Advances in Applied Mathematics 应用数学进展, 2024, 13(6), 2891-2900
Published Online June 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.136277

Hans汉斯

基于机器学习的ACS-Stacking预测模型

王改琴，王晓云*，滕凯民

太原理工大学数学学院，山西 太原

收稿日期：2024年5月28日；录用日期：2024年6月22日；发布日期：2024年6月28日

摘　要

急性冠脉综合征(Acute Coronary Syndrome, ACS)是威胁人类健康的重要疾病，其中急性心肌梗死的快速鉴别诊断技术仍需进一步研究。本研究包含了山西医科大学附属心血管病医院的813名患者的临床数据，由24个与人口统计学/合并症特征和住院并发症相关的预测变量描述。以“急性心肌梗死(Acute Myocardial Infarction, AMI)、不稳定心绞痛(Unstable Angina, UA)”二分类变量为目标变量，建立一个可解释性的机器学习(Machine Learning, ML)模型，确定显著相关指标来辅助临床医师对ACS患者进行快速有效的鉴别。训练并评估了这7种ML模型的性能，将在测试集中表现较好的Xgboost, Adaboost, Randomforest融合成表现最佳的可解释的Stacking融合模型(命名为：ACS-Stacking预测模型)。ACS融合预测模型实现了在测试集的AUC值为0.96562，在10-fold Cross-Validation下的准确率为89%。该模型有助于医生在临床诊断中结合模型预测结果、模型可视化和临床经验快速甄别出ACS患者。

关键词

ACS，机器学习，ACS-Stacking预测模型

ACS-Stacking Prediction Model Based on Machine Learning

Gaiqin Wang, Xiaoyun Wang*, Kaimin Teng

College of Mathematics, Taiyuan University of Technology, Taiyuan Shanxi

Received: May 28th, 2024; accepted: Jun. 22nd, 2024; published: Jun. 28th, 2024

Abstract

Acute Coronary Syndrome (ACS) is a significant disease that threatens human health, and the rapid differential diagnosis technology for acute myocardial infarction still requires further research.

*通讯作者。

文章引用: 王改琴, 王晓云, 滕凯民. 基于机器学习的ACS-Stacking预测模型[J]. 应用数学进展, 2024, 13(6): 2891-2900.
DOI: 10.12677/aam.2024.136277

This study involved clinical data from 813 patients at the Cardiovascular Hospital of Shanxi Medical University, described by 24 predictive variables related to demographic/comorbidity characteristics and in-hospital complications. Using "Acute Myocardial Infarction (AMI) and Unstable Angina (UA)" as binary classification variables as the target variables, an interpretable machine learning (ML) model was established to identify significant related indicators to assist clinicians in making rapid and effective identification of ACS patients. The performance of these seven ML models was trained and evaluated, and the Xgboost, Adaboost, and Randomforest models that performed better in the test set were fused into the best-performing interpretable Stacking ensemble model (named: ACS-Stacking prediction model). The ACS ensemble prediction model achieved an AUC value of 0.96562 in the test set and an accuracy rate of 89% under 10-fold Cross-Validation. This model helps doctors quickly identify ACS patients in clinical diagnosis by combining model prediction results, model visualization, and clinical experience.

Keywords

ACS, Machine Learning, ACS-Stacking Prediction Model

1. 引言

ACS 是威胁人类健康的重要疾病之一[1]，《中国心血管健康与疾病报告 2022》显示，中国心血管疾病患病人数已达到 3.3 亿人口。在居民疾病死亡构成比中，心血管疾病的死亡率居首位，高达 53.20% [2]。冠心病是心血管疾病的重要组成，依据是否存在心肌损伤，可将冠心病分为心绞痛和心肌梗死两部分，其中心肌梗死的死亡率更高、预后更差[3]。依据最新的全球心梗定义[4]，并非所有的 AMI 均由冠状动脉狭窄或堵塞后造成的心肌细胞死亡，而此类 AMI 需尽快开通血管以降低患者风险，因而如何快速、准确的识别此类患者即成为亟待解决的临床问题。

为解决该临床问题并减轻急诊工作负荷，多个临床诊疗指南均推荐使用高敏肌钙蛋白检测来快速进行 AMI 患者和 UA 患者的鉴别。但其在临床实际使用中受到诸多影响，如：患者发病时胸痛症状不典型、心电图未显示 ST 段异常等，可能导致相关生物标记物的检测率不足；此外，不同的年龄、性别对高敏肌钙蛋白的正常范围也存在影响；且生物标志物的升高受多种疾病的影响，如何进行相关疾病的鉴别仍需要临床医生的观察、解释。有文献报道，约 24%~35%的患者存在心肌梗死，但心电图未显示应有的缺血改变，从而导致治疗的延误，并最终导致 ACS 患者死亡率增加 14%~22% [5]。

近年来，人们对机器学习和深度学习算法在医学领域进行了广泛的研究[6]-[8]，其中在心肌梗死、心力衰竭等的发生发展阶段及预后的预测方面均取得了 highlight。目前指南中使用高敏肌钙蛋白的 0/1 h 方案在心肌梗死的排除方面表现亮眼，但在快速阳性诊断方面仍存在欠缺。故在本研究中，我们拟开发一种基于机器学习的 ACS-Stacking 预测模型，在已有数据分析的基础上，进一步提高 ACS-Stacking 预测模型的准确率，以实现 AMI 患者更早、更准确的预测。

2. 方法

2.1. 研究设计与人群

本研究数据来自山西省医科大学附属心血管病医院，其中 813 名患者被纳入研究，504 例患者数据

被分配到训练集，309 例患者数据被分配到测试集，训练了七种机器学习模型，将表现较好的 Adaboost，Xgboost，Randomforest 融合成可解释的 Stacking 模型。

2.2. 数据预处理

KNN 填充算法属于基于局部的单次填充算法，使用 K 个近邻样本对应的属性值判断缺失样本缺失的属性值，能简单快速地解决数据缺失问题[9]。本研究对 813 个原数据进行检测后发现存在缺失值，随后使用 KNN 填充(选取 K = 7)。

2.3. 特征工程

2.3.1. 多重共线性

在相关性分析中可能出现多个自变量之间存在高度相关性，这种现象会使得模型的解释变得困难，并降低了模型识别具有统计显著性的自变量的能力。因此，本研究通过多重共线性分析，来识别和减轻这种问题所带来的影响[10]，使用 Pearson 相关性热图(见图 1)来可视化分析。

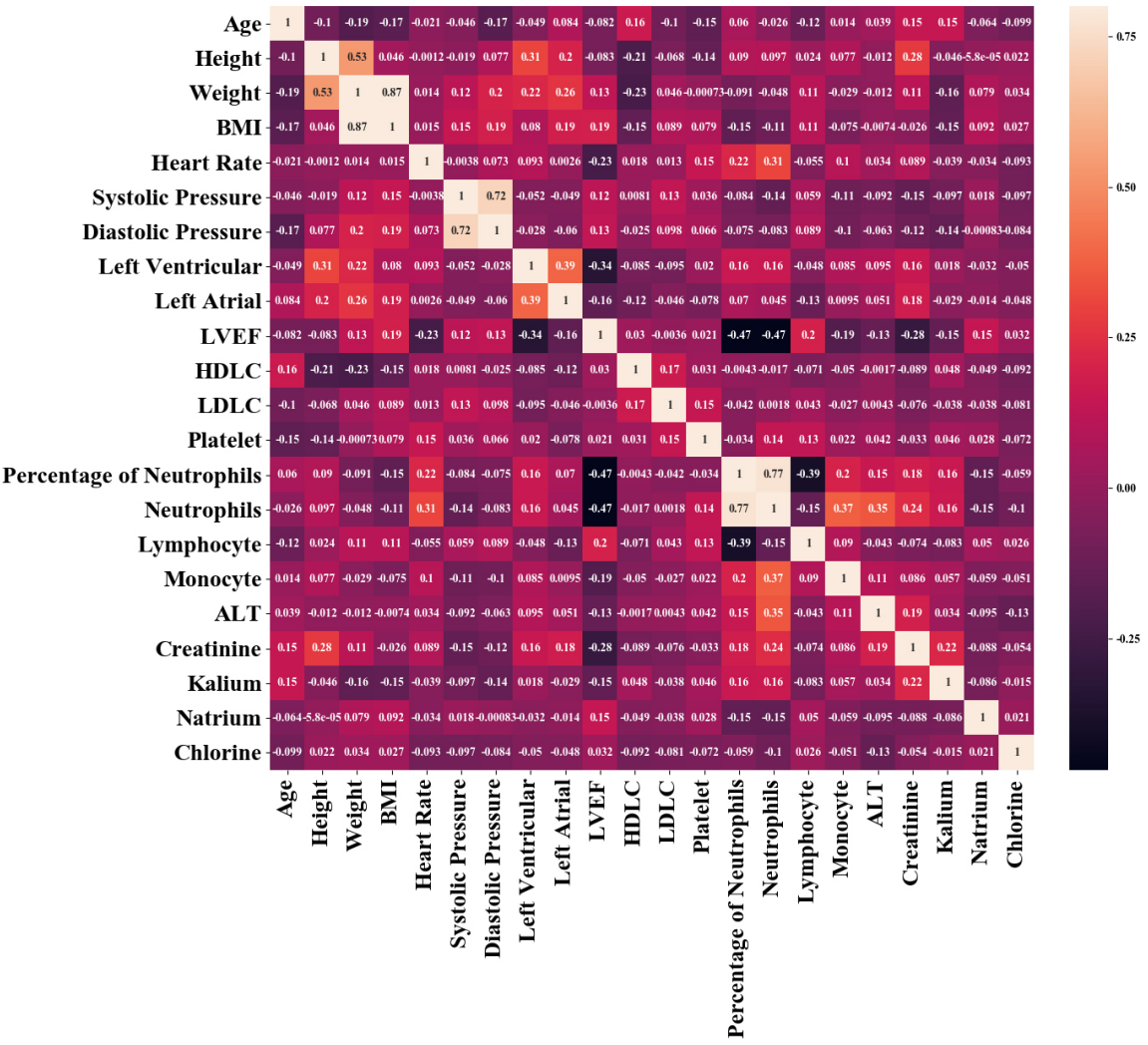


Figure 1. Pearson correlation heat map
图 1. Pearson 相关性热图

图1显示各个特征变量与目标变量之间的Pearson相关性。研究中设定阈值Pearson相关系数R = 0.7，剔除相关系数R > 0.7的两个变量中的一个。故做如下操作：删除体重，嗜中性粒细胞百分比，收缩压三个特征。考虑到收缩压是比较重要的医学指标，因此保留收缩压，只删除体重和嗜中性粒细胞百分比。

2.3.2. 特征选择

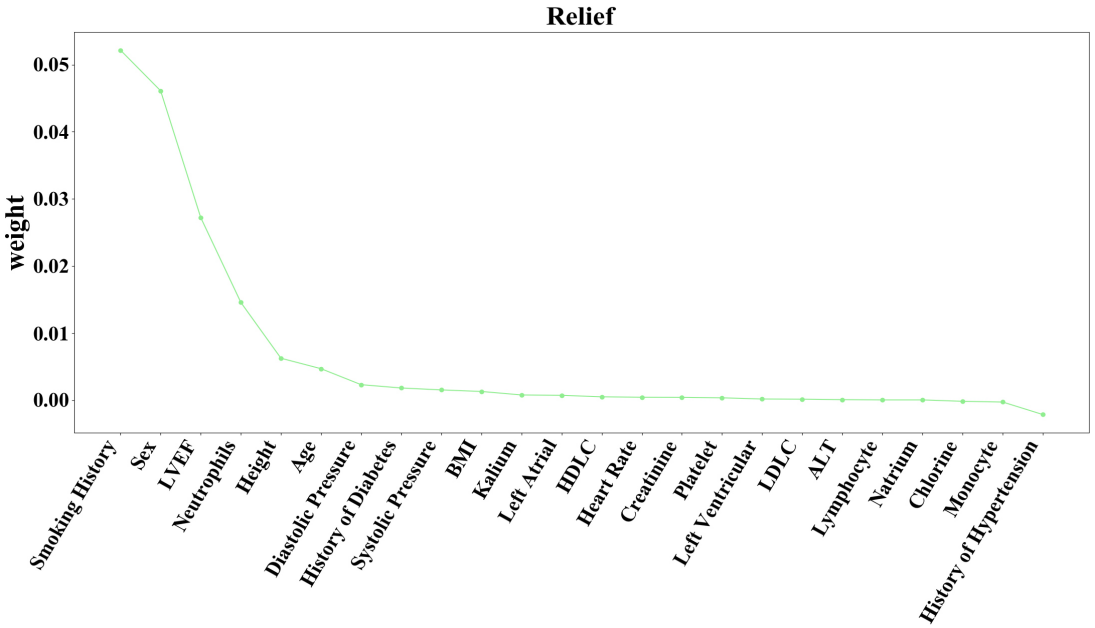


Figure 2. A line chart showing the ranking of the absolute weight coefficients of each feature variable under the Relief algorithm
图 2. 各个特征变量在 Relief 算法下的权重系数按绝对值大小进行排序的折线图

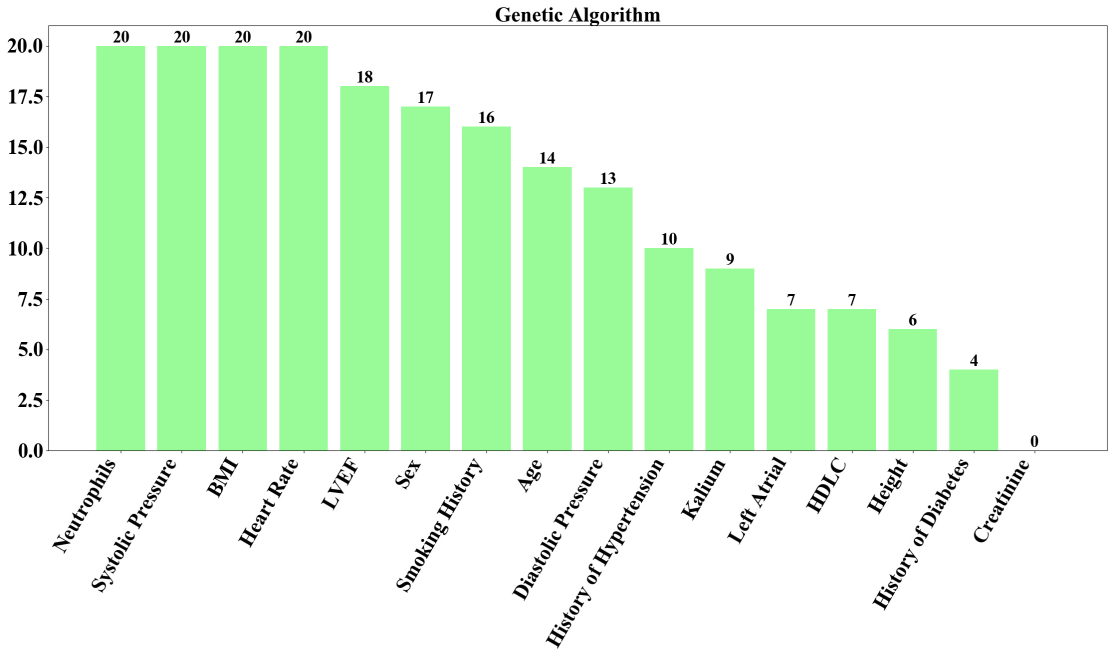


Figure 3. Histogram showing the number of times each feature is selected in the genetic algorithm
图 3. 遗传算法特征选择次数柱状图

本研究利用 Relief 算法，结合上述特征训练 Relief 模型，选择并消除最弱的特征，对剩余的特征重复这个过程，直到达到指定数量的特征[11]。对上述 22 个特征变量赋权重(见图 2)。随后对 Relief 算法筛选后的 16 个特征进行循环 20 次遗传算法之后，得到每一次选出的最佳特征，并列出循环中每个特征被选择的次数(见图 3)。

从图 3 中可以很明显地看出，嗜中性粒细胞、收缩压(mmHg)、BMI 和入院心率这四个特征变量是最为重要性的显著相关指标。

结合 RELIEF 和遗传算法，本研究选择特征权重绝对值在前 70% (16 个特征)以及在遗传算法中被选择次数在 10 次以上的特征，最终选择 10 个特征如下：嗜中性粒细胞、收缩压(mmHg)、BMI、入院心率、LVEF、性别、吸烟史、年龄、舒张压(mmHg)和高血压病史。

2.4. 模型训练与模型评估

2.4.1. 训练测试集划分

本研究采用独立测试集和交叉验证[12]两种评价方法，使用 sklearn 中的 train_test_split 进行划分，按照 9:1 划分训练集和测试集。

2.4.2. 模型训练与调参

本研究拟训练 Xgboost，Adaboost，Randomforest，KNN，SVM，Naive_Bayes，Logistic 七种机器学习分类模型，并将测试集上性能比较好的 Xgboost，Adaboost，Randomforest 融合成表现最佳的 Stacking 模型(其命名为 ACS-Stacking 预测模型)，利用 10-fold 的网格搜索方法对机器学习的超参数进行优化[13]。

7 种 ML 分类模型以及 ACS-Stacking 预测模型的最佳超参数见表 1。

Table 1. Model hyperparameters
表 1. 模型超参数表

模型	超参数
Xgboost	n_estimators = 22, max_depth = 6, alpha = 0.15, colsample_bytree = 0.9, subsample = 0.7, gamma = 0.6, eta = 0.17
Adaboost	DecisionTreeClassifier (max_depth = 9), n_estimators = 80, learning_rate = 0.83
Randomforest	n_estimators = 80, bootstrap = “true”, max_depth = 7, max_features = 0.1, min_samples_leaf = 1, min_samples_split = 2, random_state = 0
KNN	algorithm = “brute”, n_neighbors = 19, p = 1, weights = “distance”
SVM	C = 100, kernel = “rbf”, gamma = “scale”, degree = 1, probability = True
Naive_Bayes	无
Logistic	“C”:0.1, “penalty”: “l2”, “solver”: “newton-cg”
ACS-Stacking	(Randomtree: n_estimators = 80, bootstrap = “true”, max_depth = 7, max_features = 0.1, min_samples_leaf = 1, min_samples_split = 2, random_state = 0) (AdaBoost: DecisionTreeClassifier (max_depth = 7), n_estimators = 80, learning_rate = 0.1) (Xgboost: n_estimators = 80, max_depth = 6, alpha = 0.15, colsample_bytree = 0.9, subsample = 0.7, gamma = 0.6, eta = 0.17)

2.4.3. 模型评估

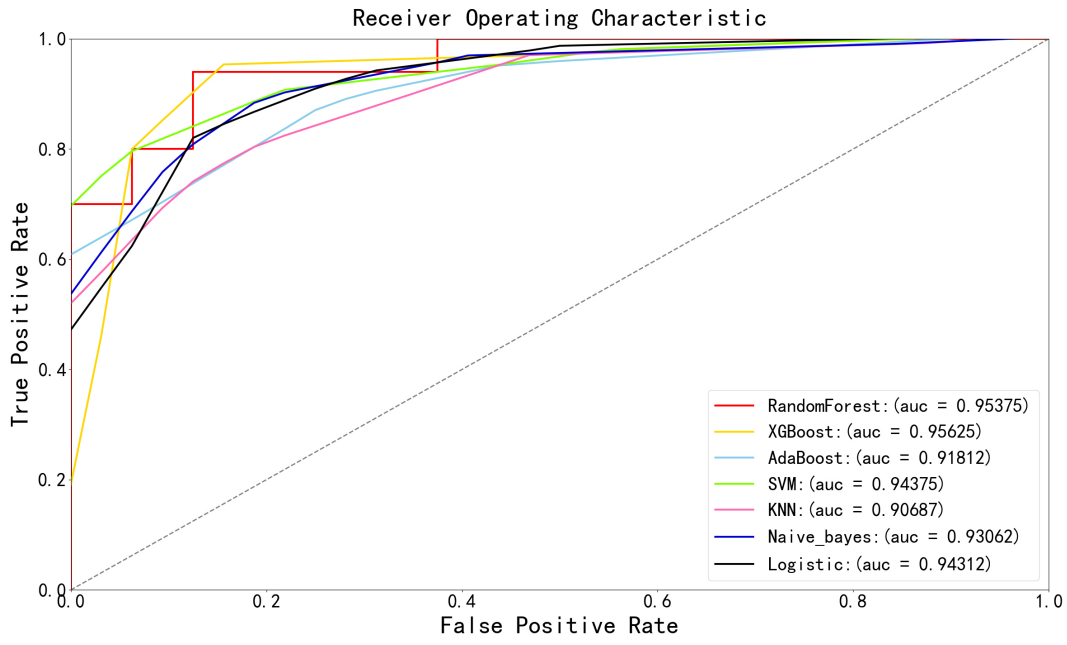
本研究采用 10-fold 验证计算曲线下面积(AUC)、f1 分数、独立测试集准确率和召回率(Recall)、敏感度来评估 8 个模型的性能。

曲线下面积(AUC)因其独立于类分布而被用作性能的预测因子。其他性能指标，包括 F1 分数、召回率，敏感度用于进一步评估模型的校准。本研究从 F1 分数、召回率、敏感度以及 10-fold-cv 下的平均准确率四个评价指标方面，评估了 7 种 ML 分类模型以及 ACS-Stacking 预测模型的预测结果。每个模型在各个评价指标下的结果见表 2，ROC 曲线见图 4。

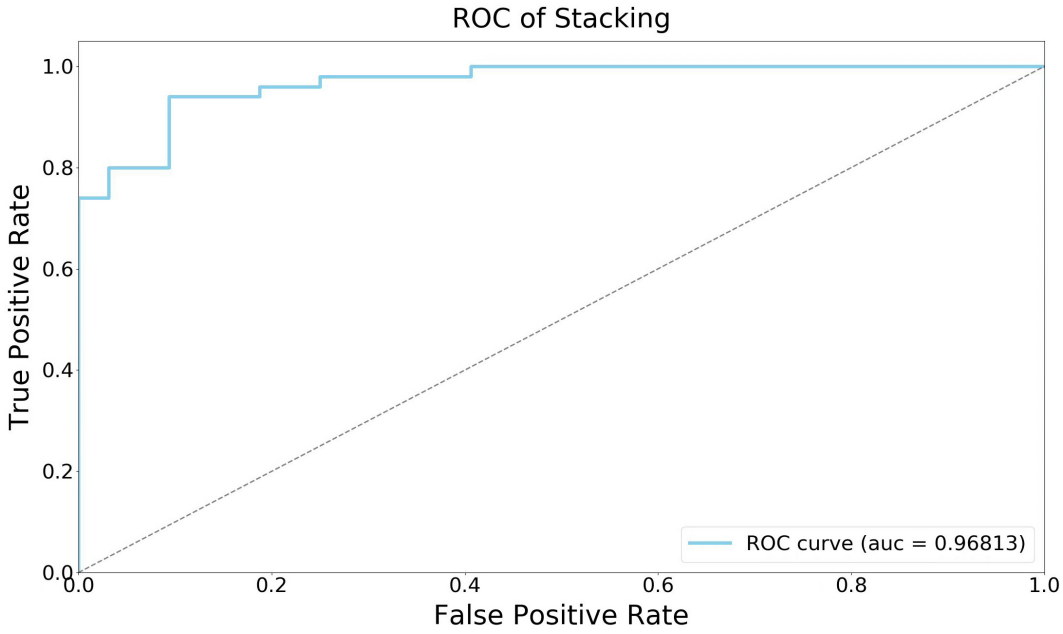
图 4(a)是本研究中训练的 7 种机器学习分类模型的 ROC 曲线图，图 4(b)是 ACS-Stacking 预测模型的 ROC 曲线图。可以看出，不论是图 4(a)中的 7 种机器学习分类模型还是 ACS-Stacking 预测模型，对应的 ROC 值所展示的效果都是较好的。对比后发现 ACS-Stacking 预测模型的效果相较其他分类模型更加优秀。

Table 2. Model evaluation results
表 2. 模型评价结果表

算法	F1 分数	召回率	敏感度	交叉验证
Xgboost	0.887	90.0%	87.5%	0.87
Adaboost	0.903	90.0%	90.625%	0.86
Randomforest	0.889	94.0%	84.375%	0.88
KNN	0.816	82.0%	81.25%	0.82
SVM	0.832	82.0%	84.375%	0.87
Naive_Bayes	0.838	78.0%	90.625%	0.84
Logistic	0.825	78.0%	87.5%	0.87
ACS-Stacking	0.923	94.0%	90.625%	0.89



(a)



(b)

Figure 4. ROC curve of fusion model
图 4. 融合模型 ROC 曲线图

2.4.4. 模型可解释性分析

本研究通过 SHAP 汇总图(见图 5)，将基于特征选择识别后的显著相关指标所得到的 SHAP 值按照图示进行排序。

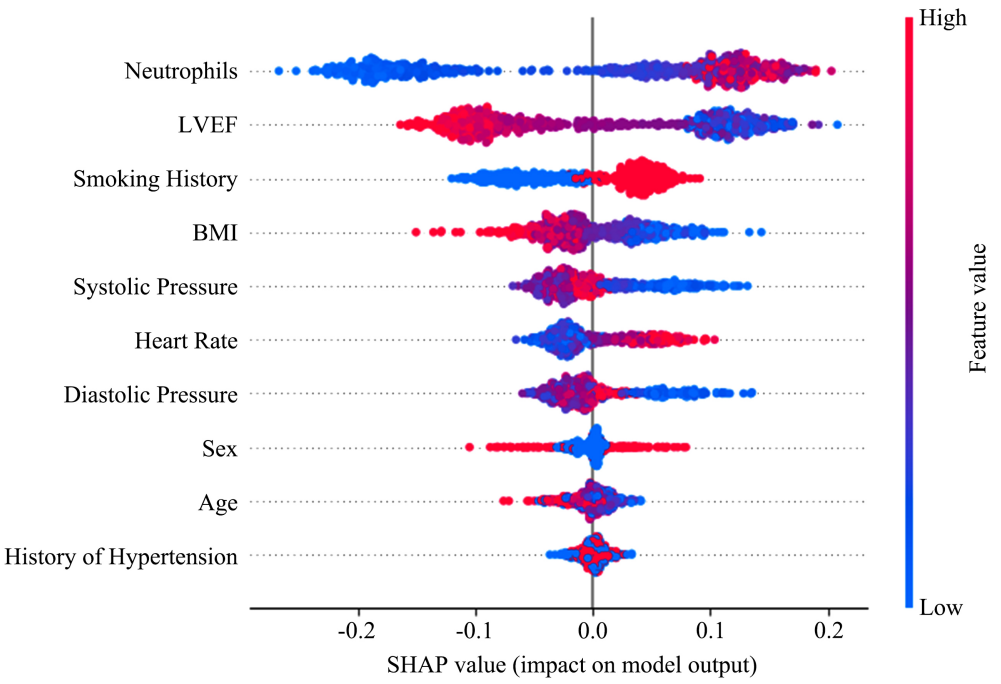


Figure 5. SHAP analysis results
图 5. SHAP 分析结果图

图 5 是按特征变量对模型预测的影响降序显示特征变量。横轴对应于单个样本的预测结果下，各个特征变量的 SHAP 值，该值衡量每个特征变量对模型输出的影响。颜色渐变表示要素的原始值，红色表示较高值，蓝色表示较低值[14]。

可以从图 5 看出连续变量嗜中性粒细胞、入院心率对模型呈正向影响，LVEF、BMI、收缩压、舒张压对模型呈负向影响。对于年龄来说，检查数据发现年轻人患心梗只是个例，因此年龄与目标变量之间并没有很强的关联性。对于分类变量有吸烟史容易患心梗，性别和高血压病史而言，并没有患心梗或心绞痛的倾向。

2.4.5. 模型决策

本研究通过 SHAP 决策图(图 6)，将模型最终预测结果的决策过程进行可视化，以便观察模型决策的准确性与指标的重要性。

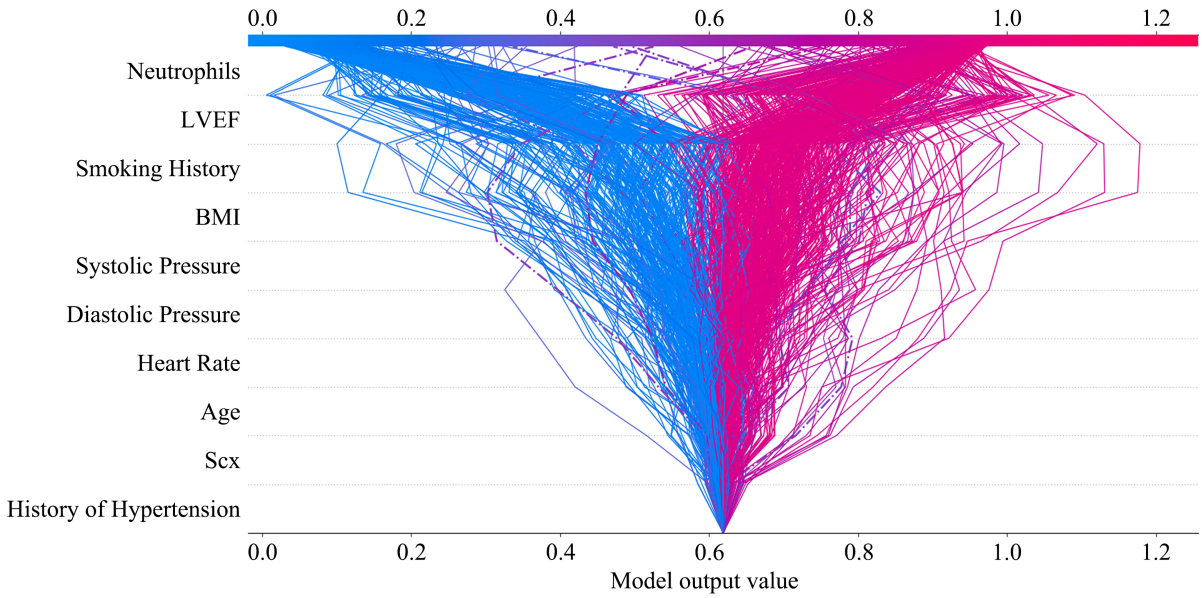


Figure 6. Model decision diagram
图 6. 模型决策图

图 6 红色代表正影响、蓝色代表负影响。从上往下，依次按照特征重要性排序。从最下方开始，通过每个特征的影响，SHAP 值不断加减，以得到最终的预测值。其中实线代表预测正确，虚线代表预测错误。可以看出，该模型的最终预测结果优秀。

3. 讨论

急性心肌梗死(AMI)是心血管疾病的最严重形式，诊断和治疗必须及时，否则可能引发不可逆的后果[3]。然而，由于以下三个方面，AMI 很容易被漏诊：不典型的临床症状、临床评估的低估，和生物标志物血清浓度的不合时宜波动。且目前指南对于急性心肌梗死患者的快速评估多以排除为主，而疾病的阳性诊断方法仍有提高的空间。本研究利用山西医科大学附属心血管病医院胸痛患者的数据进行分析，所有患者的疾病诊断都严格遵循医疗标准研究分类，并通过标准化实验室技术测试其病理数据，以确保机器学习模型中变量的准确性。本研究针对缺失的数据集会使数据中蕴涵的确定性成分较难把握，导致不可靠的输出，我们使用 KNN 填充来解决这个问题。特征工程采用了 Relief 算法初步选择和遗传算法进一

步精确选择相关性大的 10 个特征，采用选择出的 10 个特征构建了 7 个经典的机器学习模型，最终建立了 ACS-Stacking 预测模型。我们确定 ACS-Stacking 预测模型在通过 10-fold 交叉验证比较后是最优的，模型的高召回率表明它具有良好的性能。

随着心脏病患者人数的逐年增加，对现有的医疗系统也产生了巨大的压力，如何利用机器学习进行疾病方面的诊疗优化成为新的临床问题。例如在急性心肌梗死的快速诊断方案，有研究依托高敏肌钙蛋白[14]及心电图等传统诊断模型，进行机器学习相关模型的构建和验证，并明显提高了现有的诊断率。本研究构建的可解释的冠状动脉疾病融合预测模型则是依据临床简单易得的常见指标，进行急性心肌梗死患者的筛选，而并不依赖于患者是否具有心肌缺血症状、体征或心肌缺血导致的异常客观检测值。可解释的冠状动脉疾病融合预测模型可以结合高敏肌钙蛋白及心电图等常规检测，从而提高急性心肌梗死的检出率。

这种可解释的模型是解决临床医生对将 ML 输出纳入临床决策的犹豫不决的关键第一步[15]-[17]。未来的研究应侧重于在外部人群中前瞻性地应用这些模型，以进一步表征其预测性能。

4. 研究展望

未来研究将扩充高血压病史患者数据，以优化预测其患急性心肌梗死(AMI)或不稳定心绞痛(UA)倾向。同时，计划实现对患者主观描述的采集和分析，以完善特例样本在临床实际中所呈现出的异常情况。最后，以验证本研究在其他地区的适用性，扩充数据，实现基于多源多中心数据的模型优化。我们的团队将解决上述存在的问题，并通过未来的研究努力建立更好的 ACS 预测模型，以帮助医生快速准确地判断相应的预后，以改进冠状动脉疾病患者个性化治疗和管理。

参考文献

[1] 徐伟. 急性心肌梗死患者 IIb/IIIa 受体的表达及受体拮抗剂对其血小板释放反应作用的研究[D]: [硕士学位论文]. 芜湖: 皖南医学院, 2014.

[2] 中国心血管健康与疾病报告编写组, 胡盛寿, 王增武. 《中国心血管健康与疾病报告 2022》概要[J]. 中国介入心脏病学杂志, 2023, 31(7): 485-508.

[3] Anderson, J.L. and Morrow, D.A. (2017) Acute Myocardial Infarction. *New England Journal of Medicine*, **376**, 2053-2064. https://doi.org/10.1056/nejmra1606915

[4] Thygesen, K., Alpert, J.S., Jaffe, A.S., Chaitman, B.R., Bax, J.J., Morrow, D.A., *et al.* (2018) Fourth Universal Definition of Myocardial Infarction (2018). *European Heart Journal*, **40**, 237-269. https://doi.org/10.1093/eurheartj/ehy462

[5] Al-Zaiti, S.S., Martin-Gill, C., Zègre-Hemsey, J.K., Bouzid, Z., Faramand, Z., Alrawashdeh, M.O., *et al.* (2023) Machine Learning for ECG Diagnosis and Risk Stratification of Occlusion Myocardial Infarction. *Nature Medicine*, **29**, 1804-1813. https://doi.org/10.1038/s41591-023-02396-3

[6] Kumar Dubey, A., Choudhary, K. and Sharma, R. (2021) Predicting Heart Disease Based on Influential Features with Machine Learning. *Intelligent Automation & Soft Computing*, **30**, 929-943. https://doi.org/10.32604/iasc.2021.018382

[7] Saw, M., Saxena, T., Kaithwas, S., Yadav, R. and Lal, N. (2020). Retracted: Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning. 2020 *International Conference on Computer Communication and Informatics* (*ICCCI*), Coimbatore, 22-24 January 2020, 1-6. https://doi.org/10.1109/iccci48352.2020.9104210

[8] Rahman, A. and Tabassum, A. (2020) Model to Assess the Factors of 10-Year Future Risk of Coronary Heart Disease among People of Framingham, Massachusetts. *International Journal of Public Health Science* (*IJPHS*), **9**, 259-266. https://doi.org/10.11591/ijphs.v9i3.20469

[9] 李董, 迟家俊, 相博, 等. 基于 SMOTE 和 KNN 的石油数据缺失填充算法[J]. 数学的实践与认识, 2019, 49(17): 187-195.

[10] Meng, X., Rosenthal, R. and Rubin, D.B. (1992) Comparing Correlated Correlation Coefficients. *Psychological Bulletin*, **111**, 172-175. https://doi.org/10.1037//0033-2909.111.1.172

[11] 朱海洋. 基于多领域脑电特征与融合特征选择的情感识别研究[D]: [硕士学位论文]. 长春: 吉林大学, 2024.

[12] Katoch, S., Chauhan, S.S. and Kumar, V. (2020) A Review on Genetic Algorithm: Past, Present, and Future. *Multimedia Tools and Applications*, **80**, 8091-8126. https://doi.org/10.1007/s11042-020-10139-6

[13] (2022) Correction: Primer on Binary Logistic Regression. *Family Medicine and Community Health*, **10**, e001290corr1.

[14] Doudesis, D., Lee, K.K., *et al*. (2023) Machine Learning for Diagnosis of Myocardial Infarction Using Cardiac Troponin Concentrations. *Nature Medicine*, **29**, 1201-1210.

[15] Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A., *et al*. (2021) Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences*, **11**, Article 5088. https://doi.org/10.3390/app11115088

[16] Bussone, A., Stumpf, S. and O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. 2015 *International Conference on Healthcare Informatics*, Dallas, 21-23 October 2015, 160-169. https://doi.org/10.1109/ichi.2015.26

[17] Tonekaboni, S., *et al*. (2019) What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Machine Learning for Healthcare Conference*, Ann Arbor, 8-10 August 2019, 359-380.