预测糖尿病和慢性肾病中的主要不良心脏事件：西里西亚糖尿病心脏项目的机器学习研究

2025年9月2日 20:21

Predicting major adverse cardiac events in diabetes and chronic kidney disease: a machine learning study from the Silesia Diabetes-Heart Project

Hanna Kwiendacz[1†], Bi Huang[2,3†], Yang Chen[2†], Oliwia Janota[1,4], Krzysztof Irlik[1,2], Yang Liu[2,5], Marta Mantovani[2,6], Yalin Zheng[2,7], Mirela Hendel[8], Julia Piaśnik[8], Wiktoria Wójcik[8], Uazman Alam[2,9], Janusz Gumprecht[1], Gregory Y. H. Lip[2,10†] and Katarzyna Nabrdalik[1,2*†]

Background:
患有糖尿病（DM）和慢性肾脏病（CKD）的人患心血管事件（CVE）的风险很高，但是传统风险预测方法的预测性能受到限制
Methods:
1、我们利用机器学习（ML）模型来预测Silesia Diabetes-heart Project的DM和CKD的CVE，这是一个常规的护理数据集标准
2、心血管事件（CVE）定义为非致命性心肌梗死、新发心力衰竭、非致命性卒中、新发房颤、经皮冠状动脉介入治疗或冠状动脉旁路移植术、因心血管疾病住院或死亡的复合事件
3、我们构建了五个机器学习模型（逻辑回归[LR]、随机森林[RF]、支持向量机分类[SVC]、轻梯度提升机[LGBM]和极限梯度提升机[XGBM]）。比较了这五个机器学习模型的预测性能，并使用Shapley可加性解释（SHAP）评估了模型的可解释性
Results:
1、在3056名糖尿病和慢性肾脏病患者中，共纳入了1116名糖尿病和慢性肾脏病患者（中位年龄为67 [IQR为57-76]岁；57%为男性）
3.1 年的随访期间，CVE 的发生率为 14.1%（157/1,116）
2、单变量 Logistic 回归、Boruta 回归和最小绝对收缩与选择操作器 [LASSO] 回归，确定了 10 个重要特征
3、在基于这些特征的五个 ML 模型中，LGBM 的曲线下面积 [AUC] 最高（AUC = 0.740，95% 置信区间 [CI] 0.738-0.743），其次是 LR（AUC = 0.621，95% CI 0.618-0.623）、RF（AUC = 0.707，95% CI 0.704-0.709）、SVC（AUC = 0.707，95% CI 0.704-0.710）和 XGBM（AUC = 0.710，95% CI 0.707-0.713）
4、LGBM 的 Recall（0.739）、F1-score（0.820）和 G-mean （0.826）相对较高。LGBM 的 SHAP 图显示，估计肾小球滤过率（eGFR）、年龄和甘油三酯葡萄糖指数是预测 CVE 的三个最重要特征

- **LGBM (Light Gradient Boosting Machine)**：一种高效的梯度提升框架，使用基于树的学习算法。[8]
- **LR (Logistic Regression)**：逻辑回归，是一种广泛应用于二分类问题的统计模型。[2][9][10]
- **RF (Random Forest)**：随机森林，通过构建大量的决策树并综合其结果来进行预测，是一种集成学习方法。
- **SVC (Support Vector Classification)**：支持向量分类器，其目标是找到一个能将不同类别样本最佳分开的决策边界。[11][12][13][14]
- **XGBM (Extreme Gradient Boosting)**：极限梯度提升，是另一种梯度提升算法，以其高性能和高效率而闻名。[8][15][16][17]

- **单变量逻辑回归 (Univariate Logistic regression)**：这是一种统计方法，用于评估单个预测变量与一个二元结果（例如，是否发生心血管事件）之间的关联强度。[1][2][3]
- **Boruta**：这是一种基于随机森林算法的特征选择方法，它通过创建“影子特征”(原始特征的随机排列版本）来帮助识别所有与结果相关的特征，而不仅仅是那些在模型中表现最优的特征。[4][5][6][7]
- **LASSO回归 (Least Absolute Shrinkage and Selection Operator)**：这是一种线性回归的扩展，它能够在模型训练过程中将不重要的特征的系数压缩至零，从而实现自动化的特征选择。

Introduction
1、然而，糖尿病（DM）和慢性肾脏病（CKD）患者的表型具有高度异质性，这使得传统预测方法的有效性受到限制，因此在筛选出需要优化治疗的个体方面存在重大挑战
2、以往的研究已经使用传统的统计学方法建立了一些模型，但它们的预测能力有限
3、以往的研究已利用 ML 预测了单纯糖尿病或慢性肾脏病患者的 CVE，但侧重于预测糖尿病合并慢性肾脏病患者 CVE 的研究还很有限
Participants in ML analysis
本研究的参与者是患有糖尿病和慢性肾脏病的患者，我们的ML分析旨在建立一个基于临床变量的预测模型，通过以下方法对CVE高危人群进行分层：
（1）构建ML模型，将两类人群区分开来：接受过CVE的人群和随访期间未发生CVE的人群；
（2）评估ML模型的预测性能
（3）将ML结果可视化，使其具有可解释性
Feature selection for ML analysis
1、在预先设计的病例报告表中，共包含 81 个变量

2、剔除了缺失值超过总数 20% 的变量，并对其余缺失值进行了多重插补

3、如果两个变量之间存在明显的共线性（Spearman correlation > 0.6），为了避免共线性，就不再对其中一个变量进行特征选择，变量选择的原则是基于变量的临床重要性

4、将所有糖尿病和慢性肾脏病患者随机分为训练队列和验证队列，比例为 7:3。然后使用单变量 Logistic 分析、Boruta 和最小绝对收缩和选择算子（LASSO）回归筛选特征进行 ML 分析

5、单变量 Logistic 分析是一种基于 P 值的经典变量筛选方法，P 值小于 0.05 的变量被视为具有统计学意义。Boruta 算法是一种基于随机森林算法的特征排序和选择算法[15]，LASSO 回归是一种用 L1 正则进行惩罚的正则化方法

6、最终用于 ML 分析的特征以三种筛选方法为基础，同时考虑了临床重要性

Construction of ML models and performance evaluation

1、在本研究中，基于模型的多样性与代表性、在不同数据类型上的性能、稳健性与处理过拟合的能力，以及实际应用性，我们构建了五种机器学习模型逻辑回归（LR）、随机森林（RF）、支持向量分类（SVC）、轻量化梯度提升机（LGBM）和极限梯度提升机（XGBM）

- **模型的多样性与代表性（Model diversity and representation）**：他们选择了一系列不同类型、不同工作原理的模型。这就像是让一个由工程师、艺术家、医生和律师组成的团队来解决同一个问题，每个人都有不同的视角。通过比较这些不同模型的表现，可以更全面地评估哪种方法最适合当前的数据和问题。
- **在不同数据类型上的性能（Performance across different data type）**：不同的模型对不同类型的数据（例如，线性关系数据 vs. 复杂非线性关系数据）有不同的偏好。选择多种模型可以确保至少有一种能够很好地捕捉数据中潜在的复杂模式。
- **稳健性与处理过拟合的能力（Robustness and handling of overfitting）**：
  - **稳健性（Robustness）**指的是模型在面对数据中的噪声或微小变化时，其性能是否依然稳定。
  - **过拟合（Overfitting）**是机器学习中的一个常见问题，指的是模型在训练数据上表现完美，但在新的、未见过的数据上表现很差，因为它“记忆”了训练数据的特定细节，而不是学习到底层的普遍规律。选择那些内置了防止过拟合机制的模型（如随机森林、梯度提升树）是非常重要的。
- **实际应用性（Practical applicability）**：考虑模型在未来临床实践中是否易于部署和解释。有些模型虽然精确，但可能是个“黑箱”，难以解释其决策过程，这在医疗领域是不利的。

2、模型的训练与优化策略

对于每个模型，我们都在训练数据上使用5折交叉验证，对其超参数进行了优化，并在可能的情况下于训练过程中采用了早停策略

超参数优化 (Hyperparameters were optimised): “超参数”是模型在开始学习之前就需要人为设定的参数，例如决策树的深度、随机森林中树的数量等。这些参数的设置直接影响模型的性能。优化的过程就是通过系统性的尝试（如网格搜索），为每个模型找到一组能使其表现最佳的超参数组合。

5折交叉验证 (5-fold cross-validation): 这是评估和优化模型性能的黄金标准。它将训练数据随机分成5个互不重叠的部分（“折”）。然后进行5轮训练和验证：

第一轮：用第1折作为验证集，其余4折作为训练集。

第二轮：用第2折作为验证集，其余4折作为训练集。

...以此类推，直到每一折都做过一次验证集。

最后，将5轮得到的性能指标（如准确率、AUC）取平均值，作为模型最终的性能评估。这样做可以有效避免因数据划分的偶然性带来的评估偏差，使得结果更加稳健和可信。

早停 (Early stopping): 这是一种防止过拟合的有效策略，主要用于像LGBM和XGBM这样的迭代式训练模型。在训练过程中，模型会不断地在训练集上提升性能，但它在验证集上的性能可能会在某个点之后开始下降（这标志着过拟合的开始）。早停策略就是监控模型在验证集上的性能，一旦发现性能不再提升甚至开始变差，就立即停止训练，并采用性能最佳时的模型状态。

在验证队列中使用 1,000 次引导迭代，通过接收器操作特征曲线（ROC）和每个分类器的平均曲线下面积（AUC）和 95% 置信区间（CI）来评估 ML 模型的主要预测性能。此外，还计算了每个分类器的准确度、特异性、灵敏度、精确度、召回率、F1-分数和 G 平均值

3、处理数据不平衡问题

由于发生事件的样本和未发生事件的样本数量之间存在不平衡，我们在构建这些模型时使用了样本加权

- **问题：数据不平衡 (Imbalance)**: 在医学研究中，这很常见。例如，在一个预测心脏病发作的研究中，最终发病的患者（事件组, events）可能只占总人数的10%，而90%的人是健康的（非事件组, non-events）。如果直接用这样的数据训练模型，模型可能会“偷懒”，倾向于把所有人都预测为“健康”，这样也能达到90%的准确率，但这个模型对于识别真正需要关注的少数高风险患者毫无用处。
- **解决方法：样本加权 (Sample weight)**: 这是一种有效的处理策略。它的核心思想是，在模型训练时，**人为地给少数类（在这里是“事件组”）的样本赋予更高的权重**。这相当于告诉模型：“虽然这些事件样本数量少，但它们非常重要，如果你预测错了它们，将会受到更严厉的惩罚。”通过这种方式，可以迫使模型更加关注少数类样本的特征，从而学习到如何更准确地识别它们，最终提高模型对关键事件的预测能力。

Model interpretability模型的可解释性

使用shap来进行解释

Statistical analysis

第一部分：数据预处理与描述性统计

1、在原始数据集中，缺失值超过20%的变量被丢弃；对于其他存在缺失值的变量，则在R软件中使用“mice”包（版本3.16.0）进行多重插补

2、连续变量以中位数和四分位距（IQR）表示，并因其非正态分布的特征而采用Mann-Whitney U检验进行比较

3、分类变量以计数和百分比表示，并采用费舍尔精确检验或卡方检验进行比较

第二部分：模型构建与分析工具

1、整个数据集被划分为训练队列和测试队列，以及使用单变量逻辑分析、Boruta和LASSO回归进行的特征选择，这些都是通过R软件（版本4.3.3，奥地利）完成的

2、机器学习算法的实现、预测性能的评估，以及使用SHAP对LGBM模型进行的可视化，则是通过Python（版本3.11.5）实现的

3、双尾P值小于0.05被认为具有统计学显著性

Results

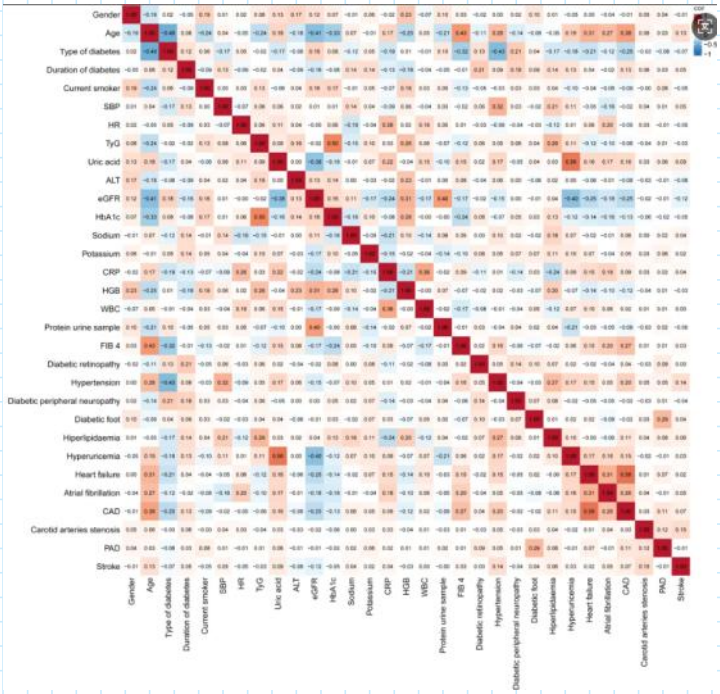
Characteristics of the studied cohort

略

Feature selection for ML

1、变量的初选择

在初始数据集的 81 个变量中，有 7 个变量（低密度脂蛋白、高密度脂蛋白、非高密度脂蛋白、甘油三酯与高密度脂蛋白比值、第一天胰岛素抵抗代谢评分（METS）、最后一天胰岛素抵抗代谢评分（METS）和致动脉粥样硬化指数）因缺失值相对较高而被舍弃。然后对剩余变量之间的共线性进行检验，两个变量之间的斯皮尔曼相关性大于 0.6 即为显著共线性。在排除了 43 个有明显共线性的变量后，最终有 31 个变量进入特征选择过程（补图 1）



2、说明训练集和测试集是可比的

数据集以 7:3 的比例随机分为训练组和测试组。训练组和测试组的所有比较参数均具有可比性（P 均大于 0.05）。训练组和测试组的基线特征见补充表 2

原文: "All the compared parameters were comparable between training and testing cohort (all P > 0.05)."
翻译: "所有被比较的参数在训练队列和测试队列之间都是可比的（所有的P值都大于0.05）。"

- **核心概念：验证划分的公平性 (Sanity Check)**
  - 这句话的目的是为了证明第一步的"随机划分"是成功的。虽然是随机的，但偶尔也可能因为运气不好，分出两组差异很大的数据（比如，训练集里老年人碰巧特别多，而测试集里年轻人特别多）。
  - 为了排除这种小概率事件，研究者必须进行一个"事后检查"。他们会对两个组的所有重要基线特征（如年龄、性别、各项生化指标、病史等）进行**统计学检验**（例如，对连续变量用t检验或Mann-Whitney U检验，对分类变量用卡方检验）。
- **"可比的 (comparable)"是什么意思?**
  - 在统计学上，"可比"意味着两个组之间在某个特征上**没有统计学上的显著差异**。
- **如何解读"all P > 0.05"?**
  - 这是判断是否存在"显著差异"的黄金标准。
  - P值代表了"如果两组实际上没有差异，我们能观测到当前差异或更大差异的概率"。
  - 当P值很小（通常以0.05为界，即P < 0.05）时，我们认为这是一个小概率事件，因此我们有理由相信两组之间确实存在显著差异。
  - 反之，当**P值很大（P > 0.05）时，我们认为观测到的差异很可能是由随机抽样误差造成的，因此我们没有足够的证据**去说两组之间有真正的差异。
  - 所以，"all P > 0.05"这句话强有力地证明了：**"我们的随机划分非常成功，训练集和测试集在所有重要特征上都非常相似，基线水平是一致的。因此，用这个测试集来评估我们用训练集训练出的模型，是完全公平和无偏的。"**

3、特征变量的选择

筛选特征的方法有三种。单变量逻辑分析确定了 13 个变量（补充表 3）。Boruta 和 LASSO 回归分析结果见图 2。6个变量（年龄、eGFR、CAD、心衰、甘油三酯血糖[TyG]指数和高血压）得到确认，6个变量（卒中史、血红蛋白A1c[HbA1c]、性别、CRP[C反应蛋白]、纤维化4评分和尿酸）为暂定变量（图 2A）。在 LASSO 回归中，选择了 12 个变量作为潜在的预测因子（年龄、性别、eGFR、CAD、心衰、HbA1c、CRP、TyG 指数、高血压、血红蛋白、中风史和心率），最佳变量数为 10（图 2B 和 C）。根据三种方法筛选出的特征并考虑到临床重要性，我们最终选择了 10 个特征（年龄、性

别、eGFR、HbA1c、CAD、TyG 指数、心衰、CRP、高血压、中风病史）进行 ML 分析。

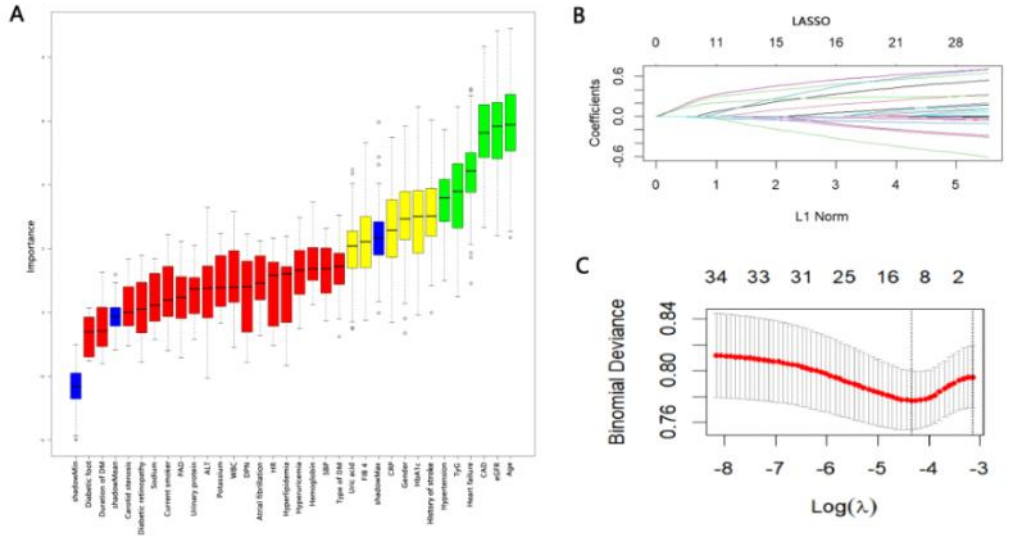


Fig. 2 Feature selection with Boruta and LASSO regression. **A** The blue plot shows minimum, average, and max shadow score. Variables having box plot in green are important, in yellow as tentative, and in red as rejected; **B** The correlation between L1 norm and different coefficients in LASSO regression. L1 norm is the regularisation term for LASSO; **C** The correlation between lamda with binomial deviance. There are two dashed lines in the graph. The left dashed line indicates the minimum mean squared error while the right one indicates one standard error away from the minimum mean squared error. ALT, alanine aminotransferase; CAD, coronary artery disease; CRP, C-reactive protein; DM, diabetes mellitus; DPN, diabetic peripheral neuropathy; eGFR, estimated glomerular filtration rate; FIB-4, fibrosis 4 score; HbA1c, haemoglobin A1c; HR, heart rate; PAD, peripheral artery disease; SBP, systolic blood pressure; TyG, triglyceride glucose; WBC, white blood cell

以下是对特征变量的选择部分进行详细的说明和解析

核心目标：从众多变量中精炼出“黄金组合”

在建立预测模型之前，研究者通常会收集大量潜在的预测变量。如果把所有变量都直接扔进模型，可能会导致**过拟合**（模型在训练数据上表现很好，但在新数据上表现很差）并降低模型的可解释性。因此，这一部分的核心目标就是：**通过三种不同原理的方法，相互验证，并结合临床专业知识，筛选出一个数量不多但预测能力最强的核心变量组合。**

第一步：三种筛选方法的应用与结果

研究者兵分三路，使用了三种不同的方法来进行特征筛选。

方法一：单变量逻辑回归 (Univariate Logistic analysis)

- **原理**：这是一种传统的统计学方法。它将每一个候选变量**单独**拿出来，与最终要预测的结果（例如，是否发生心血管事件）建立一个简单的逻辑回归模型。然后看这个模型的P值是否显著（通常 $P < 0.05$）。
- **结果**：如文字所述，这种方法初步筛选出了**13个**有潜力的变量。这是一个比较宽松的初步筛选。

方法二：Boruta算法 (对应图Fig. 2A)

- **原理**：这是一种非常聪明的、基于随机森林算法的特征选择方法。它的核心思想是：一个特征到底重不重要，不应该只看它自己的得分，而应该看它**是否比随机的“噪声”更重要**。
  1. 它为数据集中的每一个真实特征创建一个“影子特征”（Shadow Feature），这个影子特征就是把原始特征的数值完全打乱得到的。
  2. 然后，它用包含“真实特征”和“影子特征”的完整数据集去训练一个随机森林模型，并计算每个特征的重要性得分。
  3. 最后，它反复比较每个真实特征的重要性得分是否**显著高于**所有影子特征中的最高分。
- **图Fig. 2A的详细解读**:
  - **Y轴 (Importance)**: 代表了随机森林模型计算出的特征重要性得分。
  - **X轴**: 列出了所有被检验的变量。
  - **蓝色箱线图 (Blue plot - 影子特征)**: 这三个蓝色的箱线图代表了所有“影子特征”重要性得分的分布。它们设定了一个**基准线**或**“随机噪声的重要性阈值”**。shadowMin, shadowMean, shadowMax 分别代表影子特征重要性的最小值、平均值和最大值。
  - **绿色箱线图 (Green - 确认重要)**: 这些变量（如Age, eGFR, CAD, Heart failure等）的箱线图整体都**高于**shadowMax。这意味着在多次迭代中，它们的重要性始终显著地超过了随机噪声。因此，Boruta算法**确认**它们是重要的预测因子。
  - **红色箱线图 (Red - 确认不重要)**: 这些变量的箱线图整体都**低于**影子特征的基准。这说明它们的重要性与随机噪声无异甚至更差，因此被算法**拒绝**。
  - **黄色箱线图 (Yellow - 待定/犹豫)**: 这些变量的箱线图与蓝色影子特征的区域有重叠。这意味着它们的表现不稳定，时而比噪声重要，时而不如。算法无法确定它们是否真的重要，因此将它们标记为**“待定” (tentative)**。
- **结果**：如文字所述，Boruta算法**确认了8个变量**（绿色部分），同时认为**6个变量是待定的**（黄色部分）。

方法三：LASSO回归 (对应图Fig. 2B 和 2C)

- **原理**：LASSO (Least Absolute Shrinkage and Selection Operator) 是一种“惩罚性”的回归方法。在建立模型时，它会对模型的系数（coefficients）施加一个惩罚（L1惩罚），这个惩罚会迫使那些对预测结果贡献不大的变量的**系数直接变为零**。因此，LASSO在建模的同时，也自动完成了特征选择。
- **图Fig. 2B (系数路径图) 的详细解读**:
  - **Y轴 (Coefficients)**: 每个变量在LASSO模型中的回归系数。
  - **X轴 (L1 Norm / 上方的数字)**: 代表了模型的复杂程度。从左到右，施加在模型上的“惩罚”力度逐渐减小，因此允许越来越多的变量进入模型（即系数从0变为非0）。上方的数字代表在该惩罚力度下，模型中包含的非零系数变量的数量。

- **解读**：我们可以看到，在最左侧（惩罚最强时），所有变量的系数都是0。随着惩罚减弱，一些重要的变量（那些线最先“起飞”的）开始进入模型。这条图直观地展示了变量被选入模型的先后顺序和其系数的变化路径。文字中提到“12 variables were selected as potential predictors”，这可能对应图中X轴上方数字为12的那个时间点。

- **图Fig. 2C（交叉验证调优图）的详细解读**：
  - **目的**：这张图的目的是为了从图B的所有可能模型中，找到一个**最优模型**。
  - **Y轴（Binomial Deviance）**：模型的误差或“坏度”，数值越低代表模型预测得越准。
  - **X轴（Log(λ)）**：LASSO惩罚项的强度λ（取对数）。越往右，惩罚越弱，模型越复杂。
  - **红色曲线与灰色误差带**：红色曲线是经过交叉验证得到的不同λ值对应的平均模型误差。
  - **两条虚线**：
    - **左侧虚线（lambda.min）**：对应着模型误差达到**绝对最小值**的那个点。这是“最准”的模型。
    - **右侧虚线（lambda.1se）**：对应着一个“在最小误差的一个标准差范围内最简洁的模型”。这是一种更保守的选择，旨在防止过拟合。
- **结果**：文字明确指出，根据交叉验证的结果（很可能是选择了左侧虚线对应的模型），LASSO回归最终选出的**最优变量数量是10个**。

第二步：汇总与最终决策

在分别用三种方法得到结果后，研究者并没有简单地取其中一种方法的结果，而是进行了一个综合决策

- **综合分析**：研究者会看哪些变量是“三好学生”，即同时被多种方法选中（例如Age，eGFR，CAD等）。这些变量无疑是核心中的核心。
- **处理分歧**：对于那些结果不一致的变量（例如，被LASSO选中但被Boruta认为是“待定”的变量），研究者会引入第四个、也是最重要的评判标准——**临床重要性（clinical importance）**。
- **专家决策**：临床医生或领域专家会根据自己的专业知识来判断：
  - 这个变量在临床实践中是否容易获取？
  - 它是否代表了一个重要的病理生理过程？
  - 即便它在统计上稍弱，但如果临床上公认它很重要，是否也应该被保留？
- **最终结果**：经过这番“三堂会审”加上“专家评审”，研究团队最终确定了一个由**10个特征组成的“梦之队”**，用于后续更复杂的机器学习建模。这10个特征是：age，gender，eGFR，HbA1c，CAD，TyG index，heart failure，CRP，hypertension，history of stroke。

网页链接：SCI一区论文预测模型变量筛选好思路：LASSO与Boruta算法结合

Evaluation of ML models

图 3 显示了五个 ML 模型的 ROC。在五个 ML 模型中，LGBM 模型的 AUC 最高（0.740，95% CI 0.738-0.743），而 RF、SVC 和 XGBM 模型的 AUC 相近（RF，0.707，95% CI 0.704-0.709；SVC，0.707，95% CI 0.704-0.710；XGBM，0.710，95% CI 0.707-0.713）。LR 的 AUC 最低（0.621，95% CI 0.618-0.623）。五种 ML 模型的其他指标见表 2，其中 LGBM 模型的准确度（0.723）、特异度（0.739）、精确度（0.923）和 F1 分数（0.820）相对较高。

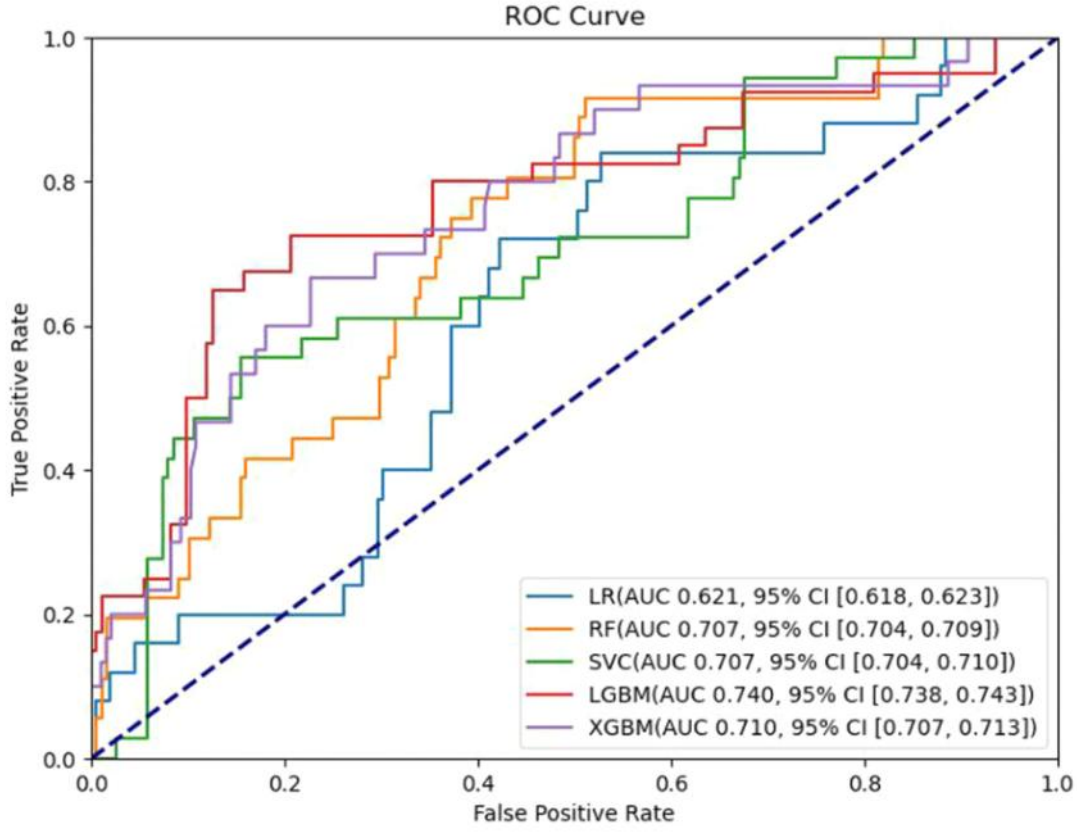
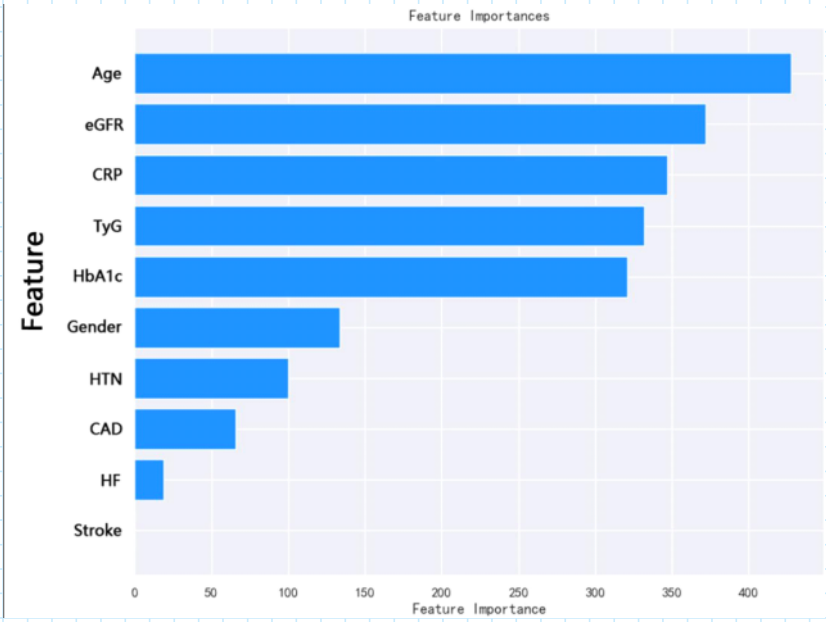


Fig. 3 The ROCs for predicting CVEs of different ML models. AUC, area under curve; LGBM, Light gradient boosting machine; LR, Logistic regression; RF, Random forest; SVC, Support vector classification; XGBM, eXtreme gradient boosting machine

Model interpretability

由于 LGBM 模型在五个 ML 模型中具有最佳预测性能，因此我们选择该模型来解释模型的输出结果。基于 LGBM 模型的特征重要性如附图 2 所示。最重要的五个特征是年龄，其次是 eGFR，然后是 CRP、TyG 指数和 HbA1c。



SHAP 值更能说明 LGBM 模式是如何预测结果的。图 4 显示了 SHAP 汇总图所总结的特征重要性。最重要的三个特征是 eGFR，其次是年龄和 TyG 指数，其中 eGFR 越低、年龄越大和 TyG 指数越高，模型的可解释性就越高。三个最重要特征（eGFR、年龄和 TyG 指数）与前七个特征的相互作用和依赖关系见补图 3。我们选择了一个具有代表性的样本，通过力图来描述输出结果，说明 SHAP 值如何解释各个模型特征（补图 4）。每个特征对预测结果的概率都有不同的影响，这共同决定了这些特征的最终贡献（补图 4A）。20 个代表性观察结果的决策图直观地显示了每个样本中的每个特征对总体预测的贡献图 4B）。

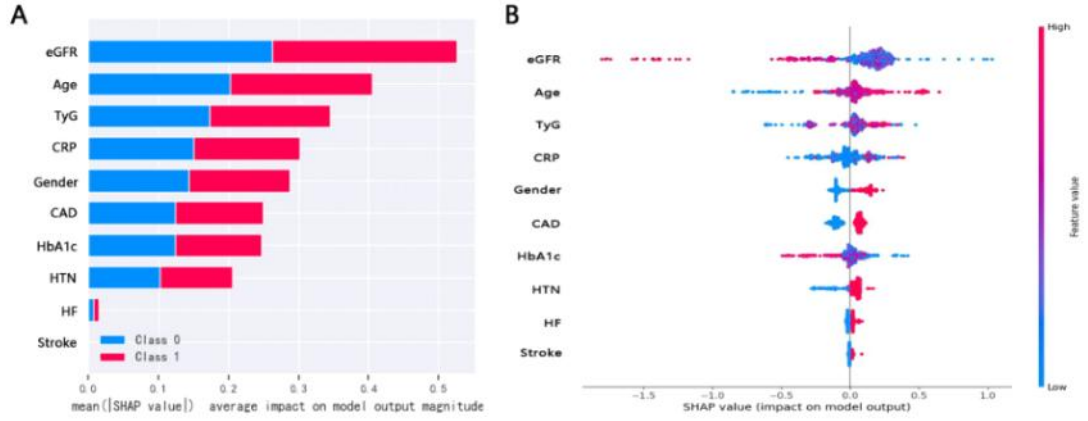


Fig. 4 The importance of features and their contribution to the output with SHAP value evaluation. **A** The importance of the ten features was shown in descending order and in two matrixes, Class 0 and Class 1, representing negative and positive classification, respectively; **B** SHAP beeswarm plot to interpret the contribution of each feature with red color increasing the interpretability while blue color decreasing the interpretability. CRP, C-reactive protein; eGFR, estimated glomerular filtration rate; HbA1c, haemoglobin A1c; HF, heart failure; HTN, hypertension; TyG, triglyceride glucose index. Stroke means history of stroke

图 A：SHAP 特征重要性图（Summary Bar Plot）

这张图的核心目的是回答一个宏观问题：**“总的来说，哪些特征对模型的预测结果影响最大？”**

如何解读？

- **Y轴**：列出了最终被模型使用的10个特征，并且按照它们**整体的重要性**从上到下进行了排序。我们可以看到，**eGFR（估算肾小球滤过率）** 是最重要的特征，其次是**Age（年龄）**，然后是**TyG（甘油三酯葡萄糖指数）**，以此类推。
- **X轴 (mean(|SHAP value|))**：代表了每个特征对模型输出影响的**平均绝对值**。这个值越大，说明该特征无论是以正面还是负面的方式影响预测，其“音量”或“影响力”都更大。
- **堆叠的条形图（蓝色 vs. 红色）**：这是这张图最精妙的地方。它不仅告诉我们一个特征有多重要，还告诉我们它是如何影响不同预测结果的。
  - **Class 1（红色）**：代表模型预测结果为“阳性”（例如，会发生心血管事件）。红色条的长度表示，当模型预测为“阳性”时，这个特征平均贡献了多大的影响力。
  - **Class 0（蓝色）**：代表模型预测结果为“阴性”（例如，不会发生心血管事件）。蓝色条的长度表示，当模型预测为“阴性”时，这个特征平均贡献了多大的影响力。

从图A能得到什么信息？

1. **特征排序**：我们可以一目了然地知道哪些是关键驱动因素。在这里，肾功能（eGFR）、年龄（Age）和胰岛素抵抗指标（TyG）是驱动模型预测的“三巨头”。

2. **影响方向的倾向性**:
    - 观察**eGFR**的条形图，红色部分（对预测“阳性”的贡献）远大于蓝色部分。这强烈暗示了eGFR主要是通过其**异常值**（通常是低eGFR）来**推高**患病风险的。
    - 观察**TyG**，同样是红色部分远大于蓝色部分，说明主要是**高TyG值**在**增加**患病风险。
    - 这个图提供了一个关于特征影响力的总体概览。

图 B：SHAP 摘要图（Beeswarm Summary Plot）

这张图则更进一步，它回答了一个更微观、更细致的问题：**“对于每一个样本，每个特征是如何具体地影响预测结果的？特征的数值高低又是如何与影响方向关联的？”**

如何解读？

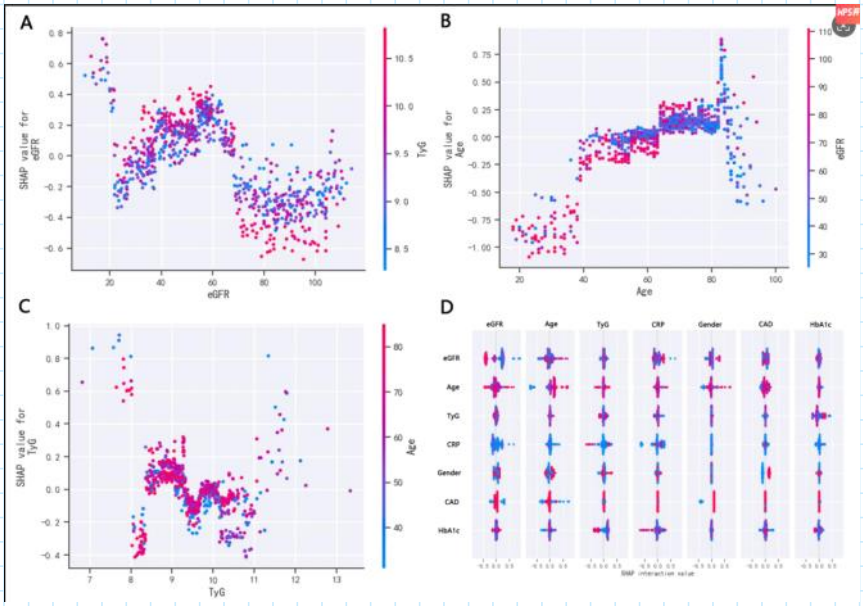
- **Y轴**：和图A一样，是按重要性排序的特征。
- **X轴（SHAP value）**：这是核心。它代表了对于某个特定的样本，一个特征将其预测结果**“推离”**基准值的量。
    - **SHAP值 > 0**：表示该特征的值将模型的预测**推向了“阳性”类别（Class 1）**。这个值越大，推力越强。
    - **SHAP值 < 0**：表示该特征的值将模型的预测**推向了“阴性”类别（Class 0）**。这个值越小（绝对值越大），推力越强。
- **图中的每一个点**：代表了数据集中的**一个样本（一个病人）**。
- **点的颜色（Feature value）**：这是这张图的点睛之笔。点的颜色代表了这个特征在那个样本中的**原始数值的高低**。
    - **红色（High）**：代表特征的数值较高。
    - **蓝色（Low）**：代表特征的数值较低。

从图B能得到什么信息？（这是信息量最大的一部分）

现在，我们可以结合X轴位置（影响方向）、点的颜色（特征值高低）和Y轴（哪个特征）来解读丰富的临床和模型信息：

1. **eGFR**:
    - **现象**：大部分的**蓝色点**（低eGFR值）都分布在X轴的**右侧（SHAP值 > 0）**，而大部分的**红色点**（高eGFR值）都分布在X轴的**左侧（SHAP值 < 0）**。
    - **解读**：**肾功能越差（eGFR值低），越会显著增加模型的预测风险；肾功能越好（eGFR值高），越会显著降低模型的预测风险。** 这与我们的临床知识完全吻合。
2. **Age**:
    - **现象**：**红色点**（年龄大）普遍在**右侧**，**蓝色点**（年龄小）普遍在**左侧**。
    - **解读**：**年龄越大，风险越高；年龄越小，风险越低。**
3. **TyG**:
    - **现象**：**红色点**（高TyG指数）在**右侧**，**蓝色点**（低TyG指数）在**左侧**。
    - **解读**：**胰岛素抵抗越严重（TyG指数高），风险越高。**
4. **CAD（冠心病史）**:
    - **现象**：这是一个二元变量。我们可以看到**红色点**（有冠心病史）集中在**右侧**，而**蓝色点**（无冠心病史）集中在**左侧**。
    - **解读**：**有冠心病史会增加预测风险，没有则会降低风险。**
5. **Gender（性别）**:
    - **现象**：我们可以看到红色点和蓝色点（分别代表两种性别）在X轴两侧有明显的分群。
    - **解读**：性别对风险有明确的影响。具体哪种颜色代表哪种性别需要参照数据编码，但很明显，其中一种性别（红色点）会增加风险，而另一种（蓝色点）会降低风险。

三个最重要特征（eGFR、年龄和 TyG 指数）与前七个特征的相互作用和依赖关系见补图 3



图A, B, C：SHAP 依赖图 (Dependence Plots)

这三张图（A, B, C）的核心目的是回答这样一个问题：**“一个特定特征（如eGFR）对模型预测的影响，是如何随着该特征自身值的变化而变化的？并且，这种影响是否还受到另一个特征（如TyG）的调节？”**

如何解读这三张图？

- **X轴**：代表了我们主要关注的那个特征的**原始数值**。例如，在图A中是eGFR的值。
- **Y轴**：代表了X轴特征的**SHAP值**。我们再回顾一下：
  - Y > 0：说明该eGFR值**增加**了模型的风险预测。
  - Y < 0：说明该eGFR值**降低**了模型的风险预测。
- **每一个点**：代表数据集中的一个样本（一个病人）。
- **点的颜色**：这是揭示**交互作用**的关键！点的颜色由**另一个**我们怀疑与之有交互的特征的数值决定。这个交互特征是SHAP自动选择的，因为它被认为是与主特征交互作用最强的变量。

具体解析

- **图A：eGFR的依赖图，颜色由TyG决定**
  - **主效应**：我们可以看到一条清晰的从左上到右下的总体趋势。当**eGFR值较低时**（X轴在左侧），其SHAP值普遍**为正**（Y > 0），意味着低eGFR会增加风险。当**eGFR值较高时**（X轴在右侧），其SHAP值普遍**为负**（Y < 0），意味着高eGFR会降低风险。这再次验证了我们之前的发现。
  - **交互作用**：现在看颜色。我们发现在eGFR值相似的垂直区域内，**红色点（高TyG）往往比蓝色点（低TyG）有更高**的SHAP值。
  - **解读**：这揭示了一个重要的交互作用：**在肾功能（eGFR）相同的情况下，如果一个患者的胰岛素抵抗更严重（TyG值高），那么模型会认为他的风险更高。** 换句话说，高TyG放大了低eGFR带来的风险。
- **图B：年龄(Age)的依赖图，颜色由eGFR决定**
  - **主效应**：总体趋势是从左下到右上。**年龄越小**，SHAP值越倾向于**为负**（降低风险）；**年龄越大**，SHAP值越倾向于**为正**（增加风险）。
  - **交互作用**：在年龄相似的垂直区域内（例如，看X=80岁附近），**蓝色点（低eGFR）的SHAP值普遍高于红色点（高eGFR）**。
  - **解读**：**在年龄相同的情况下，如果一个患者的肾功能更差（eGFR值低），模型会认为他的风险更高。** 这说明，高龄本身是风险因素，但如果高龄同时合并肾功能不全，风险会被进一步放大。
- **图C：TyG指数的依赖图，颜色由年龄(Age)决定**
  - **主效应**：趋势是从左下到右上。**低TyG值**对应**负的SHAP值**（降低风险），**高TyG值**对应**正的SHAP值**（增加风险）。
  - **交互作用**：这个图的交互作用看起来没有前两个那么清晰，但在TyG值较高的区域（例如X > 10），似乎**红色点（高年龄）**的SHAP值有更高的上限。
  - **解读**：**当患者的胰岛素抵抗已经很严重时（TyG高），如果他还同时是高龄，那么模型会给他一个非常高的风险预测。** 年龄放大了高TyG带来的风险。

图D：SHAP 交互作用值图 (Interaction Value Matrix Plot)

这张图更加宏观，它旨在一次性展示模型中**所有重要特征两两之间的交互效应有多强**。

如何解读？

- 这是一个矩阵图，行和列都是模型中最重要的七个特征。
- **对角线上的图 (Main Effect)**：对角线上的每一个小提琴图（或蜂群图）显示的是该特征的**主效应**的SHAP值分布。这和我们之前看的摘要图（图B）中对应的那一行是类似的，只是形状不同。它告诉我们这个特征本身是如何影响预测的。例如，eGFR对角线上的图，显示其SHAP值分布范围很广，有正有负，说明其主效应很强。
- **非对角线上的图 (Interaction Effect)**：非对角线上的图是关键。它显示的是**两个特征之间的交互效应**的SHAP值。
  - **Y轴**：代表了**SHAP交互作用值**。这个值衡量的是，当两个特征同时出现时，相对于它们各自独立效应的总和，还额外产生了多少“1+1>2”或“1+1<2”的效应。
  - **点的颜色**：颜色由行对应的特征值决定。例如，在第1行第2列的格子里（eGFR与Age的交互），点的颜色由eGFR的值决定。
  - **解读交互效应**：
    - 如果一个格子里的点分布很**分散**，说明这两个特征之间有**很强的交互作用**。
    - 如果一个格子里的点都**紧密地聚集在Y=0附近**，说明它们之间**几乎没有交互作用**。

从图D能得到什么信息？

1. **识别强交互**：我们可以快速浏览非对角线区域，找到那些点分布最“胖”、最分散的格子。
   - 例如，**Age和eGFR**（第1行第2列，和第2行第1列）的交互图非常分散。看第2行第1列（行是Age，颜色由Age决定），我们可以看到**红色点（高年龄）**的交互作用值有正有负，分布很广。这说明，年龄对模型的影响，会根据eGFR水平的不同而发生很大变化，反之亦然。这与我们在图B的依赖图中看到的结论是一致的。
   - **eGFR和TyG**（第1行第3列）的交互作用也很明显。
2. **识别弱交互**：
   - 例如，**Gender和CAD**（第5行第6列）的交互图里的点就非常集中在0附近，说明模型认为性别和是否有冠心病史之间没有太大的交互效应。

总结

- **图A, B, C** 深入剖析了**三个最重要特征**的个体行为，并揭示了它们是如何被另一个强相关特征所“调节”的。
- **图D** 则提供了一个“上帝视角”的**交互网络图**，让我们能一次性比较所有重要特征两两之间的交互强度，快速定位模型中最重要的“组合拳”效应。