从构想到部署：基于机器学习的智能卒中预测框架与性能评估

Leila Ismail[1,2,*]，IEEE 会员及 Huned Materwala[1,21]智能分布式计算与系统（INDUCE）研究实验室，计算机科学与软件工程系，信息技术学院，阿联酋大学阿联酋艾因，阿布扎比 15551，阿拉伯联合酋长国[2]阿联酋大学国家水与能源中心阿联酋艾因，阿布扎比，阿拉伯联合酋长国 *通讯作者：Leila Ismail（电子邮箱：leila@uaeu.ac.ae）

摘要—中风是全球第二大死因。机器学习分类算法已广泛用于卒中预测。然而，这些算法采用了不同的数据集和评估指标进行评估。此外，目前尚无全面的卒中数据分析框架。本文通过对文献中的机器学习预测算法进行审慎考察，提出了一种智能卒中预测框架。本文采用统一的设置，对用于卒中预测的五种最常用机器学习算法进行评估，以实现客观比较。比较分析和数值结果表明，随机森林算法最适用于卒中预测。

关键词—人工智能、分类算法、数据分析、电子健康、健康信息学、机器学习、卒中预测

I. 引言

中风是一种神经系统疾病，会因血流中断而导致脑细胞死亡 [1]。中风是全球第二大死因，占死亡总数的 11%[1]。根据美国疾病控制与预防中心（CDC）的数据，美国每年有超过 795,000 人罹患中风[2]。2017 至 2018 年间，中风相关卫生支出约为 530 亿美元[2]。中风患病率与人口统计、遗传、生活方式及疾病状况等风险因素的相互作用有关 [1]。如果未能及早诊断和治疗，中风可能导致视觉、情绪、临床及传染性疾病方面的问题 [2]。迄今为止，尚无简便方法可以治疗和预防中风的临床病因 [1]。机器学习分类算法在中风预测方面展现出良好的潜力。

文献中已有多项研究在不同的实验设置下评估了多种用于卒中预测的机器学习算法的性能 [3]–[8],，这使得难以对这些算法进行客观比较。此外，据我们所知，目前尚无全面描述卒中预测流程的框架。本文提出一种基于机器学习的智能卒中预测框架，以帮助辅助医疗专业人员（医生、病理学家、治疗师和医学技术人员）更好地进行诊断和预后评估 [9]，并协助医疗利益相关者（政府、制药公司和健康保险提供方）制定有效的卒中预防计划。我们使用文献中最常用的五种卒中预测机器学习方法对所提出的框架进行评估 [3]–[8] （图 1）：决策树（DT）、随机森林（RF）、支持向量机（SVM）、朴素贝叶斯（NB）和逻辑回归（LR），以便在统一设置下进行客观比较。本文的主要贡献如下。

- 我们提出了一种智能卒中预测框架，展示了从领域概念构建到模型部署的卒中数据分析流程。
- 我们对卒中预测中最常用的机器学习方法进行了评估，以便进行客观比较。
- 我们评估了数据平衡对机器学习方法性能的影响。

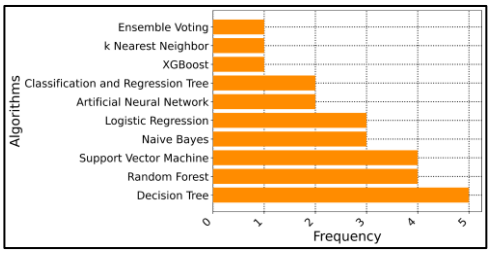

图1. 文献中用于卒中预测的机器学习分类算法的使用频率。

本文其余部分安排如下：第二节总结相关工作。第三节介绍我们提出的智能卒中预测框架。第四节介绍用于卒中预测的机器学习算法。第五节讨论实验设置、实验和数值分析。第六节总结全文。

II. 相关研究

表 I 对比了基于机器学习的卒中预测相关研究。这些研究使用了不同的数据集和性能指标进行评估。本文提出了一种智能卒中预测框架，并采用文献中最常用的五种机器学习算法对其进行评估，从而对这些算法进行客观比较。

[1] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death （最后访问于2022年4月25日）

[2] https://www.cdc.gov/stroke/facts.htm （最后访问于 25, 202 2）

表 I. 文献中基于机器学习的卒中预测研究。

Work	数据集	特征	观测值	算法 已评估	评估指标
[3]	私人 – 玛哈沙拉堪医院	性别, 婚姻状况、吸烟, 饮酒、运动、职业、血压和胆固醇	500（中风）和 500（无中风）	决策树、支持向量机、NB, and KNN	准确率、精确率*、召回率*和 F-测量*
[4]	私人 – 急诊医疗中心 忠南大学医院, 韩国	意识水平、最佳凝视, 视觉、面瘫、左右上肢运动功能、左右下肢运动功能、肢体共济失调, 感觉, best 语言, 构音障碍、消退与注意力不集中, 以及左右远端运动功能	227	人工神经网络、随机森林, XGBoost, 支持向量机、朴素贝叶斯, DT, LR, CART	准确率、精确率和召回率
[5]	MarketScan 医疗补助计划 多州数据库	4146 药物 and 医疗疾病状况、年龄和性别	17,86（中风）和 10,800（无中风）	随机森林、支持向量机、DT, LR, 和分类与回归树	曲线下面积和训练时间
[6]	卒中 预测数据集[3]	标识符、性别、年龄、高血压、心脏病、曾婚姻状况、工作类型、居住类型、平均血糖、体质指数、和吸烟状况	249（中风）和 4,861（无中风）	RF, DT, LR, and 投票法	准确率、精确率*、召回率*和 F-度量*
[7]	私人	脑部磁共振成像图像	192	RF and SVM	混淆矩阵、准确率, 敏感度、特异度、Jaccard指数和DSI
[8]	私人 – 学院物理治疗, 玛希隆大学, 泰国	性别、年龄、省份、婚姻状况、教育程度和职业	250（中风）和 250（无中风）	人工神经网络, 朴素贝叶斯, 和决策树	准确率、假阳性率、假阴性率和曲线下面积
This 论文	中风 预测数据集[4]	标识符、性别、年龄、高血压、心脏病、曾婚姻状况、工作类型、居住类型、平均血糖、体质指数, 和吸烟状况	783（中风）和 42,617（未中风）	DT, RF, 支持向量机、朴素贝叶斯和逻辑回归	准确率、精确率*、召回率*和 F-度量*、受试者工作特征、曲线下面积和执行时间，用于超参数调优和模型开发与验证

DT：决策树；SVM：支持向量机；NB：朴素贝叶斯；KNN：K近邻；ANN：人工神经网络；RF：随机森林；LR：逻辑回归；CART：分类与回归树；ROC：受试者工作特征；AUC：曲线下面积；BMI：体质指数；FPR：假阳性率；FNR：假阴性率；*适用于中风和非中风二分类；JI：Jaccard指数；DSI：Dice相似系数

III. 拟议的智能卒中预测框架

图 2 展示了我们提出的智能卒中预测框架所涉及的各个阶段，该框架基于数据分析生命周期 [10]。在接下来的小节中，我们将详细说明每个阶段。

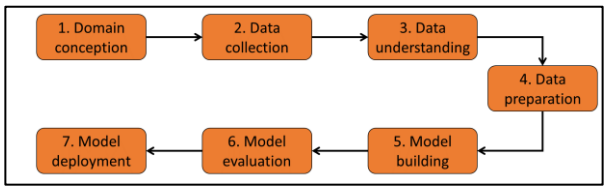

图2. 拟议的智能卒中预测框架的各阶段。

A. 领域概念构建

在这一阶段，将研究卒中预测问题，即缺血性或出血性卒中 [1]。应通过查阅文献 [1]和/或咨询专家，确定潜在的卒中风险因素。此外，还应明确说明预测模型的目标，即二分类、卒中后状况或特定风险因素预测。

B. 数据收集

在此阶段，通过加州大学欧文分校（UCI）机器学习存储库等公共存储库收集卒中数据集[5]，和/或在征得同意后使用患者的医疗记录创建数据集。应确保将卒中风险因素作为特征纳入数据集，以开发准确且具有代表性的预测模型。

C. 数据理解

在此阶段，如果数据集分散在多个文件中，应将其汇总。此外，如果风险因素使用某种编码方案表示，例如《国际疾病分类》（ICD）-9[6]，则应参考国际疾病编码系统。

D. 数据准备

在这一阶段，应对数据集进行预处理，移除患者ID等无关特征，以防止模型在构建过程中出现过拟合。此外，应采用机器学习方法 [11]和/或人工视觉检查来移除离群值，以提高预测准确率 [12]。应使用独热编码或标签编码等技术，将分类特征（例如婚姻状况）转换为数值形式 [13]。此外，应对数值型风险因素进行归一化，避免取值范围较大的风险因素相较于取值范围较小的因素产生偏差 [12]。应识别并处理数据集中的缺失值，可以移除存在缺失值的观测值，或使用统计学或机器学习方法填补合成值 [12]。机器学习模型可以使用预处理数据集中的全部风险因素开发，也可以使用通过特征选择算法 [14] 和/或专家建议选出的部分因素开发。此外，如果数据集不平衡，应采用过采样和欠采样方法使其达到平衡 [12]。

E. 模型构建

在此阶段，使用预处理数据集构建机器学习模型。可以采用 k 折

[3] https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset（最后访问于2022年4月25日）[4] https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset（最后访问于2022年4月25日） [5] https://archive.ics.uci.edu/ml/datasets.php（最后访问于2022年4月25日）[6] https://www.cdc.gov/nchs/icd/icd9cm.htm（最后访问于2022年4月25日）

交叉验证 [14] 或百分比分割方法。应使用最优超参数值来构建模型，以提升预测性能。

F. 模型评估

在此阶段，使用验证数据集对模型进行评估。对于k折交叉验证，每一折都使用数据集的1/k来验证模型。应根据准确率评估模型。如果数据集存在类别不平衡，则应评估类别特定和宏平均的精确率、召回率和F值，以及受试者工作特征（ROC）曲线和ROC曲线下面积（AUC）。应通过测量超参数调优、模型构建和验证的执行时间来评估模型的复杂度。

G. 模型部署

在此阶段，相关医疗专业人员使用已开发并经过验证的模型进行卒中预测。应根据数据集中更新和/或新增的观测值重新构建模型。

IV. 用于卒中预测的机器学习算法

本节介绍用于卒中预测的机器学习分类算法。

A. 决策树（决策树）构建

基于规则的模型，方法是构建一种树结构，其中每个卒中风险因素代表一个节点 [14]。最顶层的节点称为根节点，带有类别标签（中风和无中风）的节点称为叶节点。在每个节点，算法通过使用基于熵的信息增益或基尼指数选择信息量最大的风险因素，向下遍历至下一个节点/叶节点。给定包含风险因素 $\{R_1, R_2, \dots\}$ 的卒中数据集和卒中类别标签，风险因素 R_1 的信息增益可按公式（1）计算。

$$InfoGain_{R_1} = H_{stroke} - H_{stroke|R_1} \tag{1}$$

其中，H_{stroke} and $H_{stroke|R_1}$ represent base (表示 R_1 的条件熵 and) 公式 2 公别表示（公式 3）。

$$H_{stroke} = -\sum_{\forall s \in stroke} P(s)\log_2 P(s) \tag{2}$$

$$H_{stroke|R_1} = -\sum_{\forall r \in R_1} P(r) \sum_{\forall s \in stroke} P(s|r)\log_2 P(s|r) \tag{3}$$

其中，$s \in \{yes, no\}$，而 r 是风险因素 R_1 的取值集合。

基尼指数（GI）可使用公式（4）计算。

$$GI(R_1) = Gini(stroke) - \sum_{\forall r \in R} P(r).Gini(R_1 = r) \tag{4}$$

其中，$Gini(stroke)$ 表示卒中类别的基尼不纯度，如公式（5）所示。

$$Gini(stroke) = 1 - \left(P(stroke = yes)\right)^2 - \left(P(stroke = no)\right)^2 \tag{5}$$

参数：1）分裂准则；2）进行分裂决策时要考虑的最大特征数。

B. 随机森林

随机森林由多个决策树组成，每棵树都使用数据集和风险因素特征的随机样本构建[14]。它主要涉及三个超参数：1）决策树的数量；2）每棵决策树的分裂准则；3）分裂时要考虑的最大特征数。

C. 支持向量机 支持向量机（SVM）构建一个超平

面，将 n维风险因素特征空间分隔开来，从而把观测值划分为中风类别和非卒中类别 [14]。选择能够使两个类别的观测值之间距离（称为间隔）最大的超平面。用于最大化间隔的代价函数如公式 (6) 所示。

$$min_w \lambda \left|\left|w\right|\right|^2 + \sum_{i=1}^{n} (1 - s_i < R_i, w >)_+ \tag{6}$$

其中，e λ is the regularization parameter, w 垂直于超平面，n 是数据集中的观测值数量。

支持向量机使用核函数（线性、多项式、径向基函数或 Sigmoid函数）来构建非线性超平面[15]。它包含三个超参数：1）核函数；2）正则化参数；3）核系数。

D. 朴素贝叶斯

朴素贝叶斯描述了两个风险因素 R_1 R_2 and 与 的概率及条件概率之间的关系。给定某观测的 患病率， $R_2's$ 发生的条件概率由公式 (7) 给出。

$$P(R_2|R_1) = \frac{P(R_1 \cap R_2)}{P(R_1)} = \frac{P(R_1|R_2).P(R_2)}{P(R_1)} \tag{7}$$

For each observation $R_1, \dots, R_k$, 朴素贝叶斯以使公式 (8) 所示的条件概率 $P(s_i|R_1, \dots, R_k)$ 最大的方式预测类别标签 s_i 。

$$P(s_i|R_1, \dots, R_k) = \frac{P(R_1, \dots, R_k|s_i).P(s_i)}{P(R_1, \dots, R_k)} \tag{8}$$

采用公式 (8) 中所述贝叶斯定理的以下假设：1）对于给定的中风结局，数据集中每个风险因素在条件上均不依赖于其他任何风险因素（公式 9）；2）忽略公式 (7) 中的分母，因为它不会影响条件概率。因此，可以使用公式 (10) 计算 s_i 的条件概率。

$$P(R_1, \dots, R_k|s_i) = \prod_{j=1}^{k} P(R_j|s_i) \tag{9}$$

$$P(s_i|R_1, \dots, R_k) = P(s_i).\prod_{j=1}^{k} P(R_j|s_i) \tag{10}$$

E. 逻辑回归

逻辑回归根据公式 (11)，预测观测 i 在给定风险因素 R_i 时类别标签 s_i 的概率。

$$P(s_i) = \frac{e^{(\beta_0 + \beta_1 R_i)}}{1 + e^{(\beta_0 + \beta_1 R_i)}} \tag{11}$$

其中 β_0 和 β_1 是系数，其取值通过优化方程（12）所述的对数损失函数获得。为降低逻辑回归的方差和过拟合，在损失函数中加入惩罚项，如方程（13）所述。逻辑回归包含两个

D. algorithm mainly consists of two hyperpar

超参数：1）正则化参数和 2）求解器用于优化问题。

$$L_{log} = -\sum_{i=1}^{n} -\ln\left(1 + e^{(\beta_0+\beta_1 R_i)}\right) + s_i^{actual}(\beta_0 + \beta_1 R_i) \quad (12)$$

$$L_{log} + \lambda \sum_{i=1}^{n} \beta_i^2 \quad (13)$$

V. 性能分析

A. 实验环境 为评估性能

，在所提出的框架中使用了一个类别不平衡的卒中预测数据集[4]，以评估最常用的五种机器学习算法，即决策树、随机森林、支持向量机、朴素贝叶斯和逻辑回归。图3展示了特征与卒中患病率之间的相关性。如图所示，年龄与卒中患病率的正相关性最高，其次是心脏病。平均血糖水平、高血压、婚姻状况、体质指数和性别与卒中呈轻微正相关。居住类型与卒中几乎没有相关性，而工作类型和吸烟状况则与卒中呈轻微负相关。如图4所示，体质指数和吸烟状况分别有1462个（3.37%）和13292个（30.63%）观测值存在缺失值。

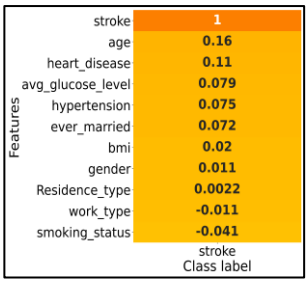

图3. 数据集特征与卒中患病率之间的相关性。

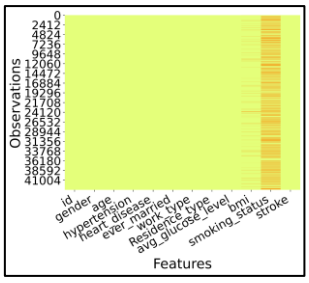

图4. 数据集中各观测值缺失情况的分布。

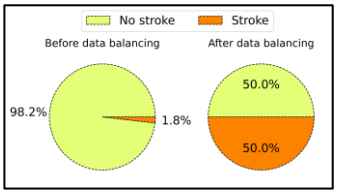

图5. 数据平衡前后数据集中卒中和未发生卒中的观测值百分比。

B. 实验

1）数据预处理：我们通过移除无关特征“ID”对数据集进行预处理。为处理缺失值，删除 BMI 和吸烟状况特征存在空值的观测值。分类特征，即性别、曾婚姻状况、工作类型、居住类型和吸烟状况，使用标签编码转换为数值。具体而言，为每个特征的不同取值分配一个介于 0 和 $v-1$ 之间的数值，其中 v表示该特征的唯一值数量。

例如，对“性别”特征执行标签编码后，分类取值 female、male 和 other 分别转换为 0、1 和 2。预处理数据集包含 10 个特征、1 个类别标签（中风/未中风）和 29072 个观测值，并且存在严重的类别不平衡，如图 5 所示，即分别有 548 个和 28,524 个观测值属于中风类和未中风类。因此，使用合成少数类过采样技术（SMOTE）[12]对少数类（中风）进行过采样。选择 SMOTE 是基于其相较于其他数据平衡方法的性能表现 [16]。平衡后的数据集包含 57048 个观测值（28524 个中风和 28524 个未中风），如图 5 所示。

2）超参数调优：为获得所研究的每种算法的最优参数，我们针对参数的不同取值进行网格搜索。表 II 列出了每种算法的超参数、我们实验和文献中使用的相应取值范围，以及网格搜索的组合总数。对于数值型参数，我们选择的范围涵盖文献中考虑的范围。对于每种算法，将宏平均 F 值达到最大值时对应的参数用于模型构建。采用 F 值进行评分，因为它能够反映分类模型识别不平衡数据集中的少数类（中风类）的能力 [12], [14]。所有实验均使用 Python 3.8[7] 完成。表中未列出的参数采用 Python 库的默认值。

3) 模型构建与评估：为开发和验证模型，采用 10 折交叉验证。我们通过准确率（公式 14）、精确率（公式 15–17）、召回率（公式 18–20）、F值（公式 21–23）、受试者工作特征（ROC）曲线、ROC曲线下面积（AUC），以及超参数调优和模型开发与验证的执行时间，评估数据平衡前后各模型的性能。

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (14)$$

$$P_{stroke} = \frac{TP}{TP + FP} \quad (15)$$

$$P_{no\ stroke} = \frac{TN}{TN + FN} \quad (16)$$

$$P_{MA} = \frac{P_{stroke} + P_{no\ stroke}}{2} \quad (17)$$

$$R_{stroke} = \frac{TP}{TP + FN} \quad (18)$$

$$R_{no\ stroke} = \frac{TN}{TN + FP} \quad (19)$$

$$R_{MA} = \frac{R_{stroke} + R_{no\ stroke}}{2} \quad (20)$$

$$F_{stroke} = \frac{2(P_{stroke} \times R_{stroke})}{P_{stroke} + R_{stroke}} \quad (21)$$

$$F_{no\ stroke} = \frac{2(P_{no\ stroke} \times R_{no\ stroke})}{P_{no\ stroke} + R_{no\ stroke}} \quad (22)$$

[7] https://devdocs.io/python~3.8/（最后访问于 2022 年 4 月 25 日）

表 II。 V ALUE(S) OF H文献和我们 L中使用的超参数 E实验所采用的 A算法 UNDER STUDY.

算法	超参数	文献中使用的值	我们实验中使用的值	总计组合
决策 Tree	分裂准则	熵 [3], [4] 和未报告 [5], [6], [8]	基尼和熵	8
	最大特征数	未报告 [3]–[6], [8]	无、自动、平方根和以 2 为底的对数	
随机森林	数量估计器/树	100 [4], 500 [5], 和未报告 [6], [7]	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200、300、400 和 500	112
	分裂准则	未报告 [4]–[7]	基尼和熵	
	最大特征数		无、自动、平方根和以 2 为底的对数	
支持向量机器	核函数	线性 [3], 径向基函数 [4], [5], 和未报告 [7]	径向基函数†	30
	正则化参数	0.1 [4], {2^{-2}, 2_0, 2_2, 2_4, 以及 2^6} [5], 和未报告 [3], [7]	0.1, 2^{-2}, 2_0, 2_2, 2_4, 以及 2^6	
	核系数	{2^{-6}, 2^{-4}, 2^{-2}, 2_0, 以及 2^2} [5] 和未报告 [3], [4], [7]	2^{-6}, 2^{-4}, 2^{-2}, 2_0, 以及 2^2	
逻辑回归	正则化参数	{2^{-6}, 2^{-4}, 2^{-2}, 2_0, 2_2, 2_4, 以及 2^6} [5] 以及未报告 [4], [6]	2^{-6}, 2^{-4}, 2^{-2}, 2_0, 2_2, 2_4, 以及 2^6	35
	求解器	未报告 [4]–[6]	Newton-CG、L-BFGS、Liblinear、SAG 和 SAGA	
	最大迭代次数		3000‡	

†Lin未考虑线性核，因为特征与中风之间的关系并非线性；‡该算法在迭代 3000 次之前未能收敛；NR：未报告 d

$$F_{MA} = \frac{2(P_{MA} \times R_{MA})}{P_{MA} + R_{MA}} \quad (23)$$

其中，TP（真正例）和 TN（真负例）分别表示被正确分类为正类（中风）和负类（无中风）的观测值数量。FP（假正例）和 FN（假负例）分别表示被错误分类为正类和负类的观测值数量。P_{stroke}、 $P_{no\ stroke}$ 和 P_{MA} 分别表示中风类、无中风类和宏平均精确率。同样，R_{stroke}、 $R_{no\ stroke}$ 和 R_{MA} 分别表示中风类、无中风类和宏平均召回率；F_{stroke}、 $F_{no\ stroke}$ 和 F_{MA} 分别表示中风类、无中风类和宏平均F值。

C. 实验结果分析

表 III 展示了数据平衡前后各算法获得的超参数最优值。图 6 展示了各算法用于卒中预测时获得的准确率。结果表明，RF 和 LR 的准确率最高，均为 98%；与其他算法相比，NB 的准确率最低，为 92%。数据平衡后，DT、RF 和 SVM 的准确率没有变化。另一方面，数据平衡后，NB 和 LR 的准确率有所下降。然而，混淆矩阵显示，在数据平衡前，DT、RF、SVM 和 LR 无法预测少数类（中风）。具体而言，DT、RF 和 SVM 分别只能正确预测少数类的 5.5 %（图 7(a)）、0.3%（图 7(b)）和 4.3%（图 7(c)）。此外，LR 将所有观测值都归为多数类（无中风）（图 7(e)）。相比之下，NB 能够正确预测少数类（中风）的 26.2%，如图 7(d) 所示。因此，与 NB 相比，数据平衡前 DT、RF、SVM 和 LR 的准确率较高（图 6），因为它们几乎将所有观测值都归为（多数类）无中风，从而获得较高的 TN 值，却无法预测少数类。然而，数据平衡后，所考虑的算法能够检测中风和无中风两类，如图 7(f)–7(j) 所示。具体而言，DT、RF 和 SVM 分别能够正确预测中风类的 97.2%（图 7(f)）、98.4%（图 7(g)）和 96.6%（图 (h)）。另一方面，NB 和 LR 分别能够正确分类中风类的 80.6%（图 (i)）和 80%（图 (j)）。

图 8(a) 和图 8(b) 分别展示了类别特定的精确率和召回率，以及宏平均精确率和召回率。对于所研究的算法，在多数类（非中风）中，精确率和召回率均高于 0.90。然而，对于少数类，决策树、随机森林、支持向量机和逻辑回归的这些指标均低于 0.1。这是因为这些算法无法识别超过 6% 的少数类样本。相比其他算法，朴素贝叶斯能够正确分类更多少数类（非中风）观测值，因此其少数类召回率较高。对所有算法进行数据平衡后，少数类的精确率和召回率均有所提高。数据平衡后，朴素贝叶斯和逻辑回归的精确率和召回率最低。图 8(c) 显示，数据平衡后，决策树、随机森林和支持向量机是最适合预测中风的算法，因为与朴素贝叶斯和逻辑回归相比，它们的类别特定 F值较高。

表 III。 我们在数据平衡前后实验中获得的超参数最优值。

算法	超参数	最优值	
		数据平衡前数据平衡	数据平衡后平衡
决策树	分裂准则	Gini	熵
	最大特征数	sqrt	none
随机森林	数量估计器/树	10	100
	分裂准则	Gini	熵
	最大特征数	none	none
支持向量机器	正则化参数	2^6	2^2
	核系数	2^{-4}	2^0
逻辑回归	正则化参数	2^6	2^2
	求解器	Newton-CG	L-BFGS

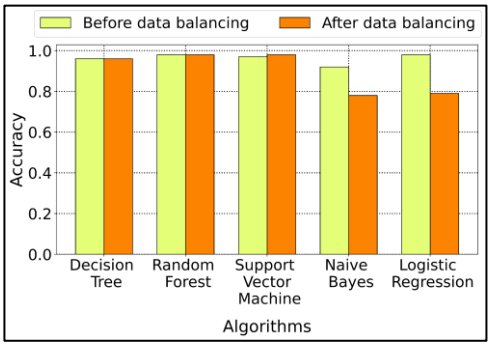

图 6. 数据平衡前后分类算法用于卒中预测的准确率。

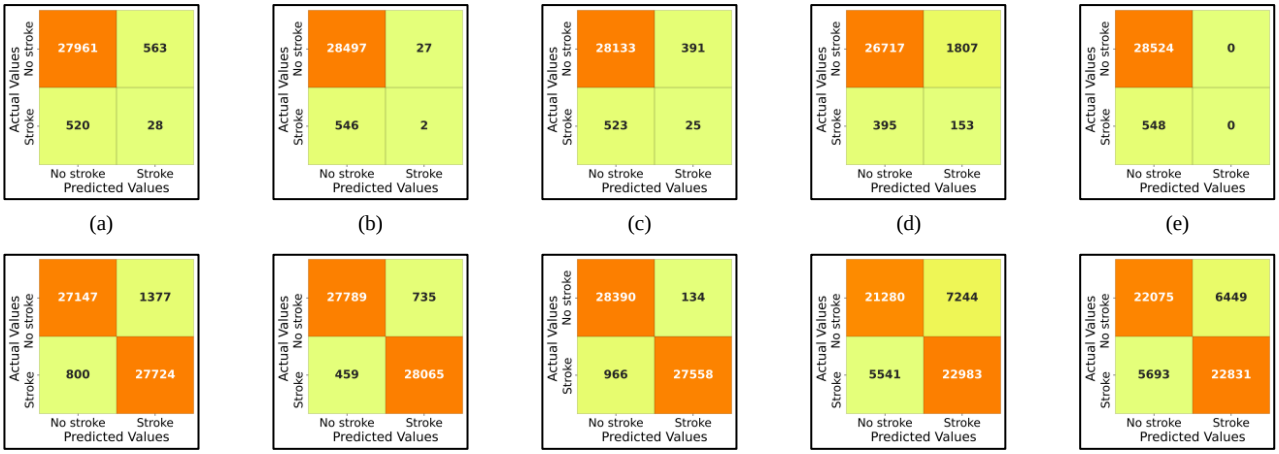

图7. 混淆矩阵：(a) 数据平衡前的决策树，(b) 数据平衡前的随机森林，(c) 数据平衡前的支持向量机，(d) 数据平衡前的朴素贝叶斯，(e) 数据平衡前的逻辑回归，(f) 数据平衡后的决策树，(g) 数据平衡后的随机森林，(h) 数据平衡后的支持向量机，(i) 数据平衡后的朴素贝叶斯，以及 (j) 数据平衡后的逻辑回归。

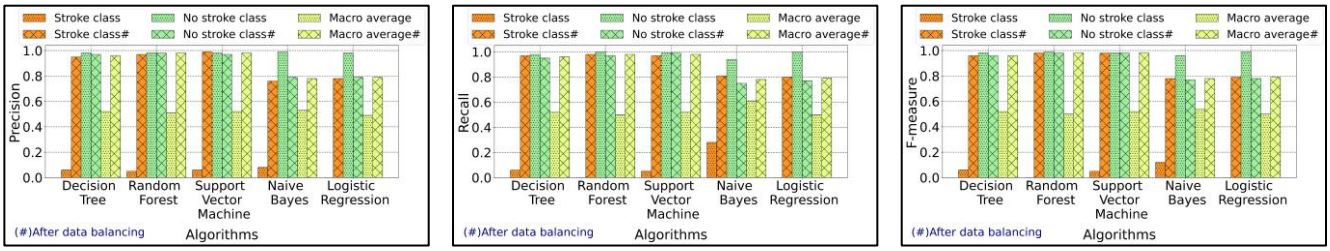

图8. 数据平衡前后分类算法用于卒中预测的类别特定及宏平均性能：(a) 精确率，(b) 召回率，以及 (c) F值。

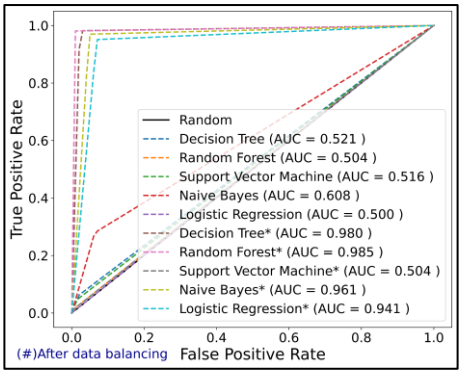

图9. 数据平衡前后分类算法用于卒中预测的受试者工作特征（ROC）曲线及ROC曲线下面积（AUC）。

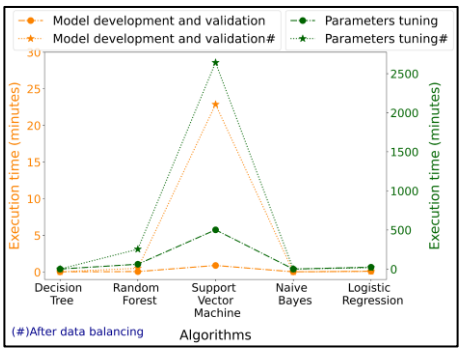

图10. 数据平衡前后分类算法进行超参数调优以及模型开发与验证的执行时间。

图 9 展示了各算法的 ROC曲线及曲线下面积值。结果表明，数据平衡前，除朴素贝叶斯外，所有算法的曲线下面积值均约为 0.500，与随机预测相同，因为这些算法无法识别少数类。另一方面，朴素贝叶斯的曲线下面积值最高，为 0.608，这表明它在数据平衡前能够识别少数类（中风）和多数类（非中风）。数据平衡后，随机森林的曲线下面积值最高，为 0.985。图 10 展示了数据平衡前后参数调优以及模型开发与验证的执行时间。结果表明，无论数据平衡前后，支持向量机的参数调优和模型开发耗时最长，而朴素贝叶斯耗时最短。数据平衡后，由于对少数类进行过采样使数据集中的观测值数量增加，执行时间也随之延长。

总而言之，在数据平衡前，朴素贝叶斯在少数类预测方面表现最佳；数据平衡后，随机森林的预测性能最佳。

VI. 结论

作为全球第二大死因和一场全球性危机，基于风险因素预测个体的卒中患病率至关重要。本文提出一种智能卒中预测框架，为辅助医疗专业人员提供预后和诊断支持，并帮助卫生利益相关者制定全国卒中预防计划。此外，本文采用一种

统一设置，对最常用于卒中预测的五种机器学习算法进行评估和比较，比较指标包括准确率、类别特定精确率、宏平均精确率、召回率、F值、受试者工作特征曲线、曲线下面积，以及超参数调优和模型开发与验证的执行时间。未来将考虑更广泛的机器学习模型。此外，还将研究深度学习模型在卒中预测中的表现。

致谢

本研究由阿联酋大学国家水与能源中心资助（资助编号：31R215）。

参考文献

[1] D. Kuriakose 和 Z. Xiao，《卒中的病理生理学与治疗：现状及未来展望》，Int. J. Mol. Sci.，第 21 卷，第 20 期，2020 年。[2] S.-M. Lai、S. Studenski、P. W. Duncan 和 S. Perera，《通过中风影响量表测量卒中的持续影响》，Stroke，第 33 卷，第 7 期，第 1840–1844 页，2002 年。[3] P. Songram 和 C. Jareanpon，《利用机器学习研究影响卒中预测的特征》，载于《人工智能多学科趋势国际会议》，2019 年，第 216–225 页，doi：10.1007/978-3-030-33709-4_19。[4] J. Yu、S. Park、H. Lee、C. S. Pyo 和 Y. S. Lee，《基于机器学习和美国国立卫生研究院卒中量表深度分析技术的老年人健康监测系统》，Mathematics，第 8 卷，第 7 期，第 1–17 页，2020 年，doi：10.3390/math8071115。[5] B. Letham、C. Rudin、T. H. McCormick 和 D. Madigan，《基于规则和贝叶斯分析的可解释分类器：构建更好的卒中预测模型》，Ann. Appl. Stat.，第 9 卷，第 3 期，第 1350–1371 页，2015 年，doi：10.1214/15-AOAS848。[6] T. Tazin、M. N. Alam、N. N. Dola、M. S. Bari、S. Bourouis 和 M. Monirujjaman Khan，《采用鲁棒学习方法进行中风疾病检测与预测》，J. Healthc. Eng.，2021 年，第 2021 卷，10 1155 2021 7633381 doi: . / / .[7] A. Subudhi 、 M. Dash 和 S. Sabut，《采用期望最大化和随机森林分类器对脑卒中进行自动分割与分类》，Biocybern. Biomed. Eng.，第 40 卷，第 1 期，第 277–289 页，2020 年，doi：10.1016/j.bbe.2019.04.004。[8] T. Kansadub、S. Thammaboosadee、S. Kiattisin 和 C. Jalayondeja，《基于人口统计数据的卒中风险预测模型》，载于《BMEiCON 2015——第八届生物医学工程国际会议》，2016 年，第 25–27 页，doi：10.1109/BMEiCON.2015.7399556。[9] L. Ismail、H. Materwala、A. P. Karduck 和 A. Adem，《生物医学护理与研究的健康数据管理系统需求：范围综述》，J. Med. Internet Res.，第 22 卷，第 7 期，2020 年，doi：10.2196/17508。[10] L. Pouchard，《重新审视大数据整理背景下的数据生命周期》，Int. J. Digit. CURATION，第 10 卷，第 2 期，2015 年，doi：https://doi.org/10.2218/ijdc.v10i2.342。[11] N. Abe、B. Zadrozny 和 J. Langford，《通过主动学习进行异常值检测》，载于《第 12 届 ACM SIGKDD 国际知识发现与数据挖掘会议论文集》，2006 年，第 504–509 页，doi：https://doi.org/10.1145/1150402.1150459。[12] L. Ismail 和 H. Materwala，《IDMPF：基于机器学习的智能糖尿病预测框架》，Appl. Comput. Informatics，2021 年，doi：https://doi.org/10.1108/ACI-10-2020- 0094。[13] H. Gupta 和 V. Asha，《高基数类别数据编码对解决预测问题的影响》，J. Comput. Theor. Nanosci.，第 17 卷，第 4197–4201 页，2020 年。[14] L. Ismail、H. Materwala、M. Tayefi、P. Ngo 和 A. P. Karduck，《人工智能与机器学习在2型糖尿病中的应用：方法与评估》，Arch. Comput. Methods Eng.，第 29 卷，第 313–333 页，2022 年，doi：https://doi.org/10.1007/s11831-021- 09582-x。[15] L. Ismail 和 H. Materwala，《2型糖尿病预测机器学习模型的比较分析》，载于《2020 年计算科学与计算智能国际会议（CSCI）》，2020 年，p. https://doi.org/10.1109/CSCI51800.2020.00095。[16] S. Susan 和 A. Kumar，《平衡技巧：不平衡数据集的优化抽样——近期最新进展简要综述》，Eng. Reports，第 3 卷，第 4 期，2021 年。