Pure Mathematics 理论数学, 2026, 16(1), 17-28
Published Online January 2026 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2026.161003

Hans汉斯

GA-RF：基于SHAP的不平衡数据中风识别的优化研究

胡译丹*，高　阳#，尹　畅，过子宽

中国石油大学(北京)理学院，北京

收稿日期：2025年11月25日；录用日期：2025年12月22日；发布日期：2026年1月8日

摘　要

在疾病初筛的场景中，数据失衡会导致分类器偏向多数类的预测偏差，对模型的性能产生影响。因此，选择合适的数据不平衡处理策略与分类器，对改进性能具有关键意义。本文分析不平衡的中风数据集，构建多种实验方案：引入11种数据不平衡处理方法，结合4种机器学习算法对中风患者进行识别(逻辑回归、SVM、CNN、随机森林)。在多组模型的对比中，得到RUS处理后的逻辑回归、SVM与随机森林优于其他方法，并引入PCA降维分析噪声数据。然后，利用PSO、GA、DE、BO对这3个模型进行优化，得到GA-RF的*AUC*为84.18%，*Recall*为91.06%，优势显著。最后，为突破解释性局限，采用SHAP对模型的特征重要性进行分析，得到年龄对中风识别的作用远超其余特征。

关键词

数据不平衡处理，中风，随机森林，优化算法，SHAP

GA-RF: Research on Unbalanced Stroke Data Recognition Based on SHAP

Yidan Hu*, Yang Gao#, Chang Yin, Zikuan Guo

College of Science, China University of Petroleum (Beijing), Beijing

Received: November 25, 2025; accepted: December 22, 2025; published: January 8, 2026

Abstract

In the scenario of primary disease screening, data imbalance will cause the classifier to bias the

*第一作者。
#通讯作者。

文章引用: 胡译丹, 高阳, 尹畅, 过子宽. GA-RF: 基于 SHAP 的不平衡数据中风识别的优化研究[J]. 理论数学, 2026, 16(1): 17-28. DOI: 10.12677/pm.2026.161003

prediction of the majority class, which will have an impact on the performance of the model. Therefore, choosing appropriate data imbalance processing strategies and classifiers is of key significance to improving performance. This article analyzes unbalanced stroke data sets and constructs various experimental plans: eleven data imbalance processing methods are introduced, and four machine learning algorithms are combined to identify stroke patients (LR, SVM, CNN, RF). In the comparison of multiple groups of models, it is found that LR, SVM and RF after RUS processing are better than other methods, and PCA dimensionality reduction was introduced to analyze noise data. Then, the methods of PSO, GA, DE and BO are used to optimize these three models. The *AUC* of GA-RF is 84.18%, and the *Recall* is 91.06%, which has significant advantages. Finally, in order to break through the explanatory limitations, SHAP is used to analyze the feature importance of these models. It is found that the role of age in stroke recognition far exceeds that of other features.

Keywords

Data Imbalance Handling, Stroke, Random Forest, Optimization Algorithm, SHAP

1. 引言

在类别不平衡的数据集中，其中一类别的样本数量明显少于另一类别，这种情况多见于疾病诊断、信用风险评估、网络入侵监测、窃电检测等领域[1]。此类数据会导致标准的分类器产生类别偏差，在分类时更多地关注多数类样本，进而导致对少数类样本的识别性能不足，并且噪声数据可能会对少数类样本的识别产生影响，进一步增加数据分类的困难度。因此，如何对训练集数据进行预处理，是决定不平衡数据分类效果的关键。

数据不平衡处理方法可分为数据层面的过采样方法、欠采样方法，以及算法层面的集成算法。在疾病初筛研究中常存在严重的失衡现象，即患病样本远少于正常样本，国内外学者针对医疗数据的失衡展开众多研究：MATLOOB KHUSHI 等[2]对比 23 种不平衡数据的分类方法与 3 种分类器，在 PLCO 与 NLST 两种肺癌数据集上进行实验，得到过采样方法比欠采样更稳定的结论；Tri Huynh 等[3]引入 ABCL 方法，根据类别概率的自适应，调整一致性损失的目标类别分布，解决医疗数据偏移问题，并与其他方法进行对比，UAR 有了明显提升；Xu Zhaozhao 等[4]结合 SMOTE 与 k 均值聚类，创新性地提出一种 KNSMOTE 过采样方法，并在 8 个 UCI 数据集的对比中灵敏度为 99.84%，特异性指标达到 99.56%；Vinod Kumar 等[5]应用 6 种分类器在对 5 种临床数据集上进行实验，采用多种数据失衡处理方法，得到 SMOTEENN 普遍优于其他技术的结论；Debapriya Banik 等[6]针对医学诊断中影像数据失衡引发的各种问题，分析不同平衡策略，给出最佳的训练方案；Yao Peng 等[7]研究小规模数据集的皮肤病诊断，融合改进版随机增强、多权重损失函数和累计学习策略，使得单一的深度卷积网络的准确率优于多种集成模型。

选择合适的分类模型对问题的研究同样至关重要：Mr.J.A.Jevin 等[8]利用关联规则技术处理分布在不同场所的医疗数据，并利用逻辑回归、SVM、XGBoost 与 RF 四种方法，对心脏病的诊断进行研究，得到随机森林的准确率达到 90%，高于其他算法；Moloud Abdar 等[9]利用集成学习算法 Boosted C5.0 与决策树算法 CHAID 研究肝脏疾病的识别，证得 DB、ALB、SGPT、TB 及 A/G 等指标对肝病预测具有显著影响；Ding Huanfei 等[10]利用正则化回归、LR、RF、决策树与 XGBoost 等方法从 525 名患者中提取与肝病相关的临床参数，其中 RF 的 *AUC* 与召回率分别达到 0.999 与 0.843 表现最佳；M.Sivaram Chowdary

等[11]将 CNN 与模糊逻辑方法对比，评估芒果叶片病害识别的性能，CNN 的准确率为 95.2%，优势显著；Raghav Agarwal 等[12]研究 CNN 算法在皮肤病诊断中的应用，采用 11 种不同的网络方法，其中 ResNet152 在召回率、准确性与精确度均优于其他深度学习方法。

分析以上研究可以看出，逻辑回归、集成算法、神经网络等机器学习方法广泛应用于疾病的诊断研究，本文基于来自 Kaggle 的中风数据集，利用 11 种数据不平衡处理方法，结合逻辑回归、SVM、CNN、随机森林 4 种机器学习算法，对中风患者的识别进行研究。全文共分为 5 个部分：第 1 部分是相关理论，包括不平衡处理的方法、机器学习算法与模型的评价指标；第 2 部分是实验数据，包括数据来源与统计性分析；第 3 部分是实验，包括数据的预处理、失衡处理与模型结果对比；第 4 部分是模型的优化、SHAP 特征重要性分析；第 5 部分是总结。

2. 相关理论

本节将介绍数据不平衡的处理方式、实验应用的机器学习算法、优化算法以及模型的评估指标。

2.1. 数据失衡的处理

数据失衡的处理方法已经广泛应用于基因表达、医疗疾病诊断和药物副作用预测等多个场景[13]-[15]，数据层面的不平衡处理方法利用重采样技术对数据重新分配，可以最大限度地改变数据集的结构缓解数据失衡的现象，且这种方法与分类器相互独立，互不影响。其主要分为过采样方法、欠采样方法以及混合方法，下面我们将介绍这些方法。

(1) 合成少数类过采样(SMOTE)

SMOTE 对每个少数类样本，找到 k 个近邻，在样本与每个近邻的连线上随机选取插值点，生成新的少数类样本，直到数据平衡。

(2) 自适应合成过采样(ADASYN)

ADASYN 由 SMOTE 改进而来，它先通过 k 近邻评估样本被多数类包围的程度即误分概率，误分概率越高，相应地通过线性插值合成的少数类样本越多。

(3) 随机过采样方法(Random Over Sampling, ROS)

ROS 不改变多数类样本，从少数类样本中随机复制，使得少数类样本数量与多数类一致。

(4) 随机欠采样方法(Random Under Sampling, RUS)

RUS 与 ROS 方法对应，不改变少数类样本，随机删除多数类样本，直到不同类别的数据量一致。

(5) 托梅克链接(TomekLinks)

TomekLinks 通过删除多数类边界样本进行欠采样，不改变少数类样本，从而减少不同类别的重叠区域。

(6) 过采样合成少数类+编辑最近邻规则(SMOTEENN)

SMOTEENN 先利用 SMOTE 生成少数样本，再利用编辑最近邻规则(ENN)筛选所有样本，若少数类样本的 k 近邻中多数类样本更多，或者多数类样本的 k 近邻中少数类样本更多，则判定为噪声并删除。

(7) 合成少数类过采样 + 托梅克链接(SMOTETomek)

SMOTETomek 先利用 SMOTE 生成少数类样本，再利用 TomekLinks 删除边界的多数类样本，结合过采样与欠采样改善数据失衡与边界清晰度。

(8) 边界合成过采样(Borderline-SMOTE)

Borderline-SMOTE 通过分析样本 k 近邻的特点，若少数类样本多，则不合成，若多数类样本多，则插值合成少数类样本，若全为多数类样本，则判断为噪声并删除。

(9) 支持向量机合成少数类过采样(SVMSMOTE)

SVMSMOTE 先训练 SVM 模型，提取少数类支持向量，再对这些少数类支持向量计算 k 近邻，在其近邻间插值生成新样本。

(10) K 均值合成少数类过采样(KMeansSMOTE)

KMeansSMOTE 对所有样本用 Kmeans 聚类，得到 K 个簇，计算每个簇中少数类样本的占比，对少数类被多数类包围的簇进行 SMOTE 插值。

(11) 近邻清洁规则(Neighborhood Cleaning Rule, NCR)

NCR 对所有样本计算 k 个近邻，若样本的近邻投票结果与自身类别不同，则判定为噪声样本进行删除操作，本质是提高数据质量，适合数据失衡不严重的情形，在失衡严重的情况下不能使样本很好的平衡。

2.2. 机器学习算法

(1) 逻辑回归(LR)

逻辑回归是经典的统计学模型，常用于二分类任务，它通过 Sigmoid 函数将特征的线性组合映射到[0, 1]区间，再以最小化交叉熵损失函数为目标，采用梯度下降优化参数，最终完成二分类。

(2) 支持向量机(SVM)

支持向量机是经典的监督学习算法，以寻找最优分离超平面为核心，它通过核函数令特征线性可分，将问题转化为带约束的凸二次规划问题求解，令不同样本到超平面的间隔最大。

(3) 卷积神经网络(CNN)

卷积神经网络是一类包含卷积运算且具有深度结构的前馈神经网络，它通常由输入层、卷积层、池化层与全连接层四部分组成。本研究针对一维时序数据建立了精简 CNN 模型，其中输入层接收变长时序信号，卷积层采用 16 个 1 维卷积核提取局部时间特征，池化层使用自适应全局平均池化，将任意长度特征统一压缩为固定尺寸，解决输入长度不一致问题。随后通过 32 维全连接层(含 Dropout 正则化)和 1 维输出层，经 Sigmoid 函数输出二分类概率。

(4) 随机森林(RF)

随机森林通过构建多个决策树并将它们的预测结果结合，以提高模型的泛化能力。在分类任务中，它先通过 Bootstrap 采样从训练集中随机有放回地抽取样本，得到多个子数据集(决策树)；然后特征采样，在树的每个节点分裂时，随机选择一部分特征进行最佳分裂点的选择，增加模型的多样性；最后每棵决策树在对应的样本和特征子集上训练，并采取多数投票法输出结果。

2.3. 模型评估指标

我们令中风为正类(取值 1)，不中风为负类(取值 0)。在医疗数据极度不平衡的情况下，将准确率作为评估指标已经不再可靠，需要体现模型综合性能的指标，因此本文选择 *AUC*、召回率作为模型的评价指标。混淆矩阵的四个值真阳性(TP)、假阳性(FP)、假阴性(FN)、真阴性(TN)是评价指标的基础，各评价指标的计算见下式：

$$FPR = FP/(FP+TN),\quad TPR = TP/(TP+FN), \tag{1}$$

$$Recall = TP/(TP+FN), \tag{2}$$

其中，*AUC* 又叫作 *ROC* 曲线下面积，用于衡量模型对于正负样本的排序能力，可以在数据不平衡的情形下不受阈值的影响，我们可以根据公式(1)得到 *ROC* 曲线的横轴假正率 *FPR* 与纵轴真正率 *TPR*。式(2)是召回率，评估的是模型识别正类的能力(即中风)。

3. 实验数据

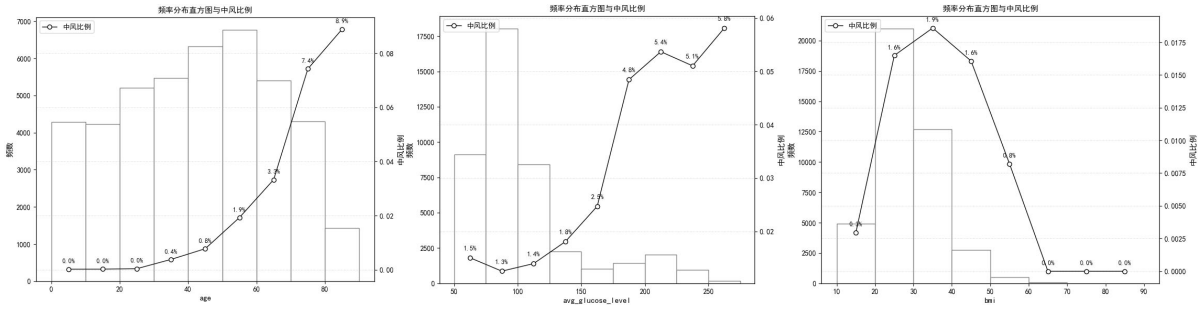
本节将介绍数据来源，并对特征相关性进行初步分析与探讨。

3.1. 数据来源

实验数据选择来自 Kaggle 的公开中风数据集，共包含 43400 条样本，12 个字段，包括数值型自变量 4 个(编号 id、年龄 age、血液葡萄糖含量 avg_glucose_level、身体指数 bmi)，类别型自变量 7 个(性别 gender、有无高血压 hypertension、有无心脏病 heart_disease、是否结婚 ever_married、工作类型 work_type、所在地区 Residence_type、抽烟等级 smoking_status)，因变量 1 个，为是否中风 stroke，中风与不中风的样本量比值为 783:42617，约为 1:54.43，存在严重的不平衡。

3.2. 数据统计性分析

为实现中风样本的精准识别，首先需分析特征与中风的关联性。我们将数据 stroke.csv 导入 Python，将数值型特征进行分箱，id、年龄、血糖、身体指数的分箱规则分别为[0, 3650, 73000]、[0, 10, 100]、[50, 25, 300]、[10, 10, 100]，并绘制分箱后除 id 外各特征的分布图，以及不同箱的中风占比，见图 1。

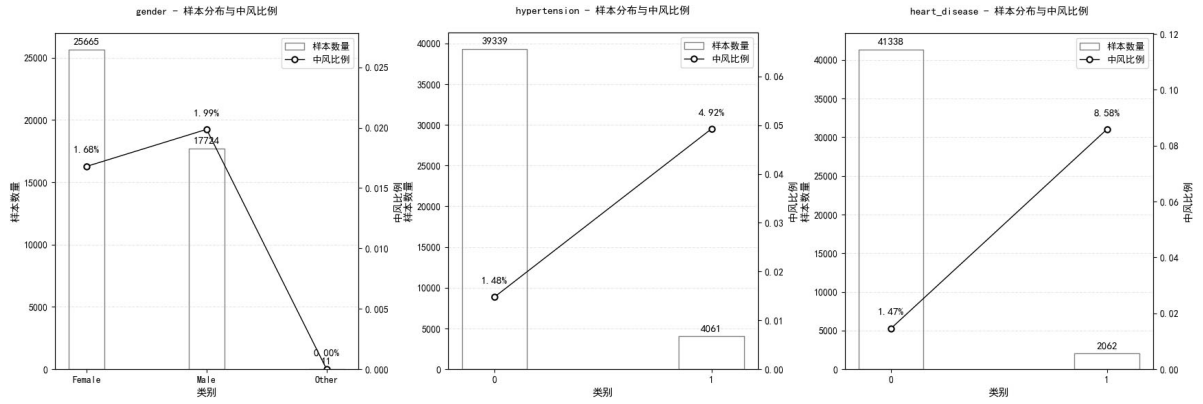


(a) (b) (c)

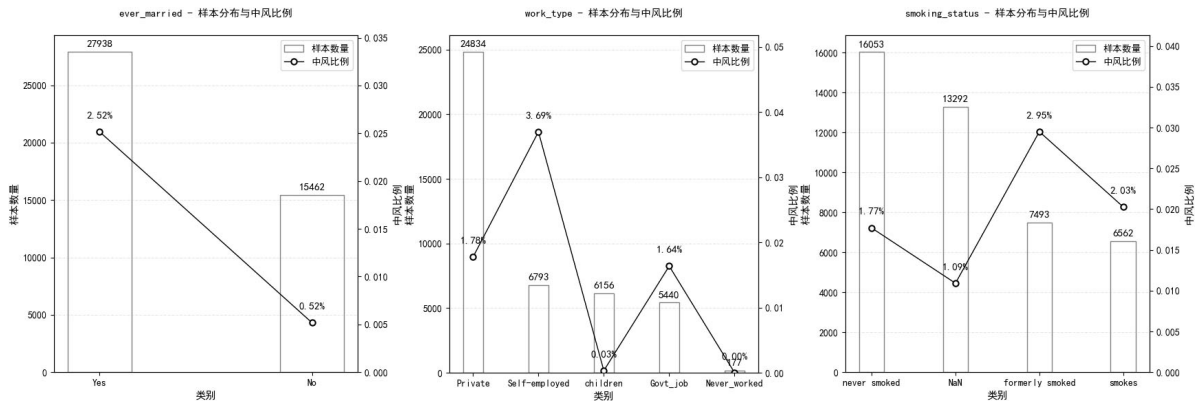
Figure 1. Frequency distribution histogram and stroke proportion of numerical features
图 1. 数值型特征的频率分布直方图与中风比例

由图 1(a)(b)可以看出，年龄、血糖与中风比例呈高度正相关，从(c)看出，身体指数在 30 到 40 范围内的中风比例最高，但是中风比例基本在 2%以下，相关性不高。

接着分析各类别型特征的取值分布以及不同取值下中风的比例，见图 2。



(a) (b) (c)



(d) (e) (f)

Figure 2. Value distribution of categorical features and stroke proportion
图 2. 类别型特征的取值分布与中风比例

由图 2 的中风比例可以看出，患有高血压的比不患高血压的高出 3.44%，患有心脏病的比不患心脏病的高出 7.11%，结婚比不结婚高出 2.00%，自由职业的中风比例最高，吸烟的中风比例略高于不吸烟的，性别与中风不存在高度相关。

综上，通过分析数值型特征与类别型特征的样本特点，可以初步得出年龄、血糖、高血压、心脏病与中风相关度较高，对识别中风有至关重要的作用。

为进一步观察特征之间的关系，我们做出自变量与中风的 Pearson、Spearman、Kendall 三种相关系数见图 3，分析图 3，各个特征与中风之间的相关系数比较：年龄 > 心脏病 > 高血压 > 血糖 > 婚姻 > 工作类型，而 id、性别、所在地、身体指数与中风基本不相关，与初步分析相符。

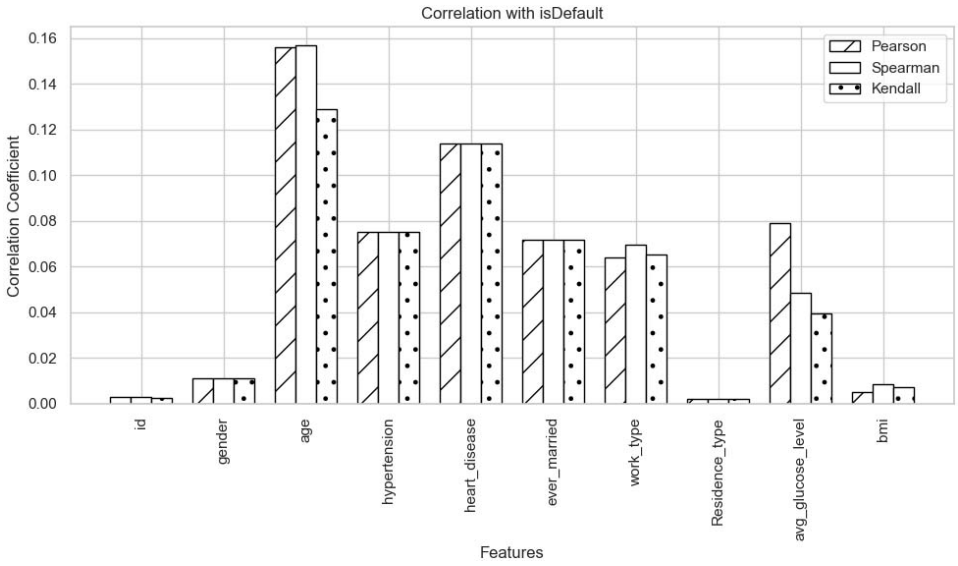


Figure 3. Three correlation coefficients
图 3. 三种相关系数

4. 实验

本节将对实验数据进行处理，将机器学习算法与不平衡处理相结合。

4.1. 数据预处理

在建立分类模型之前，我们需要对数据中的类别特征编码，并对数据的缺失值、异常值进行处理。经过分析，身体指数有 1462 个缺失值，抽烟等级有 13292 个缺失值，总数为 14754 个，我们将身体指数的缺失值用中国人平均 bmi 水平 23.5 填补，将抽烟等级的缺失值记为新的标签，解释为“unknown”。

缺失值处理后，我们对 id 进行分箱(箱数为 10)，接着对类别型数据进行标签编码后识别异常值，得到存在异常值的特征为高血压(4061 个异常值，中风比例 4.92%)、心脏病(2062 个异常值，中风比例 8.58%)、血糖(4978 个异常值，中风比例 5.08%)、身体指数(1084 个异常值，中风比例 1.01%)。

由于原数据集的中风比例为 1.80% (783/43400)，大多数存在异常值的特征，其异常值的中风比例是整体中风比例的 2 到 3 倍，且类别型特征的异常值删除后，特征将失去意义，因此对数据的异常值进行保留。

在机器学习中，由于数据存在严重的失衡，会导致模型出现严重偏向，因此选择合适的数据不平衡处理方法，是决定识别模型性能优良的关键。我们将预处理后的数据集按照 7:3 的比例随机划分为训练集与测试集，利用 LR、SVM、CNN 与随机森林 4 种机器学习算法进行实验。

4.2. 失衡处理与模型对比

我们应用 11 种数据层面的不平衡处理方式，处理训练集数据，处理后的类别失衡比率见表 1。

Table 1. Imbalance ratio after data imbalance processing
表 1. 数据不平衡处理后的不平衡比率

方法	不处理	SMOTE	ADASYN	ROS	RUS	TomekLinks
比例	54.43	1.00	0.997	1.00	1.00	53.96
方法	SMOTEENN	SMOTETomek	BorderlineSMOTE	SVMSMOTE	KMeansSMOTE	NCR
比例	0.90	1.00	1.00	1.79	1.00	51.97

由表 1 观察可观察到：经过 SMOTE、ADASYN、ROS、RUS、SMOTETomek、BorderlineSMOTE、SVMSMOTE、KMeansSMOTE 处理后的数据基本可以实现平衡，而经过 SMOTEENN 处理后的数据，会生成较多的合成样本，最终少数类样本会比多数类样本多，TomekLinks、NCR 处理后的数据去除了噪声样本，数据质量有所提高，适合轻微失衡的情境，对于数据极度不平衡的作用微乎其微。

数据处理后，分别将其输入不同的分类模型，得到模型的结果见表 2。

Table 2. Comparison of model results
表 2. 模型结果的对比

	LR		SVM		CNN		RF	
方法	*AUC*	*Recall*	*AUC*	*Recall*	*AUC*	*Recall*	*AUC*	*Recall*
不处理	84.68	0	63.74	0	79.44	0	84.22	0
SMOTE	84.79	79.15	77.72	52.77	79.80	71.91	81.40	41.28
ADASYN	**84.82**	**80.00**	77.50	53.62	78.61	65.53	81.29	41.28
ROS	**84.81**	**80.43**	78.17	59.57	**80.34**	79.15	82.57	53.19
RUS	**84.76**	**82.13**	**82.98**	**79.57**	77.52	71.49	**84.17**	**82.98**
TL	84.67	0	63.38	0.43	79.11	0	84.17	0
SMOTEENN	**84.74**	**80.85**	77.92	57.45	**80.12**	72.34	81.84	55.32
SMOTETomek	84.79	79.57	77.39	52.34	78.93	67.23	81.42	41.70
BorderlineSMOTE	84.42	74.47	79.36	37.45	**80.58**	63.40	82.09	28.51

续表

SVMSMOTE	84.44	61.28	79.07	31.06	79.46	37.45	82.71	18.72
KMeansSMOTE	83.53	20.85	71.34	18.72	78.11	25.53	83.07	9.36
NCR	84.57	0	63.71	0.43	79.83	0	84.05	0

观察表 2，从机器学习算法的角度，当数据极度不平衡时，CNN 与 ROS 结合性能最好，*AUC* 为 80.34%，召回率为 79.15%，但运行时间为 702.35 s。而 LR、SVM、RF 与 RUS 结合所得模型的 *AUC* 分别为 84.76%、82.98%、84.17%，召回率分别为 82.13%、79.57%、82.98%，运行时间分别为 0.54 s、0.26 s、1.46 s，和 CNN 相比优势明显。不仅如此，LR 经过各种方法处理后模型都得到了较大的提升，因此 LR 与数据失衡处理方法的适配度高于其他算法。

为了对比各种数据不平衡处理方法，我们依据表 2 做出图 4 与图 5。

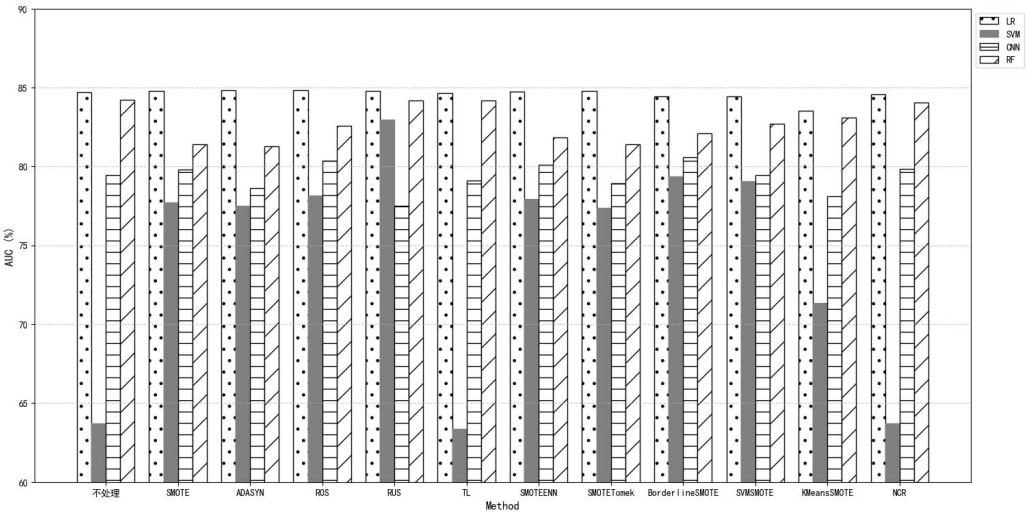


Figure 4. AUC after data imbalance processing
图 4. 数据不平衡处理后的 AUC

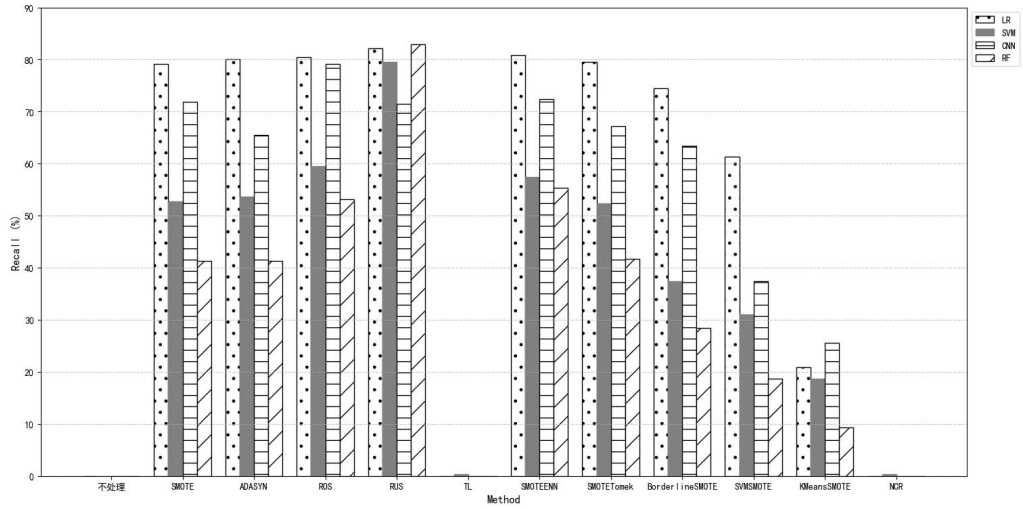


Figure 5. Recall after data imbalance processing
图 5. 数据不平衡处理后的 Recall

从数据失衡处理的角度，由图 4 可以看出，除 TL、KMeansSMOTE、NCR 外，不同算法经过不平衡处理后的 *AUC* 均超过 75%，由图 5 可以看出，只有 RUS 处理后的 *Recall* 都超过 70%，说明模型中风的漏判率低，而 ROS 与 SMOTEENN 处理后的模型 *Recall* 都超过 50%，表现次之。

为了进一步分析 RUS 在研究中表现的出色性能，我们通过主成分分析对数据降维，并做出 RUS、SMOTEENN 与 ADASYN 的噪声样本量的对比图，见图 6。

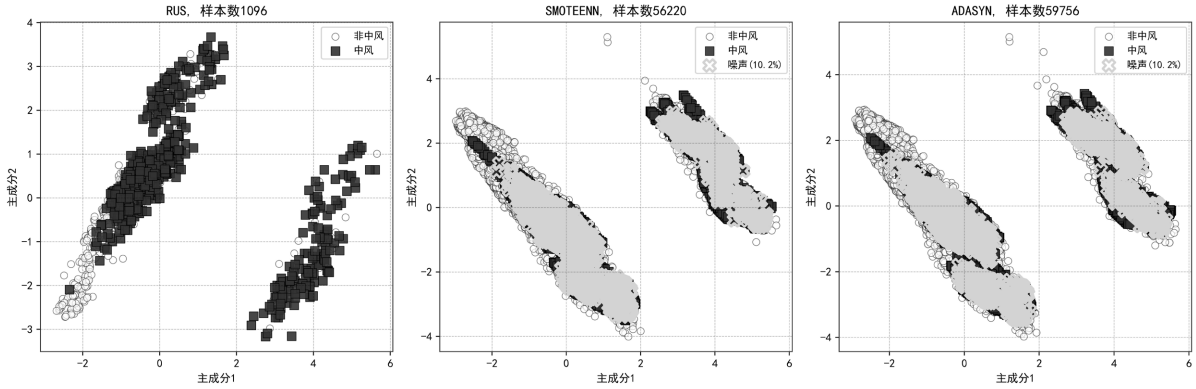


Figure 6. Comparison of noise sample size
图 6. 噪声样本量对比

由图 6 可以看出，由于 RUS 是简单地删除原有多数样本，从而平衡数据，所以并没有产生多余的噪声样本，保留了原始样本的真实可靠，且 RUS 显著提升计算效率，而以 SMOTEENN 与 ADASYN 为代表的其余方法，在平衡数据的过程中产生了大量的噪声样本，影响了模型的分类性能。

5. 模型优化

在初始参数下，逻辑回归、SVM 与随机森林的模型表现良好，为进一步提升模型性能，我们应用 PSO、GA、DE、BO 四种方法进行参数优化。

5.1. 优化算法

我们先对这 4 种方法进行简单的介绍：

(1) 粒子群优化算法(PSO)

PSO 是模拟鸟群觅食行为的、基于群体协作的随机搜索算法。它通过随机初始化粒子群，计算适应度，然后更新个体、全局的最佳适应度，最后更新粒子的速度和位置，是一种概率型的全局优化方法。

(2) 遗传算法(GA)

GA 通过选择、交叉、变异等基本操作模拟达尔文进化论自然选择与生物遗传进化的过程，多应用于解决复杂、非线性和多峰值的优化问题。

(3) 差分算法(DE)

DE 是一种高效的全局优化算法，模拟生物进化的随机模型，进化流程与遗传算法相似，都包括变异、杂交和选择操作，GA 的通用性更强，但计算复杂度高，DE 更适合连续空间优化，且操作简单。

(4) 贝叶斯优化算法(BO)

BO 基于贝叶斯定理，使用概率模型对目标函数进行建模和推断，也是全局优化算法，尤其适用于目标函数复杂的问题，对于高维、黑盒和不可导的优化问题具有很好的适用性。

5.2. 优化实验

我们将 *AUC* 作为优化目标，并采用 10 折交叉验证，下面分析优化结果。

Table 3. LR optimization results
表 3. LR 优化结果

	PSO	GA	DE	BO
LR	84.28	84.22	**84.31**	84.22
SVM	**84.50**	84.37	84.28	83.91
RF	84.13	**84.16**	84.05	83.99

逻辑回归有 4 个重要参数：正则化类型 penalty 通过给模型参数添加惩罚项来避免过拟合，正则化参数 C 是正则化强度的反向调节参数，越大越容易导致过拟合，最大迭代次数 max_iter 控制训练的最大迭代步数，class_weight 调整权重。由表 3，DE-LR 效果最好：L1 正则化，C = 0.1340685，max_iter = 304，class_weight = 1: 1.7714181 得到测试集的 *AUC* 为 84.74%，召回率 87.23%。

SVM 以找到最大间隔超平面为目标，有 4 个重要参数：惩罚系数 C 越大，允许误分类的数量越少，核函数类型 kernel 将低维非线性数据映射到高维空间，实现线性可分，核函数参数 gamma 越小，边界越平滑，越抗过拟合，class_weight 调整权重。由表 3，PSO-SVM 效果最好：C = 2.0848617，kernel 为径向基核 rbf，gamma = 0.0065456，class_weight = 1:1，得到测试集的 *AUC* 为 84.68%，召回率 81.70%。

RF 核心是多棵决策树投票表决，有 5 个重要参数：决策树数量 n_estimators 决定森林规模，单棵决策树最大深度 max_depth 越大越容易过拟合，节点分裂的最小样本数 min_samples_split 限制小样本节点的分裂，叶节点的最小样本数 min_samples_leaf 防止过拟合，class_weight 调整权重。由表 3，GA-RF 效果最好：n_estimators = 250，max_depth = 6，min_samples_split = 15，min_samples_leaf = 8，class_weight = 1:2.1827791，得到 *AUC* 为 84.18%，召回率 91.06%。

在疾病诊断的情境中，低漏判率是模型优化的核心目标。根据上述结果，模型经过优化后，*AUC* 都可以达到 84%，但召回率存在较大的差异。为了得到最佳模型，通过公式 $\overline{X_1}=\frac{AUC+Recall}{2}$、$\overline{X_2}=\sqrt{AUC\cdot Recall}$、$\overline{X_3}=\frac{2\cdot AUC\cdot Recall}{AUC+Recall}$ 分别计算 DE-LR、PSO-SVM、GA-RF 的 *AUC* 与召回率的算术平均、几何平均、调和平均见表 4，可以看出，GA-RF 的三种均值都高于 87%，分别比 DE-LR、PSO-SVM 高出 2%与 4%，因此 GA-RF 综合性能最佳，且其在漏判率上具有显著优势。

Table 4. Mean calculation
表 4. 均值计算

	算术均值	几何均值	调和均值
DE-LR	85.99	85.98	85.97
PSO-SVM	83.19	83.18	83.16
GA-RF	**87.62**	**87.55**	**87.48**

5.3. SHAP 特征重要性

为了进一步增强模型 GA-RF 的可解释性，我们利用 SHAP 对特征的影响程度进行分析，并得到特征重要性的排序见图 7。

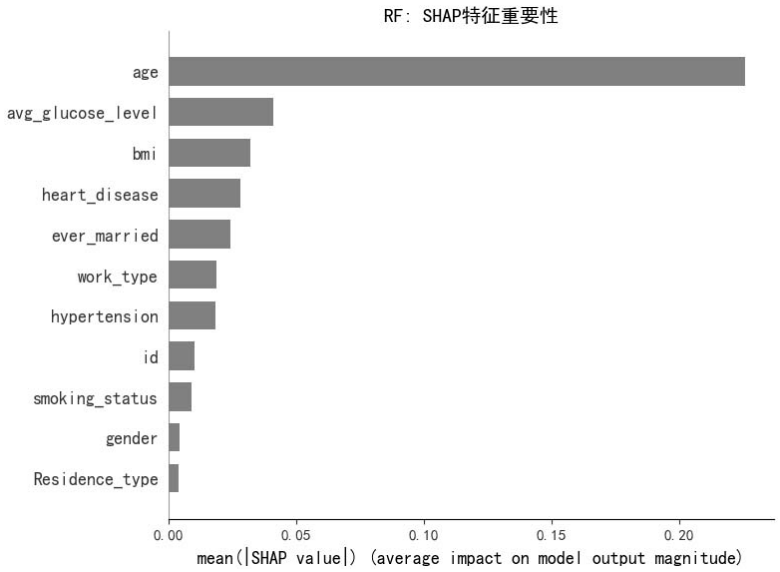


Figure 7. SHAP feature importance analysis of GA-RF
图 7. GA-RF 的 SHAP 特征重要性分析

由图 7 可以看出，年龄对模型结果的作用最大，远超其他特征。除此之外，血糖、bmi、心脏病等也对模型有不同程度的影响，而 id、抽烟、性别、居住环境的影响较小。该结果与第 2 节数据的统计性分析中，对特征与中风的相关性分析基本一致。

6. 总结

当数据存在不平衡时，机器学习模型通常无法充分关注少数类样本，导致模型输出结果向多数类倾斜，准确率变得不再可靠。为了解决这个问题，本文先对来自 Kaggle 的公开中风数据集进行了统计性分析，讨论特征之间的相关性，并对缺失值、异常值进行识别与处理。然后，分别研究 11 种不平衡数据处理方法与逻辑回归、SVM、CNN、随机森林的组合模型，对中风数据分类，并对比模型的 *AUC*、召回率、运行时间，证得 RUS 与逻辑回归、SVM、随机森林结合的优越性。随后利用 PSO、GA、DE、BO 四种方法对逻辑回归、SVM 与随机森林进行优化，以 *AUC* 为优化目标，引入 10 折交叉验证，通过计算 *AUC* 与召回率的算术均值、几何均值与调和均值，评估模型综合性能，得到 GA-RF 的性能最佳。最后，利用 SHAP 探讨了各个特征对 GA-RF 识别结果的贡献，进一步增强了模型的可解释性，证得年龄对模型的作用远大于其他特征。

针对以上研究，文章还存在如下不足与改进方向：

(1) RUS 的数据失衡处理易导致数据信息丢失，且数据极度比轻微不平衡的场景更加适用，使得模型的泛化能力不足，适用于特定数据集，后续将在多个独立的外部数据集上验证，检验其泛化能力；

(2) 机器学习算法不足。可以采用更多单一或混合的机器学习算法进行中风识别，助力疾病诊断。

参考文献

[1] Pereira, J. and Saraiva, F. (2021) Convolutional Neural Network Applied to Detect Electricity Theft: A Comparative Study on Unbalanced Data Handling Techniques. *International Journal of Electrical Power & Energy Systems*, **131**, Article ID: 107085. https://doi.org/10.1016/j.ijepes.2021.107085

[2] Khushi, M., Shaukat, K., Alam, T.M., Hameed, I.A., Uddin, S., Luo, S., *et al.* (2021) A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access*, **9**, 109960-109975.

https://doi.org/10.1109/access.2021.3102399

[3] Huynh, T., Nibali, A. and He, Z. (2022) Semi-Supervised Learning for Medical Image Classification Using Imbalanced Training Data. *Computer Methods and Programs in Biomedicine*, **216**, Article ID: 106628. https://doi.org/10.1016/j.cmpb.2022.106628

[4] Xu, Z., Shen, D., Nie, T., Kou, Y., Yin, N. and Han, X. (2021) A Cluster-Based Oversampling Algorithm Combining SMOTE and K-Means for Imbalanced Medical Data. *Information Sciences*, **572**, 574-589. https://doi.org/10.1016/j.ins.2021.02.056

[5] Kumar, V., Lalotra, G.S., Sasikala, P., Rajput, D.S., Kaluri, R., Lakshmanna, K., *et al.* (2022) Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques. *Healthcare*, **10**, Article 1293. https://doi.org/10.3390/healthcare10071293

[6] Banik, D. and Bhattacharjee, D. (2021) Mitigating Data Imbalance Issues in Medical Image Analysis. In: Rana, D.P. and Mehta, R.G., Eds., *Data Preprocessing*, *Active Learning*, *and Cost Perceptive Approaches for Resolving Data Imbalance*, IGI Global, 66-89. https://doi.org/10.4018/978-1-7998-7371-6.ch004

[7] Yao, P., Shen, S., Xu, M., Liu, P., Zhang, F., Xing, J., *et al.* (2022) Single Model Deep Learning on Imbalanced Small Datasets for Skin Lesion Classification. *IEEE Transactions on Medical Imaging*, **41**, 1242-1254. https://doi.org/10.1109/tmi.2021.3136682

[8] Jevin, M.J., Jayant, H., Sanjay, R., *et al.* (2023) Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. *Heart Disease*, **10**, 2322-2327.

[9] Abdar, M., Zomorodi-Moghadam, M., Das, R. and Ting, I. (2017) Performance Analysis of Classification Algorithms on Early Detection of Liver Disease. *Expert Systems with Applications*, **67**, 239-251. https://doi.org/10.1016/j.eswa.2016.08.065

[10] Ding, H., Fawad, M., Xu, X. and Hu, B. (2022) A Framework for Identification and Classification of Liver Diseases Based on Machine Learning Algorithms. *Frontiers in Oncology*, **12**, Article 1048348. https://doi.org/10.3389/fonc.2022.1048348

[11] Sivaram Chowdary, M. and Puviarasi, R. (2022) Accuracy Improvement in Disease Identification of Mango Leaf Using CNN Algorithm Compared with Fuzzy Algorithm. *ECS Transactions*, **107**, 11889-11903. https://doi.org/10.1149/10701.11889ecst

[12] Agarwal, R. and Godavarthi, D. (2023) Skin Disease Classification Using CNN Algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, **9**, 1-8. https://doi.org/10.4108/eetpht.9.4039

[13] Al-Azani, S., Alkhnbashi, O.S., Ramadan, E. and Alfarraj, M. (2024) Gene Expression-Based Cancer Classification for Handling the Class Imbalance Problem and Curse of Dimensionality. *International Journal of Molecular Sciences*, **25**, Article 2102. https://doi.org/10.3390/ijms25042102

[14] Chen, C., Wu, X., Zuo, E., Chen, C., Lv, X. and Wu, L. (2023) R-GDORUS Technology: Effectively Solving the Raman Spectral Data Imbalance in Medical Diagnosis. *Chemometrics and Intelligent Laboratory Systems*, **235**, Article ID: 104762. https://doi.org/10.1016/j.chemolab.2023.104762

[15] Wang, J., Yu, L. and Zhang, X. (2022) Explainable Detection of Adverse Drug Reaction with Imbalanced Data Distribution. *PLOS Computational Biology*, **18**, e1010144. https://doi.org/10.1371/journal.pcbi.1010144