WHO）。https://doi.org/10.1038/s41598-024-66976-0 。 https://apps.who.int/iris/bitstream/10665/246208/1/9789241549165-V1-eng.pdf[18] Sauthier, N.,Bouchakri, R., Carrier, F.M., Sauthier, M., Mullie, L.-A., Cardinal,H., Fortin, M.-C., Lahrichi, N., Chass´e, M.: 利用时序机器学习模型自动筛查潜在器官供体。《科学报告》（2023年）https://doi.org/10.1038/s41598-023-35270-w[19] Mohan, S., Husain, S.A., Schold, J.D., Reese, P.P., Stewart, D., Kadatz, M., Chow, D.S.,Khurana, K.K., Axelrod, D., Mulligan, D.C., Formica, R.N., Roberts, J.P., Segev, D.L., Locke, J.E., Rees, M.J., Matas, A., Stegall, M.L., Cooper, M., Stock, P.G., Ellis, M.J., Heeger, P.S., Cohen, D.J., Danovitch, G.M., Mont- gomery, R.A., Bromberg, J.S., Redfield, R.R., Gaston, R.A., Gill, J., Kasiske, B.L., Kaplan, B.: 用于肾脏移植的机器学习驱动虚拟活检系统。《自然·通讯》（2024年） https://doi.org/10.1038/s41467-023-44595-z[20] Scikit-learn: IterativeImputer。（2024年）。访问日期: 2025-09-26。 htt ps://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

[21]Scikit-learn：HistGradientBoostingClassifier。（2024年）。访问日期：2025-03- 29。
h
t
t
p
s://
scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html[22] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.：《分类与回归树》。Chapman & Hall/CRC，博卡拉顿、伦敦、纽约和华盛顿哥伦比亚特区（1984年）。
https://doi.org/10.1201/9781315139470。https://www.taylorfrancis.com/books/9781351460491[23] Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.：《应用逻辑回归》。John Wiley & Sons[24] Breiman, L.：随机森林。《机器学习》**45**(1), 5–32（2001年）
https://doi.org/10.1023/A:1010933404324[25] Chen, T., Guestrin, C.：XGBoost。载于：Krishnapuram, B.（编）《第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集》。ACM数字图书馆，第785–794页。ACM，纽约州纽约市（2016年）。
https://doi.org/10.1145/2939672.2939785[26] Freund, Y., Schapire, R.E.：在线学习的决策理论泛化及其在提升法中的应用。《计算机与系统科学杂志》**55**(1), 119–139（1997年）[27] Wu, J., Li, Y., Ma, Y.：类别平衡数据集上XGBoost与神经网络模型的比较。载于：2021年IEEE第三届信息与计算机前沿技术国际会议，第457–461页。IEEE，皮斯卡塔韦，新泽西州（2021年）。
https://doi.org/10.1109/ICFTIC54370.2021.9647373[28] Grinsztajn, L., Oyallon, E., Varoquaux, G.：为什么基于树的模型在典型表格数据上的表现仍优于深度学习？《神经信息处理系统进展》**35**，507–520（2022年）[29] Goodfellow, I., Bengio, Y., Courville, A.：《深度学习》。MIT出版社，???（2016年）。http://www.deeplearningbook.org[30] Holland, J.H.：遗传算法。《科学美国人》**267**(1), 66–73（1992年）[31]Taha, Z.Y., Abdullah, A.A., Rashid, T.A.：利用遗传算法优化特征选择：方法与应用综述。《知识与信息系统》，1–40（2025年）[32] Liu, H., Setiono, R.：特征选择与分类——一种概率包装器方法。载于：《工业与工程中的人工智能及专家系统应用》，第419–424页。CRC出版社

[33]Ji, S., Carin, L.：成本　-敏感特征获取与分类。《Pattern》

模式识别 **40**(5), 1474–1485 (2007)

[34] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.：快速且精英的多目标遗传算法：NSGA-II。《IEEE 进化计算汇刊》**6**(2),182–197 (2002)[35] Hamano, R., Saito, S., Nomura, M., Shirakawa, S.：带边界的 CMA-ES：为混合整数黑盒优化设定边缘概率下限。载于：遗传与进化计算会议论文集，第639–647页（2022）[36]Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.：Optuna：新一代超参数优化框架。载于：第25届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第2623–2631页（2019）[37] Bergstra, J., Bengio, Y.：用于超参数优化的随机搜索。《机器学习研究期刊》**13**(1), 281–305 (2012)[38] Matthews, B.W.：T4 噬菌体溶菌酶预测与观测到的二级结构比较。《生物化学与生物物理学报》（BBA）—蛋白质结构 **405**(2),442–451 (1975) https://doi.org/10.1016/0005-2795(75)90109-9。访问日期：2025年2月16日[39]Chicco, D., Jurman, G.：马修斯相关系数（MCC）应取代 ROC 曲线下面积，成为评估二分类的标准指标。《生物数据挖掘》**16(1)**, 4 (2023) https://doi.org/10.1186/s13040-023-00322-4 。访问日期：2025年2月16日[40] Chicco, D., Jurman, G.：在二分类评估中，马修斯相关系数（MCC）相较于 F1分数和准确率的优势。《BMC 基因组学》**21(1)**, 6 (2020)https://doi.org/10.1186/s12864-019-6413-7 。访问日期：2025年2月16日[41] Atallah, D.M., Badawy, M., El-Sayed, A.：采用改进的 k近邻进行肾移植预测的智能特征选择。《SN Applied Sciences》**1**(10), 1297 (2019)[42]Chen, Y., Gao, J., Wu, J.：通过强化学习在医疗预测性监测中进行动态特征选择。arXiv 预印本 arXiv:2405.19729 (2024)[43] Bergstra, J., Bardenet, R., Bengio, Y., K´egl, B.：超参数优化算法。《神经信息处理系统进展》**24** (2011)[44] Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z., Guyon, I.：贝叶斯优化在机器学习超参数调优方面优于随机搜索：2020年黑盒优化挑战赛分析。载于：

神经信息处理系统大会 2020年竞赛与演示专题，第 3–26 页（2021年）。PMLR

[45] Hollmann, N., M¨uller, S., Eggensperger, K., Hutter, F.：TabPFN：一种能在一秒内解决小型表格分类问题的Transformer。arXiv预印本 arXiv:2207.01848（2022年）[46] Bardes, A., Ponce, J., LeCun, Y.：VICReg：用于自监督学习的方差-不变性-协方差正则化。arXiv预印本 arXiv:2105.04906（2021年）[47]Gorishniy, Y.V., Rubachev, I., Khrulkov, V., Babenko, A.：重新审视用于表格数据的深度学习模型。arXiv **abs/2106.11959**（2021年）[48]Defazio, A., Bach, F., Lacoste-Julien, S.：SAGA：一种支持非强凸复合目标函数的快速增量梯度法。《神经信息处理系统进展》**27**（2014）[49] Breiman, L.：装袋预测器。《机器学习》**24(2)**，123–140（1996）[50] Glorot, X., Bengio, Y.：理解深度前馈神经网络训练的困难。载于：第十三届国际人工智能与统计学会议论文集，第249–256页（2010）。JMLR研讨会与会议论文集[51]He, K., Zhang, X., Ren, S., Sun, J.：深入探索整流器：在ImageNet分类中超越人类水平的表现。载于：IEEE国际计算机视觉会议论文集，第1026–1034页（2015）[52] Desai, C., Desai, C.：权重初始化技术对神经网络效率和性能的影响：以MNIST数据集为例。国际工程与计算机科学杂志**13(04)**（2024年）[53] Christen, P., Hand, D.J., Kirielle, N.：F度量综述：其历史、性质、批评及替代方案。ACM计算调查**56(3)**，1–24（2023年）[54] Hanley, J.A., McNeil, B.J.：受试者工作特征（ROC）曲线下面积的含义与用途。《放射学》**143(1)**，29–36（1982）[55] Hanczar, B., Hua, J., Sima,C., Weinstein, J., Bittner, M., Dougherty, E.R.：ROC相关估计的小样本精确率。《生物信息学》**26(6)**，822–830（2010）[56] Lobo, J.M., Jim´enez-Valverde, A., Real, R.：曲线下面积：预测分布模型性能的一种误导性度量。《全球生态学与生物地理学》**17(2)**，145–151（2008）[57] Girden, E.R.：ANOVA：重复测量，第84卷。Sage

[58] Nanda, A., Mohapatra, B.B., Mahapatra, A.P.K., Mahapatra, A.P.K., Mahapa- tra, A.P.K.：Tukey诚实显著差异（HSD）多重比较检验：置信水平是否控制I类错误。《国际统计与应用数学杂志》6(1)，59–65（2021年）[59] Haller, M.C., Aschauer, C., Wallisch, C., Leffondr´e, K., Smeden, M., Oberbauer, R., Heinze, G.：活体器官移植预测模型的开发、报告和验证均不充分：系统综述。《临床流行病学杂志》145，126–135（2022年）[60] Ali, H., Shroff, A., F¨ul¨op, T., Molnar, M.Z., Sharif, A., Burke, B., Shroff, S., Briggs, D., Krishnan, N.：人工智能辅助器官移植风险预测：英国活体供肾移植结局预测工具。《肾衰竭》47(1)，2431147（2025年）[61]Sala¨un, A., Knight, S., Wingfield, L., Zhu, T.：使用可解释机器学习模型预测肾移植后的移植物和患者结局。《科学报告》14(1)，17356（2024年）[62]Brier, G.W., Allen, R.A.：天气预报的验证。载于：《气象学概要：在气象学概要委员会指导下编写》，第841–848页。施普林格[63] Murphy, A.H.：概率评分的一种新向量分解。《应用气象学与气候学杂志》12(4)，595–600（1973）[64] Ojeda, F.M., Jansen, M.L., Thi´ery, A., Blankenberg, S., Weimar, C., Schmid, M., Ziegler, A.：校准概率估计的机器学习方法：全面比较。《医学统计学》42(29)，5451–5478（2023年）https://doi.org/10.1002/sim.9921[65] Rufibach, K.：使用布里尔分数评估二元预测。《临床流行病学杂志》63(8)，938–939（2010）[66] Platt, J., et al.：支持向量机的概率输出及与正则化似然方法的比较。《大间隔分类器进展》10(3)，61–74（1999）[67] Zadrozny, B., Elkan, C.：将分类器评分转换为准确的多分类概率估计。载于：第八届ACM SIGKDD国际知识发现与数据挖掘会议论文集，第694–699页（2002）[68] Lundberg, S., Lee, S.-I.：解释模型预测的统一方法。arXiv。arXiv:1705.07874 [cs]（2017年）。https://doi.org/10.48550/arXiv.1705.07874 。http://arxiv.org/abs/1705.07874 访问日期：2025-02-20[69] Shapley, L.S.：n人博弈中的一个值。《博弈论贡献》

[70] Zheng, Q., Wang, Z., Zhou, J., Lu, J.: Shap-CAM：基于Shapley值的卷积神经网络视觉解释。在：欧洲计算机视觉会议，第459–474页（2022年）。施普林格[71] Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., Groh, G.: 基于SHAP的解释方法：自然语言处理可解释性综述。在：第29届国际计算语言学会议论文集，第4593–4603页（2022年）[72]Rodr´ıguez-P´erez, R., Bajorath, J.: 使用Shapley值解释机器学习模型：在化合物效力和多靶点活性预测中的应用。《计算机辅助分子设计杂志》34(10)，1013–1026（2020年）[73]Saarela, M., Jauhiainen, S.: 特征重要性度量作为分类模型解释的比较。SN Applied Sciences 3(2)，272（2021年）[74] Saeed, W., Omlin, C.: 可解释人工智能（XAI）：当前挑战与未来机遇的系统性元综述。基于知识的系统 263，110273（2023年） https://doi.org/10.1016/j.knosys.2023.110273。访问日期：2025年3月19日[75] Scholbeck, C.A., Molnar, C., Heumann, C., Bischl, B., Casalicchio, G.: 采样、干预、预测与聚合：一种模型无关解释的通用框架。在：机器学习与数据库知识发现：ECML PKDD 2019国际研讨会，德国维尔茨堡，2019年9月16–20日，论文集，第一部分，第205–216页（2020年）。施普林格[76]Decruyenaere, A., Decruyenaere, P., Peeters, P., Vermassen, F., Dhaene, T., Couckuyt, I.: 肾移植后移植肾功能延迟恢复的预测：逻辑回归与机器学习方法的比较。《BMC医学信息学与决策制定》15(1)，83（2015年）[77]Esmaily, H., Tayefi, M., Doosti, H., Ghayour-Mobarhan, M., Nezami, H., Amirabadizadeh, A.: 决策树与随机森林在确定2型糖尿病相关风险因素方面的比较。《健康科学研究杂志》18(2)，412（2018年）[78] Slonim, D.K.: 从模式到通路：基因表达数据分析日渐成熟。《自然-遗传学》32(4)，502–508（2002年）[79] Tolstyak, Y., Zhuk, R., Yakovlev, I., Shakhovska, N., Gregus ml, M., Chopyak, V., Melnykova, N.: 用于预测肾移植后生存情况的机器学习方法集成。《应用科学》11(21)，10380（2021年） https://doi.org/10.3390/app112110380。编号：21；出版者：多学科数字出版机构。访问日期：2025年2月16日

[80] Reeve, J.,B¨ohmig, G.A., Eskandary, F., Einecke, G., Gupta, G., Madill-Thomsen, K., Mackova, M., Halloran, P.F., INTERCOMEX MMDx-Kidney研究组：结合分子测量与机器学习分类器集成模型生成自动化肾移植活检报告。《美国移植杂志》：美国移植学会与美国移植外科医师学会官方杂志 **19(10)**，2719–2731（2019年）https://doi.org/10.1111/ajt.15351[81] Wilks, D.S.: 连续降水时段预测概率的组合。《天气与预报》**5(4)**，640–650（1990年）[82] W¨uthrich, M.V., Merz, M.:《精算学习及其应用的统计基础》。施普林格[83] Niculescu-Mizil, A., Caruana, R.: 使用监督学习预测良好概率。在：Dˇzeroski, S.（编）《第22届国际机器学习会议论文集》。ACM其他会议，第625–632页。ACM，纽约州纽约市（2005年）。https://doi.org/10.1145/1102351.1102430[84] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: 现代神经网络的校准问题。在：国际机器学习会议，第1321–1330页（2017年）。PMLR[85] Sageshima, J., Than, P., Goussous, N., Mineyev, N., Perez, R.: 利用结构化供体特征和非结构化供体叙述预测高风险供体的肾脏弃用和未恢复使用。《JAMA外科》**159(1)**，60–68（2024年）[86] Sauthier, N., Bouchakri, R., Carrier, F.M., Sauthier, M., Mullie, L.-A., Cardinal, H., Fortin, M.-C., Lahrichi, N., Chass´e, M.: 使用时序机器学习模型自动筛查潜在器官供体。《科学报告》**13(1)**，8459（2023年）[87] Shiri, F.M., Perumal, T., Mustapha, N., Mohamed, R.: 深度学习模型全面综述与比较分析：卷积神经网络、循环神经网络、长短期记忆网络、门控循环单元。arXiv预印本 arXiv:2305.17473（2023年）[88] Noor, M.H.M., Ige, A.O.: 深度学习应用现状与挑战综述。《人工智能工程应用》**159**，111225（2025年）[89] Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: 深度神经网络与表格数据：综述。《IEEE神经网络与学习系统汇刊》**35(6)**，7499–7519（2022年）[90] Huang, X., Marques-Silva, J.: Shapley值在可解释性方面的不足。《国际近似推理杂志》**171**，109112（2024年）

[91] Kumar, I., Scheidegger, C., Venkatasubramanian, S., Friedler, S.：Shapley 残差：量化 Shapley 值用于解释的局限性。神经信息处理系统进展**34**，26598–26608（2021年）

[92] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer,P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.：Scikit-learn：Python 中的机器学习。机器学习研究期刊**12**，2825–2830（2011）

[93] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L.,等：PyTorch：一种命令式风格的高性能深度学习库。神经信息处理系统进展**32** （2019年）

补充信息。

附录 A 软件与硬件

所有使用 sklearn 训练的模型均采用 AMD EPYC-Milan处理器 [92]版本 1.5.2（决策树、随机森林、逻辑回归）和 xgboost 版本 2.1.2。我们使用 pytorch [93] 版本 2.5.0 和 CUDA 版本 12.2，在 NVIDIA A100显卡上训练多层感知机模型。我们使用 shap 软件包 [68] 版本 0.46.0 实现可解释性。

附录 B 特征工程

B.1 特征变换

• **糖尿病诊断时年龄**供体的年龄通过其出生时间戳和糖尿病诊断日期计算得出。

• **糖尿病持续时间** 供体的糖尿病持续时间根据诊断时间戳和死亡时间计算。

• **饮酒持续时间**此特征的计算取决于是否提供了相应供体的结束日期。如果有结束日期，则饮酒持续时间（以天为单位）按开始日期与结束日期之差计算。如果未提供结束日期，则假设住院日期为该供体的终止日期。

• **末次饮酒** 此特征以类别形式表示距供体末次饮酒所经过的时间。类别划分依据记录的酒精使用结束日期与住院日期之间的差值。计算方法如下：

– 未记录饮酒情况：如果未提供饮酒开始日期，则假定供体不存在与酒精相关的问题，并归入类别 0。

– 目前仍在饮酒：如果有开始日期但未提供结束日期，则假定供体一直饮酒至住院。在这种情况下，时间差为 0 天，供体归入最高类别，即类别 5。

– 入院前已停止饮酒：如果有饮酒开始和结束日期，则计算停止酒精使用至住院之间的天数。随后将这一时间跨度分为四个类别。分类区间根据训练数据集中的分布动态确定；较近期停止饮酒的供体归入较高类别（例如类别 4），而最后一次饮酒距今更久的供体则归入较低类别（例如类别 1）。

• 对于一些取值罕见的分类变量，我们将其归并为正向和负向，以将它们转化为二元特征。例如，“心电图QRS波差异结果”被二分为“否”（未报告差异）和“是”（报告为“其他”“心肌梗死样”“右束支传导阻滞”“左束支传导阻滞”或“双分支传导阻滞”）。

• 最后一小时的利尿量（dlh）使用体重（bw）进行了标准化（见公式B1）。

• 在时间窗口（dt）内的利尿量总量（t）按比例换算为24小时的数值，并根据体重进行标准化（见公式B2）

• 我们将肌酐值转换为毫克/分升，方法是将这些值乘以系数0.011312（该系数见主数据库中的 $\mu mol/L$ ）。

$$dlh_{norm} = \frac{dlh}{bw} \tag{B1}$$

$$d_{24h/bw} = \frac{dt}{t\ (in\ hours)} \cdot \frac{24}{bw} \tag{B2}$$

附录C 超参数结果

C.1 超参数空间

	mean	std	min	25%	50%	75%	max
最大深度	17.19667	11.33016	1	8	15	24	49
叶节点最小样本数	13.67667	4.15246	1	12	14	16	20
最小分裂样本数	9.70333	5.67120	2	5	8	15	20

表 C1 决策树超参数搜索的汇总统计：300 次试验中最大深度、最小叶节点样本数和最小分裂样本数的均值、标准差、最小值和最大值。

	mean	std	min	25%	50%	75%	max
C	12.03738	20.69000	0.01497	0.59561	2.10518	15.78112	97.24803
L1比率	0.77262	0.20096	0.00505	0.70081	0.81645	0.91059	0.99963

表 C2 逻辑回归超参数搜索的汇总统计：300 次试验中 C 和 L1比率的均值、标准差、最小值和最大值。

	mean	std	min	25%	50%	75%	max
最大深度	28.70333	11.99032	1	20	30	37	50
叶节点最小样本数	7.69667	4.11352	1	5	7	9	20
最小分裂样本数	15.90667	5.05931	2	15	18	20	20
估计器数量	393.70333	130.46514	10	367.75000	445.50000	484	500

表 C3 随机森林超参数搜索的汇总统计：300 次试验中估计器数量、最大深度、最小叶节点样本数和最小分裂样本数的均值、标准差、最小值和最大值。

	mean	std	min	25%	50%	75%	max
按树列采样	0.80872	0.05063	0.75021	0.76957	0.79353	0.83285	0.99825
早停轮数	44.51667	27.23028	5	20	45	65	100
学习率	0.02292	0.02188	0.00108	0.00856	0.01462	0.03076	0.09703
最大深度	4.74000	2.92591	2	3	4	6	15
最小子节点权重	11.76333	5.77626	1	6	13	17	20
估计器数量	1013.66667	414.85367	100	700	1150	1350	1500
L1正则化系数	0.54814	2.13458	0.00103	0.00491	0.02079	0.13249	19.89581
L2正则化系数	0.87691	3.01173	0.00102	0.00361	0.02605	0.21597	22.85764
子采样	0.90318	0.07581	0.75181	0.84068	0.92029	0.97013	0.99973

表 C4 XGBoost 超参数搜索汇总统计：300 次试验中 colsamply bytree、早停轮数、学习率、最大深度、最小子节点权重、估计器数量、L1/L2正则化系数以及子采样的均值、标准差、最小值和最大值。

	mean	std	min	25%	50%	75%	max
批归一化	0.766667	0.423659	0	1	1	1	1
丢弃法	0.367333	0.113398	0	0.300000	0.400000	0.450000	0.500000
隐藏维度	838.613333	350.107032	1010	5800	8580	1065.250000	15000
初始学习率	0.002675	0.010163	0.000100	0.000145	0.000255	0.000677	0.090003
层数	4.676667	2.948633	20	30	40	60	150
权重衰减	$1.009 \times e^{-7}$	$2.1 \times e^{-7}$	$1.002 \times e^{-10}$	$2.34 \times e^{-10}$	$1.9 \times e^{-9}$	$7.02 \times e^{-8}$	$9.91 \times e^{-7}$

表 C5 多层感知机超参数搜索的汇总统计：300 次试验中批归一化（使用或不使用）、丢弃概率、隐藏维度、初始学习率和权重衰减的均值、标准差、最小值和最大值。

超参数	值
最大深度	7
最大特征数	sqrt
叶节点最小样本数	18
最小分裂样本数	15

表 C6 超参数搜索得出的最佳决策树配置

超参数	值
C	0.07085
L1比率	0.88392

表 C7 超参数搜索得出的最佳逻辑回归配置

超参数	值
最大深度	7
最大特征数	sqrt
叶节点最小样本数	18
最小分裂样本数	15

表C8 超参数搜索得到的最佳随机森林配置

超参数	值
按树列采样	0.80955
早停轮数	75
学习率	0.01551
最大深度	3
最小子节点权重	4
估计器数量	1300
L1正则化系数	0.01442
L2正则化系数	2.46270
子采样	0.99142
树方法	hist

表 C9 超参数搜索得到的最佳 XGBoost 配置

超参数	值
激活函数	ELU
批归一化	1
类别权重	0
丢弃法	0.35000
隐藏维度	1121
初始学习率	0.00012
层数	3
权重衰减	$1 \times e^{-10}$

表 C1 通过超参数搜索得到的最佳多层感知机配置

C.2 最终架构

附录 D 进一步的 Tukey 结果

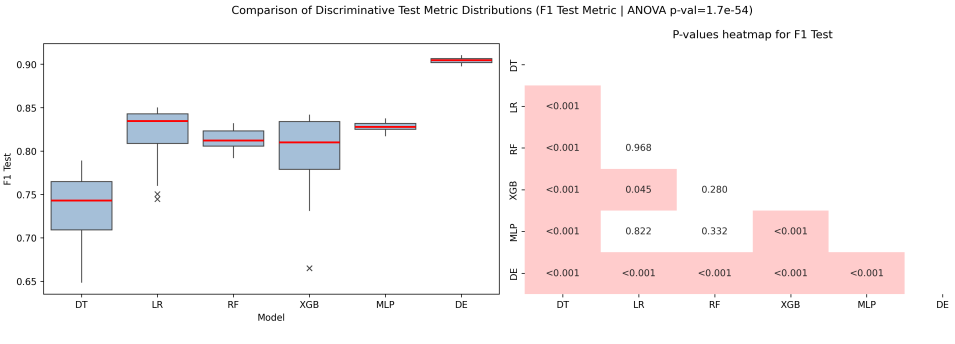

图 D1 左：对六种不同机器学习模型进行判别性能评估。每个面板报告多种方法的测试 F1分数（Y 轴，越高越好）：XGBoost、随机森林、多层感知机、逻辑回归、决策树和差分进化。箱线图展示每种方法的性能分布。**右：**Tukey 校正的 p 值下三角热图，展示模型之间的两两统计显著性。

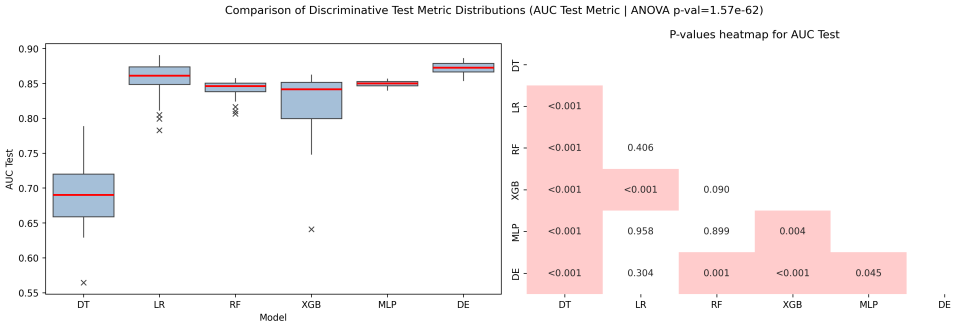

图 D2 左：六种不同机器学习模型的判别性能评估。每个面板展示多种方法的测试集曲线下面积（纵轴，越高越好）：XGB、RF、MLP、LR、DT 和 DE。箱线图展示了各方法的性能分布。**右：**显示模型间两两统计显著性的 Tukey 校正 p 值下三角热图。

附录 E 校准

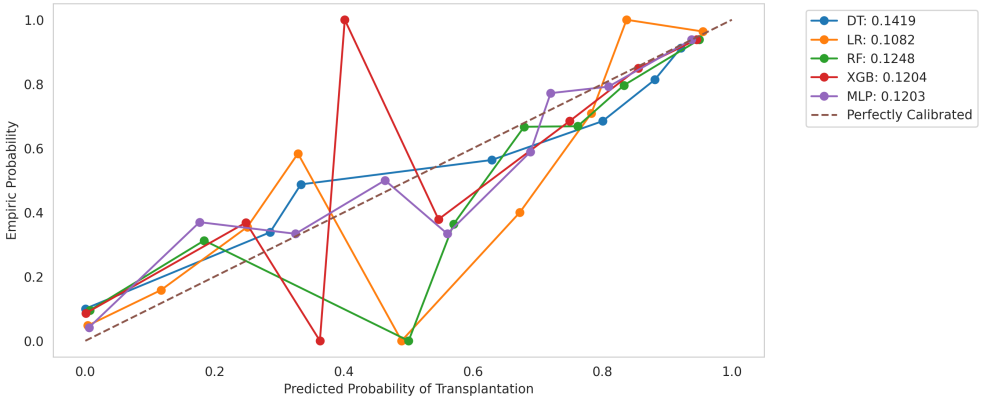

图 E3 DT、RF、LR、XGB 和 MLP 的保序校准曲线。

附录 F 可解释性

F.1 逻辑回归

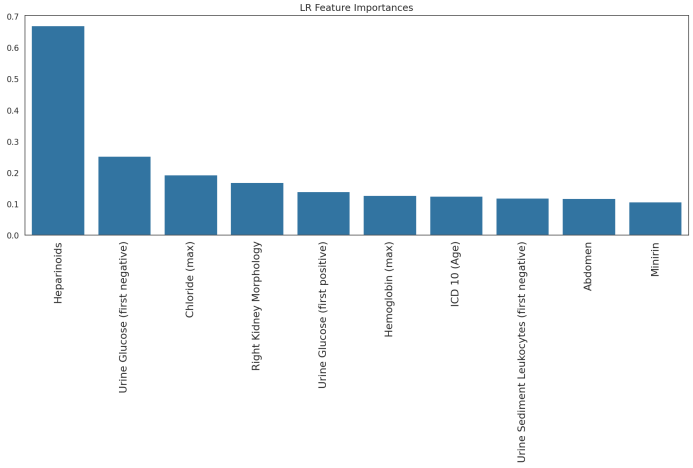

图 F4 通过逻辑回归系数（纵轴）实现模型特定可解释性，突出显示肾功能指标的时间序列聚合和用药信息等重要特征。

F.2 决策树

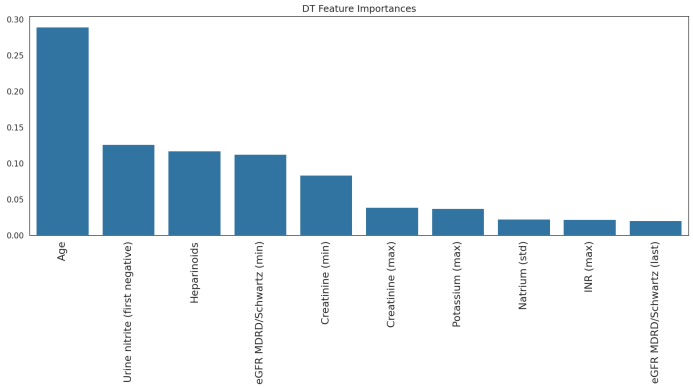

图 F5 通过决策树特征重要性（纵轴）实现模型特定可解释性，突出显示肾功能指标的时间序列聚合和用药信息等重要特征。

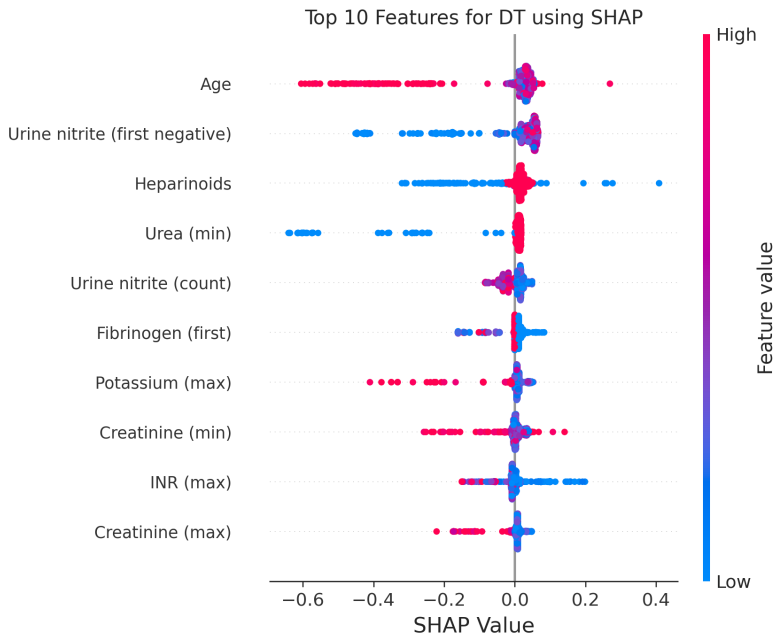

图 F6 SHAP值蜂群图展示了特征在DT供体移植预测中的分布和影响。每个点代表一个独立样本（测试集），其位置由SHAP值决定，颜色则表示特征的实际值，展示了特征影响的大小和方向。排名靠前的特征包括年龄、肾功能指标的时间序列聚合以及用药信息。

F.3 随机森林

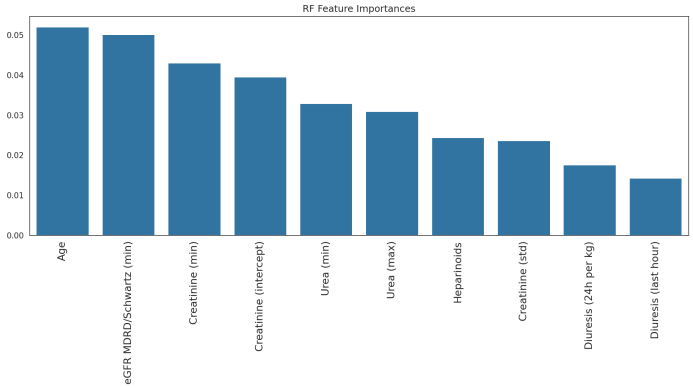

图 F7 通过RF特征重要性（纵轴）进行模型特定可解释性分析，突出显示了年龄、肾功能指标的时间序列聚合以及用药信息等重要特征。

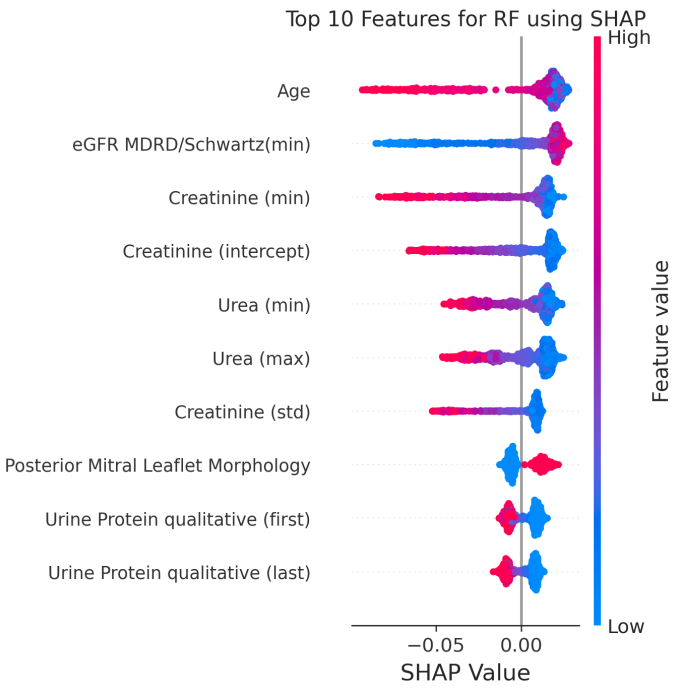

图 F8 SHAP值蜂群图，展示了特征对随机森林供体移植预测的分布及影响。每个点代表一个独立样本（测试集），其位置由 SHAP值决定，颜色表示特征的实际值，体现了特征影响的大小和方向。排名靠前的特征包括年龄、肾功能指标的时间序列聚合以及用药信息。

F.4 XGBoost

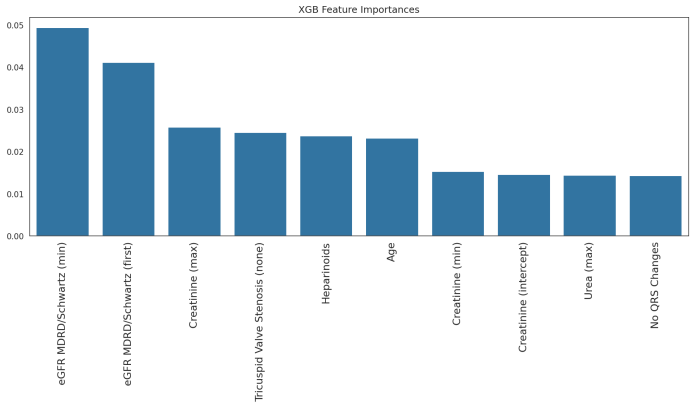

图 F9 通过 XGBoost 特征重要性（y 轴）进行模型特定可解释性分析，突出显示了年龄、肾功能指标的时间序列聚合以及用药信息等重要特征。

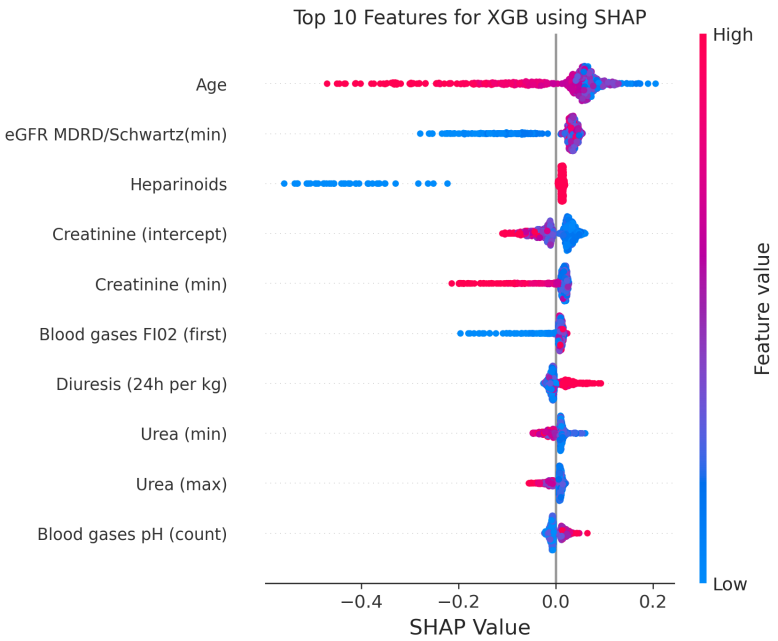

图 F10 展示 SHAP 值的蜂群图，显示特征对 XGBoost 供体移植预测的分布和影响。每个点代表一个单独的实例（测试集），其位置由 SHAP 值决定，颜色则表示特征的实际值，从而同时展示特征影响的大小和方向。排名靠前的特征包括年龄、肾功能测量值的时间序列聚合以及用药信息。

F.5 多层感知机

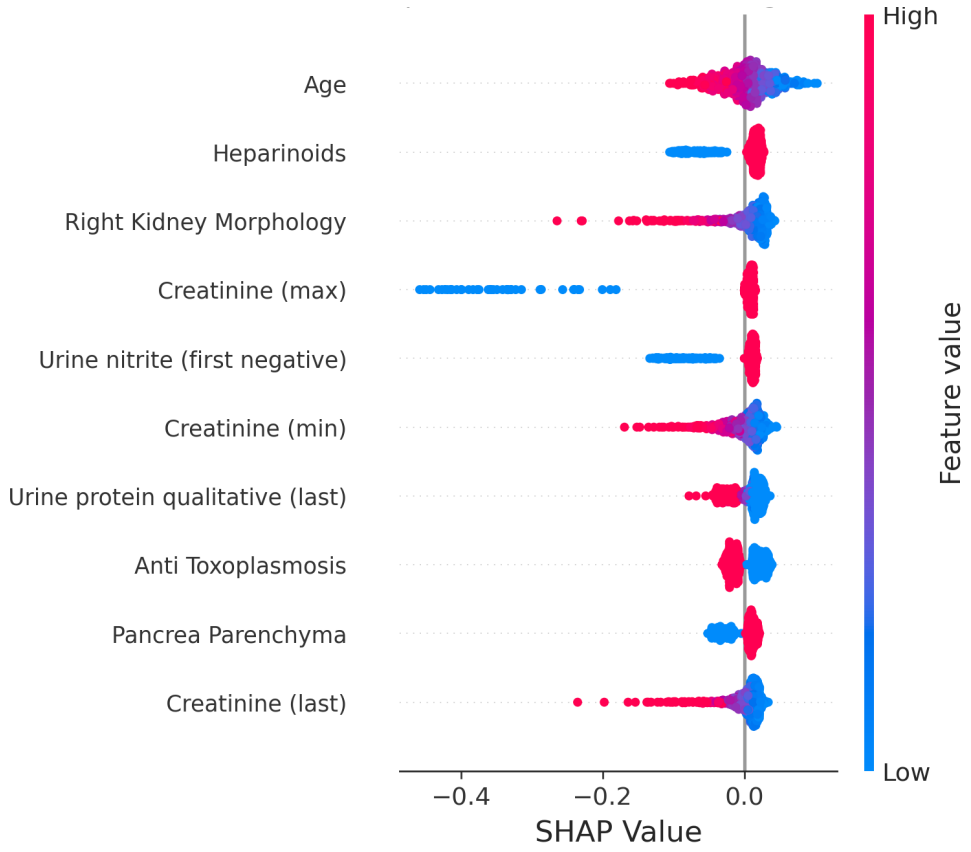

图 F11 展示多层感知机供体移植预测中各特征分布及影响的 SHAP 值蜂群图。每个点代表一个独立样本（测试集），其位置由 SHAP 值决定，颜色则表示该特征的实际值，体现了特征影响的大小和方向。重要特征包括年龄、肾功能指标的时间序列聚合以及用药信息。