

基于机器学习和不平衡数据处理的AMI-MC预测新方法

李刘琪^{1*}, 刘静祎^{2*}, 王晓云^{1#}

¹太原理工大学数学学院, 山西 太原

²山西心血管医院心脏病学部, 山西 太原

收稿日期: 2025年3月28日; 录用日期: 2025年4月23日; 发布日期: 2025年4月29日

摘要

目的: 本研究目的是寻求处理AMI-MC这类高度不平衡数据的方法以及快速准确预测模型的构建。背景: 急性心梗后机械并发症是严重威胁患者生命的疾病, 早识别、早处理是改善患者预后的关键。但实际工作中对于该类患者的识别往往存在滞后性, 因此快速预测高危人群是亟待解决的临床问题。方法: 基于山西省心血管病医院的509例急性心肌梗死患者数据, 通过多种机器学习分类器, 建立了预测模型, 并进行了系统的模型训练和评估。此外, 针对数据集中存在的不平衡问题, 我们采用了先进的不平衡处理策略, 以优化模型的预测性能。结果: 509例AMI-MC患者数据中, 仅有45例MC数据, 经不平衡处理后, 各算法准确率以及召回率等评价指标得分均得到了明显提升。最高准确率(0.987)、最高敏感性(0.999)、最高精确率(0.982)、最高F1分数(0.991)、最高敏感性(0.967)。并且各个模型在处理数据上的AUROC得分较原数据而言均有明显提升(0.69521→0.96495)。LVEF是最影响机械并发症的特征, 结合其他6个重要特征构建模型, 在保持高召回率(1.0)的同时, 准确率也达到了0.941、敏感性为0.935、F1分数为0.966。结论: 本研究为提早预测急性心肌梗死的机械并发症提供了新的视角和方法, 发现了LVEF在其中的关键影响, 以及不平衡处理在提升模型性能中的显著作用。

关键词

急性心肌梗死, 机械并发症, 不平衡处理, 机器学习

A New Method for AMI-MC Prediction Based on Machine Learning and Unbalanced Data Processing

Liuqi Li^{1*}, Jingyi Liu^{2*}, Xiaoyun Wang^{1#}

¹School of Mathematics, Taiyuan University of Technology, Taiyuan Shanxi

*共第一作者。

#通讯作者。

文章引用: 李刘琪, 刘静祎, 王晓云. 基于机器学习和不平衡数据处理的 AMI-MC 预测新方法[J]. 应用数学进展, 2025, 14(4): 881-891. DOI: 10.12677/aam.2025.144213

Abstract

Objective: The purpose of this study was to seek methods for handling highly unbalanced data such as AMI-MC and the construction of a fast and accurate prediction model. **Background:** Mechanical complications after acute cardiac infarction are serious life-threatening diseases, and early recognition and management are the key to improving patients' prognosis. However, in practice, there is often a lag in the identification of such patients, so rapid prediction of high-risk groups is a clinical problem that needs to be solved urgently. **Methods:** Based on the data of 509 patients with acute myocardial infarction in Shanxi Provincial Cardiovascular Disease Hospital, a prediction model was established by using various machine learning classifiers and systematic model training and evaluation was performed. In addition, for the imbalance problem existing in the dataset, we adopted an advanced imbalance processing strategy to optimize the prediction performance of the model. **Results:** In the data of 509 AMI-MC patients, there are only 45 cases of MC data. After imbalance processing, the accuracy, recall rate, and other evaluation metrics of each algorithm were significantly improved. The highest accuracy (0.987), the highest sensitivity (0.999), the highest precision (0.982), the highest F1 score (0.991), and the highest sensitivity (0.967). Moreover, the AUROC scores of various models on the processed data have significantly improved compared to the original data (0.69521→0.96495). LVEF is the feature that most affects mechanical complications, and the model was constructed by combining the other six important features, which achieved an accuracy of 0.941, a sensitivity of 0.935, and an F1 score of 0.966, while maintaining a high recall (1.0). **Conclusion:** This study provides new perspectives and methods for early prediction of mechanical complications in acute myocardial infarction, identifies the key influence of LVEF in this, and the significant role of imbalance treatment in enhancing model performance.

Keywords

Acute Myocardial Infarction, Mechanical Complications, Imbalanced Treatment, Machine Learning

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

急性心肌梗死(acute myocardial infarction, AMI)是心血管疾病的急危重症,随着再灌注治疗以及基层医疗保健的发展,AMI的发病率和病死率有了明显下降,且机械并发症(mechanical complications after acute myocardial infarction, AMI-MC)仍然是AMI死亡的主要原因之一[1]-[3]。研究显示,虽然AMI-MC整体发生率低,约为1‰~3‰[4],但却显著影响患者预后,存活患者整体存活率约40%[5]。AMI-MC具有进展迅速、治疗效果欠佳的困局,且实际工作中对于该类患者的识别往往存在滞后性;此外,还有研究发现其对医院资源利用也存在重大风险相关[6]。因此如何快速识别高危患者、协助医生高效分配临床资源即成为亟待解决的临床问题。

近年来,机器学习在医学领域的应用日益广泛,其在处理复杂数据和识别非线性模式方面的独特优势使其成为预测疾病进展和并发症的有力工具。但由于AMI-MC较低的发病率,使得其临床样本量无法

满足平衡数据要求。而在机器学习任务中, 使用平衡的数据集对于模型的训练和性能评估至关重要, 数据集的不平衡性可能导致模型对多数类别预测准确度较高, 而对少数类别的预测能力不足。这在医学数据中则表现为将病人误判为正常人, 这会导致致命性失误。一些经典方法已经被用于处理不平衡数据处理, 这些方法背后的主要动机不仅仅是为了平衡训练数据, 还要去除位于决策边界错误一侧的噪声样例[7][8]。但目前这些处理方法在低发病率数据中(也即高度不平衡数据集中)存在平衡度不够高、处理后数据在模型训练时不够理想的缺点[9]-[11]。有研究证明在心律失常领域的高度不平衡数据上, 应用 SmoteENN 算法[12][13]处理数据, 再经过多种机器学习算法训练后呈现出更好的效能[14]。但在 AMI-MC 数据处理中, SmoteENN 算法是否仍有效不得而知。故, 本研究拟采用采样 + 机器学习的方式进行数据平衡处理, 这可能有助于不平衡数据更好的分类。

基于此, 我们希望通过机器学习探索并建立一种高效的预测急性心肌梗死后机械并发症(AMI-MC)模型; 同时基于此类小样本疾病的高度不平衡数据集, 采用 Smote 算法对 AMI-MC 数据进行不平衡处理, 以便提高机器学习精度。我们希望通过不同采样策略与机器学习算法的组合, 能找到最佳的解决方案, 从而为医生提供更简洁、迅速的辅助诊断手段。

2. 方法

2.1. 数据分布和处理

本研究使用的数据由山西省心血管病医院收集, 患者的一般临床资料包括生理生化等指标 56 项, 纳入 509 例 AMI 患者, 其中 45 例发生机械并发症(标记为正样本、阳性); 454 例患者不发生机械并发症(标记为负样本、阴性)。收集的数据中不存在缺失值。异常值检测如图 1。

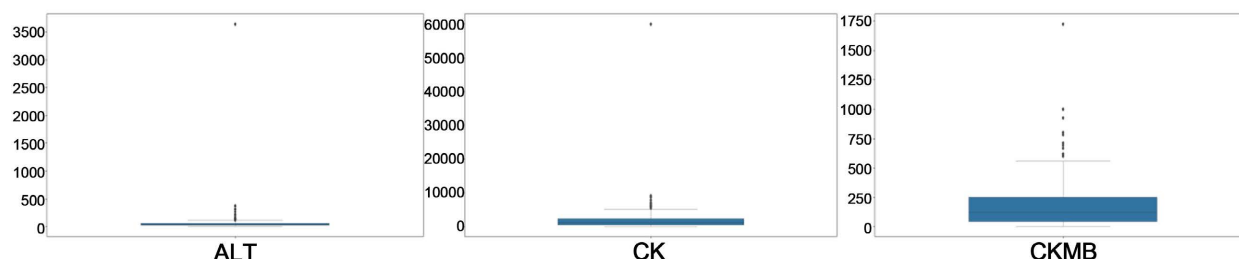


Figure 1. Box plot of ALT, CK, CKMB. The blue box represents the majority distribution interval of the data, and the black dots represent outliers

图 1. ALT、CK、CKMB 箱形图。蓝色箱形代表数据的多数分布区间, 黑色点代表离群点

图 1 表明, ALT、CK、CKMB 各存在一例数值过高的数据, 考虑到数据量较少, 我们使用均值填充。

经过异常值处理、去除共线性(相关系数 > 0.75 的两个变量被视作具有多重共线性)、归一化处理, 最终数据集包括 509 个样本(45:454), 特征数减少为 50 个, 其中连续指标 35 个、离散指标 15 个。

2.2. 不平衡处理

考虑到数据集分布, 我们在实验中特别关注了数据集的平衡性, 并采取了相应的措施来处理。对于训练集, 我们使用多种不平衡数据处理方法进行不平衡处理。

2.3. 模型构建

利用多种机器学习分类器, 包括逻辑回归、朴素贝叶斯估计、支持向量机、树模型和 K 近邻, 进行模型训练。树模型包括自适应提升算法(Adaboost)[15], 随机森林(RF)和极端梯度增强(xgboost)[16]。

实验中所使用的所有分类器均来自 Python 的 sklearn 工具箱, 确保了实验的一致性和可重复性。

为了客观评估模型的预测性能, 我们采用了交叉验证方法。为了更全面地评估模型的整体性能, 我们通过计算准确率、召回率(TPR)、敏感性(TNR)、精确率(PPV), F1 分数、ROC 曲线下面积(AUROC)这些关键指标, 从而更准确地判断不同采样策略在不同分类器下的性能表现。

3. 结果

3.1. 不平衡处理结果

我们通过十折交叉验证来检验不平衡数据处理后的数据在各个模型上的性能, 评分结果如表 1 及图 2 所示。

表 1 表明, 原始数据训练的模型在面临两类样本比率极度不平衡(40:418)的问题时, 普遍展现出不佳的性能。具体来说, 各个模型的准确率最高才达到了 0.917, AUC 值最高为 0.76315, 与此同时, 这些模型的召回率极低, 甚至达到了 0, 而敏感性却异常高, 有时接近或达到 1。KNN 分类器中精确率由于召回率为 0, 敏感性为 1, 导致计算精确率的分子为 0 从而无法计算(NAN)。这种表现显然不符合实际需求, 因此, 需要我们进一步解决不平衡数据的处理问题。

Table 1. Cross-validation results of different processing methods in different models
表 1. 不同处理方法在不同模型中的交叉验证结果

算法	模型	ACC	TPR	TNR	PPV	F1	AUC
None	adaboost	0.891	0.225	0.954	0.258	0.239	0.67999
	randomtree	0.917	0.050	1.000	0.200	0.080	0.67911
	xgboost	0.915	0.150	0.988	0.350	0.207	0.70968
	SVM	0.900	0.100	0.976	0.150	0.119	0.68600
	gaussianNB	0.798	0.475	0.829	0.242	0.313	0.76315
	KNN	0.913	0.000	1.000	NAN	0.000	0.66641
	Logistic	0.897	0.125	0.971	0.233	0.154	0.68213
SMOTE	adaboost	0.946	0.966	0.926	0.930	0.947	0.98433
	randomtree	0.977	0.988	0.967	0.967	0.978	0.99869
	xgboost	0.970	0.983	0.957	0.959	0.971	0.99486
	SVM	0.839	0.928	0.749	0.790	0.852	0.87604
	gaussianNB	0.810	0.926	0.693	0.757	0.831	0.91994
	KNN	0.709	0.993	0.426	0.635	0.774	0.94479
	Logistic	0.833	0.876	0.790	0.810	0.840	0.88412
SMOTEENN	adaboost	0.932	0.967 0.937	0.855	0.938	0.952	0.95128
	randomtree	0.987	0.999	0.957	0.982	0.991	0.99961
	xgboost	0.969	0.993	0.914	0.963	0.978	0.99388
	SVM	0.944	0.983	0.856	0.939	0.960	0.94031
	gaussianNB	0.913	0.974	0.777	0.908	0.939	0.93416
	KNN	0.813	0.998	0.401	0.789	0.881	0.99002
	Logistic	0.937	0.986	0.829	0.928	0.956	0.94540

不同处理方法和不同模型组合达到的最优性能加粗表示。

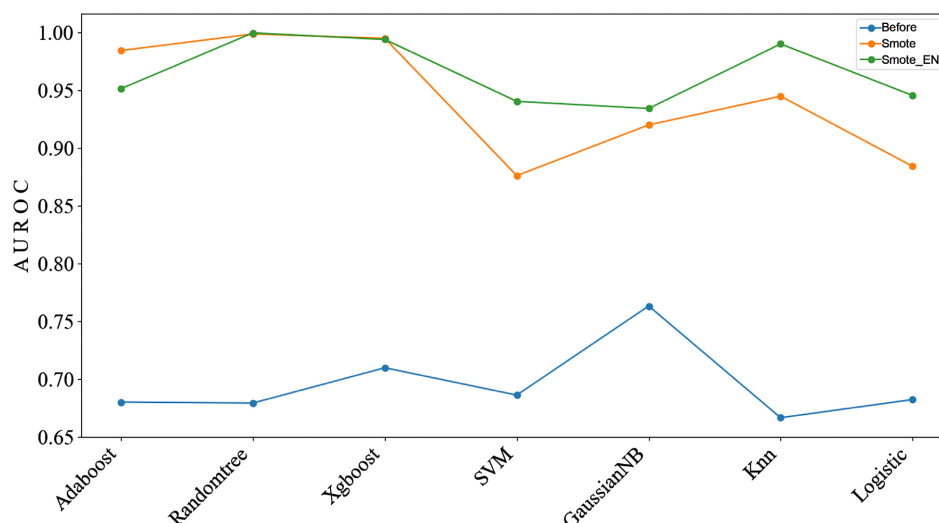


Figure 2. AUROC scores of different processing methods on different models. The blue line represents the raw unprocessed data across model scores; the orange line represents Smote-processed; the green line represents SmoteENN-processed

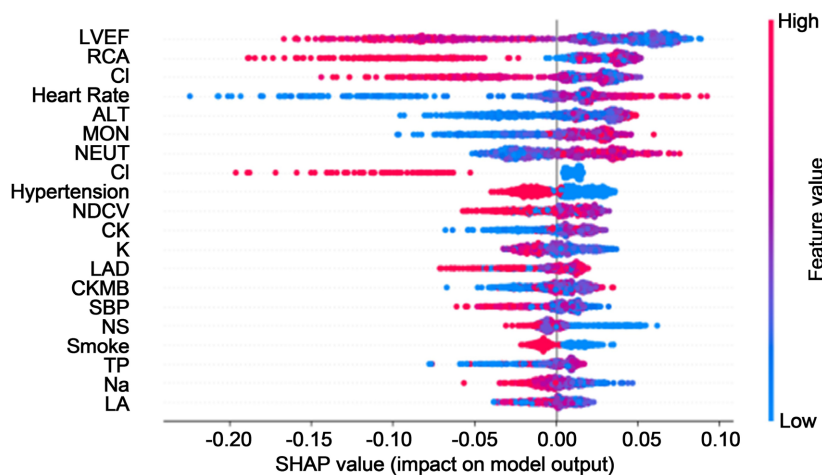
图 2. 不同处理方式在不同模型上 AUROC 得分。蓝色线代表未处理原始数据在各个模型得分；橘色线代表经 Smote 处理；绿色线代表经 SmoteENN 处理

此外,从表 1 中可以看出,在经不平衡处理后,各算法准确率以及召回率等评价指标得分均得到了明显提升。其中 SmoteENN 和 RF 联合使用的准确率最高(0.987)。同时具有最高的敏感性(0.999)、精确率(0.982)以及最高 F1 分数(0.991)。经 Smote 处理后数据在 RF 分类器上取得了最高的敏感性(0.967)。从表 1 可以看出,SMOTE + RF 和 SmoteENN + RF 的组合是所有模型中综合性能最好的。

图 2 给出了 AUC 得分比较,各个模型在处理数据后的 AUROC 得分较原数据而言均有明显提升,其中 SmoteENN 和 RF 算法的组合达到了最高的 AUC 值(0.99961)这表明经不平衡处理后的数据在各个模型上的性能都得到了提升,证明了不平衡处理方法的有效性,为之后的机器学习算法构建了数据基础。

3.2. 模型构建结果

基于相关系数以及 SHAP 算法[17],我们对上述最优模型进行特征重要性排序。特征排序结果如图 3、图 4。



(a)

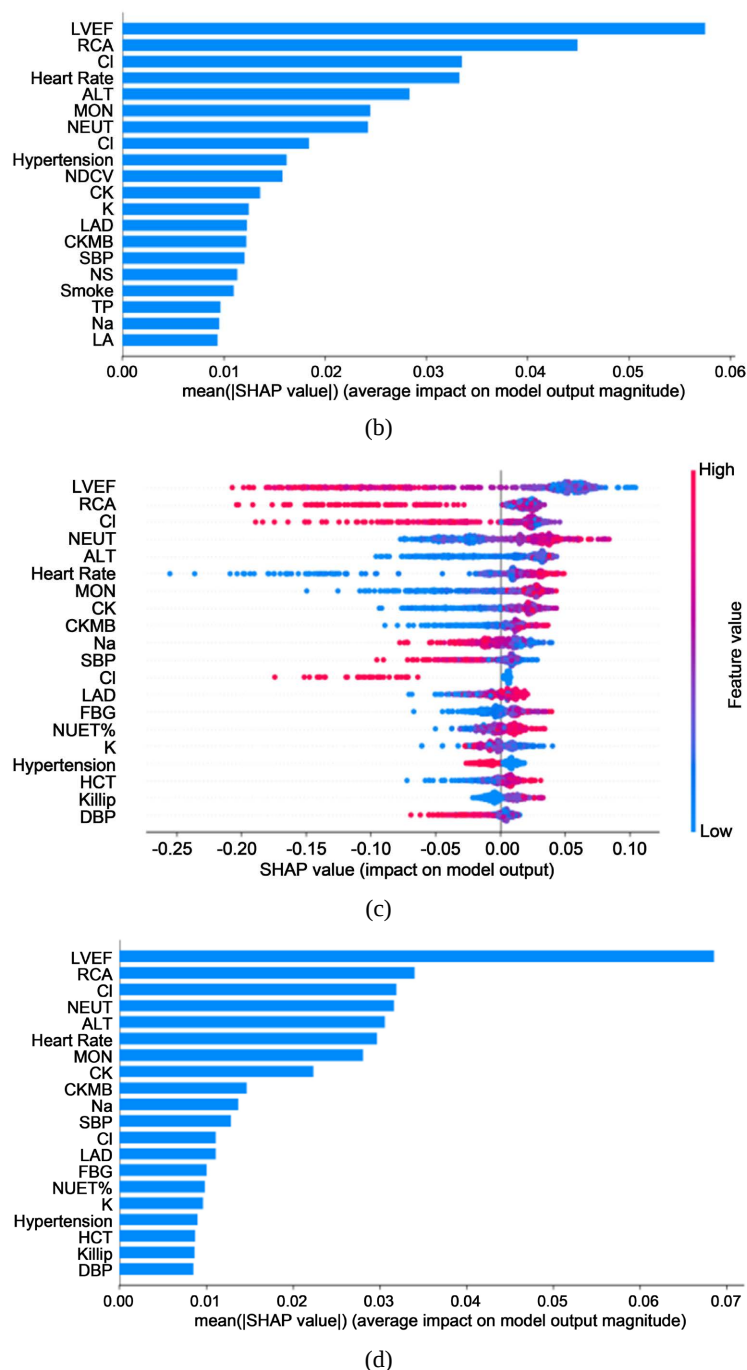


Figure 3. Top twenty highest features of SHAP algorithm. The plot (a) and plot (c) represent density scatter plots of all samples. The vertical view displays the feature importance in order of decreasing value, from top to bottom. The horizontal view presents the color bar, with the color indicating the magnitude of the value of the feature. Red indicates a high value of the feature, while blue indicates a low value of the feature. Wide areas indicate a large number of samples clustered in that region. The LVEF samples exhibit a pronounced concentration in the region exhibiting a positive shape value. As illustrated in plot (b) and plot (d), according to the importance of features, it decreases from top to bottom, and the more important features belong to the longer color bar

图 3. SHAP 算法最高前二个特征。(a)和(c)为密度散点图, 代表了所有样本。纵向来看, 按照特征重要性从上往下依次减弱。横向来看, 每个颜色条中, 颜色表示特征的值的, 红色表示特征值高, 蓝色表示特征值低。宽的地方表示有大量的样本聚集。图中 LVEF 的样本大量在 shap value 为正的区域内聚集。(b)、(d)图表示按照特征重要性从上往下依次减弱, 越重要的特征所属颜色条越长

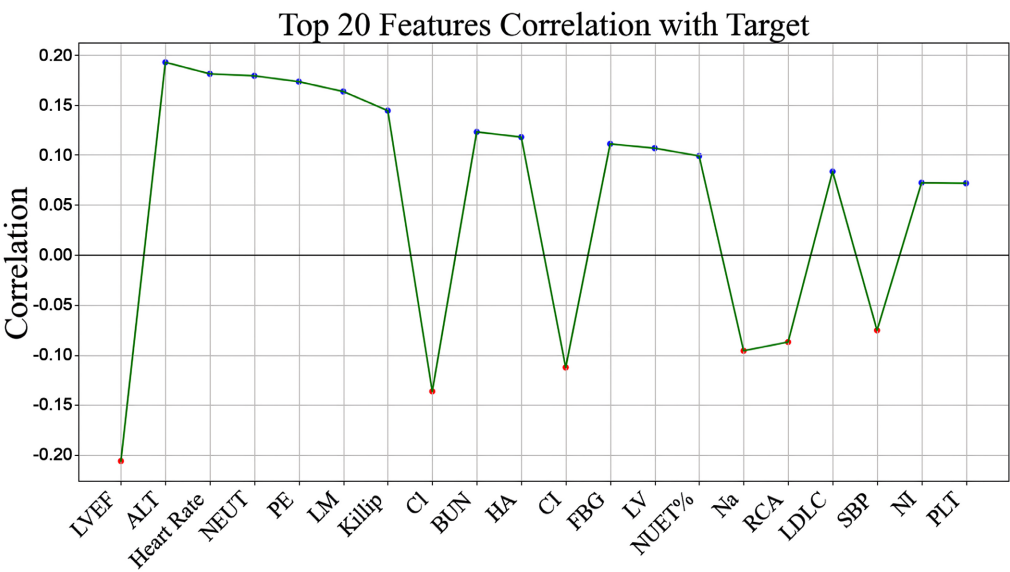


Figure 4. Top twenty features with the highest correlation coefficients. The presence of a feature is indicated by a red dot if it is negatively correlated with the endpoint event, and by a blue dot if it is positively correlated
图 4. 相关系数最高前二十个特征。与终点事件负相关的特征标记为红点、正相关为蓝点

结果表明, LVEF 在两种排序方法中都具有最高的重要性, 这意味着它是最影响机械并发症的特征。此外, 两个模型的 SHAP 排序结果的前 10 个特征中有 7 个是相同的, 包括相关度排名前 10 个特征中的 5 个。这表明特征排序方法在同一数据集上结果相似。然而, SHAP 排序的前 20 个特征几乎没有离散型特征, 这是由于 SHAP 算法可能会对基数较大的特征(如连续特征)表现出偏好。因此, 它不倾向于将离散特征视为重要特征。对于相关系数排序, 我们可以看到心包积液、killip 分级是与机械并发症最相关的两个离散特征。

为了构建模型, 我们在排名结果基础上, 选择每个排名方法的前 K 个特征来训练模型。结果如表 2。

Table 2. Cross-validation results of Top k feature training model
表 2. Top k 特征训练模型交叉验证结果

模型	方法	K	ACC	TPR	TNR	F1	AUC
SMOTE + RF	SHAP	5	0.909	0.933	0.885	0.911	0.97005
		10	0.938	0.955	0.921	0.939	0.98564
		20	0.964	0.974	0.955	0.964	0.99577
	Correlation	5	0.853	0.900	0.806	0.859	0.93758
		10	0.908	0.938	0.878	0.911	0.97092
		20	0.959	0.976	0.943	0.960	0.99304
SMOTEENN + RF	SHAP	5	0.949	0.981	0.877	0.964	0.98185
		10	0.962	0.990	0.898	0.973	0.99223
		20	0.982	0.998	0.946	0.987	0.99843
	Correlation	5	0.924	0.979	0.801	0.947	0.96099
		10	0.940	0.979	0.855	0.958	0.98394
		20	0.967	0.993	0.908	0.977	0.99548

排名前 5 特征在模型上达到的最优性能加粗表示。

表 2 展示了使用两种特征排序方式以及两种模型进行训练所报告的结果。相较于相关系数, SHAP 算法选择的前 5 个特征在 SHAP 算法下达到了更好的性能(ACC: 0.949, TPR: 0.981, TNR: 0.885, F1: 0.964, AUC: 0.98185)。

将 SHAP 算法和相关系数结合起来, 选取在两个模型的 SHAP 排序中都排名前 10 的特征: 射血分数(LVEF)、入院心率、右冠、氯、ALT、单核细胞; 以及在相关系数排序中最重要的两个离散特征, 即心包积液和 Killip 分级作为主要特征, 利用上述采样策略 + 机器学习框架, 以选取的 7 个特征构建模型, 模型的评价结果如表 3 所示。两个模型均表现出较好的精准性和敏感性。

Table 3. Evaluation results of seven feature construction models
表 3. 7 个特征构建模型评价结果

模型	ACC	TPR	TNR	PPV	F1	AUC
Smote_RF	0.927	0.952	0.902	0.908	0.929	0.98163
SmoteENN_RF	0.955	0.988	0.882	0.947	0.968	0.99071

3.3. 真实世界模拟

真实世界患者的数据仍然遵循原始数据集的分布, 患病率 5.24%。受限于本研究使用的数据集的规模, 为了充分利用数据, 我们使用利用分层抽样方法抽取的测试集(51 个样本, 5 例患者), 对真实世界进行模拟。模型结果如表 4 所示。

Table 4. Evaluation results of the model on simulated data sets
表 4. 模型在模拟数据集上的评价结果

模型	ACC	TPR	TNR	PPV	F1
Smote + RF	0.941	1.0	0.935	0.625	0.966
SmoteENN + RF	0.686	1.0	0.652	0.263	0.789

最优性能加粗表示。

通过表 4 我们可以看出, 两个模型均达到了 1.0 的召回率, 说明两个模型都可以对正样本进行正确的预测, 然而 SmoteENN + RF 的准确率仅有 0.686、敏感性也仅有 0.652、F1 分数 0.789; 说明该模型存在过拟合问题, 泛化能力不足。相反, Smote + RF 在保持高召回率的同时, 准确率也达到了 0.941、敏感性为 0.935、F1 分数为 0.966; 另外, 其精确率较低(0.625), 这是因为模型仅将三个负类样本预测为了正类, 正类全部正类预测, 故 $PPV = 5/8 = 0.625$ 。说明该模型在能够准确预测正样本(这是我们最需要的)的同时, 保证了正确预测负类样本的能力, 泛化性能良好。

经过全面评估, 我们决定采用 Smote + RF 模型作为预测机械并发症的首选模型。该模型在处理 AMI-MC 数据集固有的不平衡问题上表现卓越, 其高召回率、准确率、敏感性以及 F1 分数均证明了其有效性。我们深信, 我们构建的 AMI-MC 不平衡处理模型非常适合临床应用, 并能作为 AMI-MC 的可靠预测工具。

4. 讨论

尽管当前的医疗实践已经显著降低了急性心肌梗死机械并发症的发生率, 依据近些年研究估测, STEMI 后, 有 0.27%~0.91% 的患者出现了机械性并发症, 其中乳头肌断裂(PMR)、心室游离壁破裂(FWR)和室间隔穿孔(VSR)的发生率分别为 0.05%~0.26%、0.01%~0.52%和 0.17%~0.21%。过去的 20 年里, 机械

并发症患者的死亡率并没有明显下降, 住院死亡率比没有机械并发症的患者高出 4 倍以上[18]。所以尽管机械并发症发生率很低, 但因其超高的死亡率, 故仍然是决定心梗预后的重要因素。这表明, 尽管我们在急性心肌梗死的治疗策略上取得了进步, 但在预测和识别心梗后机械并发症方面仍面临挑战。但关于 AMI-MC 事件的研究很少, 建立 AMI 患者准确、快速的风险预测模型对预防 AMI-MC 至关重要。

GRACE 和 TIMI 风险评分等临床预测模型已被广泛应用于心肌梗死临床实践, 主要用于预测不良事件的发生[19][20]。但这些风险评分均无法预测 AMI-MC, 故探索新的预测工具和方法显得尤为迫切。本研究通过对山西省心血管病医院 509 例急性心肌梗死患者的数据进行深入分析, 发现 LVEF (左室射血分数) 与急 AMI-MC 之间存在显著的相关性。我们特别关注了数据高度不平衡问题, 采取了相应的处理措施。通过对比各种组合的性能和效率, 我们成功识别出了一套适合处理此类数据集的方法, 并基于这一方法结合 SHAP 可解释性算法, 构建了一个快速、准确的 AMI-MC 预测模型(Smote + RF MC Model, SRMC)。通过对真实世界的模拟验证了模型的鲁棒性(ACC: 0.941, TPR: 1.0)。预测模型的构建, 不仅能辅助医生进行 MC 的预测、从而高效利用有限的临床资源; 还为处理相似数据集提供了思路。

本研究发现, LVEF 是预测 AMI-MC 发生的重要生理指标, 且在临床简单易得, 对于低 LVEF 而言, 过去更多地关注其与心力衰竭之间的关系[21][22]。本研究则提示, 在 LVEF 较小, 还需要同时关注心梗后机械并发症。低 LVEF 可能与就诊时间晚、心肌损伤面积大、缺血时间长相关。而这些也均是机械并发症的危险因素。此外, 我们的研究发现, 对 LVEF 进行动态监测, 也能预测机械并发症的发生。因此, 在临床实践中, 监测 LVEF 的变化对于预防和治疗 AMI-MC 具有重要意义。

在数据不平衡问题上, 过去的一些简单采样或使用平衡分类器等不平衡处理方式在一般不平衡数据集上表现良好, 但在类似 AMI-MC 这样低发病率的高度不平衡数据集上却存在着这样那样的问题, 要么训练效果不够良好(ACC: 0.843, TPR: 0.700), 要么泛化能力不足(训练测试结果: ACC: 0.999~0.922、TPR: 1.000~0.200)。本研究使用的两种数据处理方法在保持了高效能训练的同时, 增强了模型的泛化能力。

在模型构建方面, 我们采用了多种机器学习分类器, 并结合了不平衡数据处理方法, 以提高模型的预测性能。实验结果表明, 我们构建的模型在预测 AMI-MC 方面具有较高的准确性(0.941)、灵敏度(0.935)和特异性(1.000)。这一结果不仅验证了 LVEF 作为预测指标的有效性, 也展示了 SRMC 模型在处理复杂医疗数据方面的优势。

虽然通过对真实世界的模拟证明了所构建模型的性能, 但该模型仅在我们的数据上得到了验证, 需要进一步的验证来评估其在大规模研究中的预测性能。另外, 我们的数据量较少, 未来将会采用更多的数据来检验并进行模型优化。

本研究不仅为解决 AMI-MC 数据集的不平衡问题提供了新思路, 还为构建高效、准确的预测模型奠定了坚实基础。且此类不平衡数据的处理思路, 也为其他类似的低发病率但高死亡率的疾病预测模型的构建提供了实践依据, 如以本研究为基础, 进行主动脉夹层、电风暴等更小众却更凶险的疾病预测模型的构建。所以, 本研究不仅为 AMI-MC 的临床决策提供了简捷而有力的辅助诊断支持, 也为未来相关研究提供了新的思路和方法。

基金项目

山西省回国留学人员科研资助项目(2022-074), 医学数据驱动多模态融合急性心肌梗死数学建模及分析。

参考文献

- [1] Puerto, E., Viana-Tejedor, A., Martínez-Sellés, M., Domínguez-Pérez, L., Moreno, G., Martín-Asenjo, R., *et al.* (2018)

- Temporal Trends in Mechanical Complications of Acute Myocardial Infarction in the Elderly. *Journal of the American College of Cardiology*, **72**, 959-966. <https://doi.org/10.1016/j.jacc.2018.06.031>
- [2] Birnbaum, Y., Fishbein, M.C., Blanche, C. and Siegel, R.J. (2002) Ventricular Septal Rupture after Acute Myocardial Infarction. *New England Journal of Medicine*, **347**, 1426-1432. <https://doi.org/10.1056/nejmra020228>
 - [3] Elbadawi, A., Elgendy, I.Y., Mahmoud, K., Barakat, A.F., Mentias, A., Mohamed, A.H., *et al.* (2019) Temporal Trends and Outcomes of Mechanical Complications in Patients with Acute Myocardial Infarction. *JACC: Cardiovascular Interventions*, **12**, 1825-1836. <https://doi.org/10.1016/j.jcin.2019.04.039>
 - [4] Goldsweig, A.M., Wang, Y., Forrest, J.K., Cleman, M.W., Minges, K.E., Mangi, A.A., *et al.* (2018) Ventricular Septal Rupture Complicating Acute Myocardial Infarction: Incidence, Treatment, and Outcomes among Medicare Beneficiaries 1999-2014. *Catheterization and Cardiovascular Interventions*, **92**, 1104-1115. <https://doi.org/10.1002/ccd.27576>
 - [5] French, J.K., Hellkamp, A.S., Armstrong, P.W., Cohen, E., Kleiman, N.S., O'Connor, C.M., *et al.* (2010) Mechanical Complications after Percutaneous Coronary Intervention in ST-Elevation Myocardial Infarction (from Apex-Ami). *The American Journal of Cardiology*, **105**, 59-63. <https://doi.org/10.1016/j.amjcard.2009.08.653>
 - [6] Damluji, A.A., van Diepen, S., Katz, J.N., Menon, V., Tamis-Holland, J.E., Bakitas, M., *et al.* (2021) Mechanical Complications of Acute Myocardial Infarction: A Scientific Statement from the American Heart Association. *Circulation*, **144**, e16-e35. <https://doi.org/10.1161/cir.0000000000000985>
 - [7] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. <https://doi.org/10.1613/jair.953>
 - [8] Kubat, M. and Matwin, S. (1997) Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *International Conference on Machine Learning*, **97**, 179-186.
 - [9] Peng, M., Zhang, Q., Xing, X., Gui, T., Huang, X., Jiang, Y., *et al.* (2019) Trainable Undersampling for Class-Imbalance Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 4707-4714. <https://doi.org/10.1609/aaai.v33i01.33014707>
 - [10] Ling, C.X. and Li, C. (1998) Data Mining for Direct Marketing: Problems and Solutions. *KDD*, **98**, 73-79.
 - [11] Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J. and Napolitano, A. (2010) Rusboost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, **40**, 185-197. <https://doi.org/10.1109/tsmca.2009.2029559>
 - [12] Batista, G.E.A.P.A., Prati, R.C. and Monard, M.C. (2004) A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter*, **6**, 20-29. <https://doi.org/10.1145/1007730.1007735>
 - [13] Wilson, D.L. (1972) Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, **2**, 408-421. <https://doi.org/10.1109/tsmc.1972.4309137>
 - [14] Manju, B.R. and Nair, A.R. (2019) Classification of Cardiac Arrhythmia of 12 Lead ECG Using Combination of SMOTEENN, XGBoost and Machine Learning Algorithms. 2019 9th International Symposium on Embedded Computing and System Design (ISED), Kollam, 13-14 December 2019, 1-7. <https://doi.org/10.1109/ised48680.2019.9096244>
 - [15] Freund, Y. and Schapire, R.E. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55**, 119-139. <https://doi.org/10.1006/jcss.1997.1504>
 - [16] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. <https://doi.org/10.1145/2939672.2939785>
 - [17] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., *et al.* (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2**, 56-67. <https://doi.org/10.1038/s42256-019-0138-9>
 - [18] Gong, F.F., Vaitenas, I., Malaisrie, S.C. and Maganti, K. (2021) Mechanical Complications of Acute Myocardial Infarction. *JAMA Cardiology*, **6**, 341-349. <https://doi.org/10.1001/jamacardio.2020.3690>
 - [19] Tang, E.W., Wong, C. and Herbison, P. (2007) Global Registry of Acute Coronary Events (GRACE) Hospital Discharge Risk Score Accurately Predicts Long-Term Mortality Post Acute Coronary Syndrome. *American Heart Journal*, **153**, 29-35. <https://doi.org/10.1016/j.ahj.2006.10.004>
 - [20] Morrow, D.A., Antman, E.M., Charlesworth, A., Cairns, R., Murphy, S.A., de Lemos, J.A., *et al.* (2000) TIMI Risk Score for St-Elevation Myocardial Infarction: A Convenient, Bedside, Clinical Score for Risk Assessment at Presentation: An Intravenous nPA for Treatment of Infarcting Myocardium Early II Trial Substudy. *Circulation*, **102**, 2031-2037. <https://doi.org/10.1161/01.cir.102.17.2031>
 - [21] Curtis, J.P., Sokol, S.I., Wang, Y., Rathore, S.S., Ko, D.T., Jadbabaie, F., *et al.* (2003) The Association of Left Ventricular Ejection Fraction, Mortality, and Cause of Death in Stable Outpatients with Heart Failure. *Journal of the American*

- College of Cardiology*, **42**, 736-742. [https://doi.org/10.1016/s0735-1097\(03\)00789-7](https://doi.org/10.1016/s0735-1097(03)00789-7)
- [22] Maeder, M.T. and Kaye, D.M. (2009) Heart Failure with Normal Left Ventricular Ejection Fraction. *Journal of the American College of Cardiology*, **53**, 905-918. <https://doi.org/10.1016/j.jacc.2008.12.007>