Hans Journal of Biomedicine 生物医学, 2025, 15(3), 545-554
Published Online May 2025 in Hans. https://www.hanspub.org/journal/hjbm
https://doi.org/10.12677/hjbm.2025.153063

Hans汉斯

基于多视图数据的出血性脑卒中智能诊疗预测模型研究

马广昊，于长青*

西京学院电子信息学院，陕西 西安

收稿日期：2025年4月9日；录用日期：2025年5月20日；发布日期：2025年5月30日

摘　要

由于出血性脑卒中起病急、进展快，预后较差，给社会患病患者及家属带来了沉重的负担，近年来引起临床广泛关注。出血性脑卒中主要有血肿扩张和血肿周围水肿发生和发展，早期发现并且提供有效的防治措施对患者治疗以及改善预后有很重要的意义。本题通过对于不同时期出血性脑卒中临床诊断的数据进行分析建模，使用聚类算法以及随机森林等方法比较出拟合程度最好的模型。

关键词

出血性脑卒中，聚类算法，随机森林算法，机器学习

Research on an Intelligent Diagnostic and Prognostic Prediction Model for Hemorrhagic Stroke Based on Multi-View Data

Guanghao Ma, Changqing Yu*

School of Electronic Information, Xijing University, Xi'an Shaanxi

Received: Apr. 9th, 2025; accepted: May 20th, 2025; published: May 30th, 2025

Abstract

Due to the rapid onset, progression, and poor prognosis of hemorrhagic stroke, it has brought a

*通讯作者。

文章引用: 马广昊, 于长青. 基于多视图数据的出血性脑卒中智能诊疗预测模型研究[J]. 生物医学, 2025, 15(3): 545-554. DOI: 10.12677/hjbm.2025.153063

heavy burden to patients and their families in society, and has attracted widespread clinical attention in recent years. Hemorrhagic stroke mainly involves the occurrence and development of hematoma dilation and perihematoma edema. Early detection and effective prevention and treatment measures are of great significance for patient treatment and improving prognosis. This question analyzes and models clinical diagnosis data of hemorrhagic stroke at different stages, and compares the best fitting model using clustering algorithms and random forest methods.

Keywords

Hemorrhagic Stroke, Clustering Algorithm, Random Forest Algorithm, Machine Learning

1. 引言

出血性脑卒中是一种严重的脑血管事件，由非外伤性脑实质内血管破裂引发，占全部脑卒中发病率的 10%~15%。它的发病急、进展快，急性期内的病死率高达 45%~50%，且大多数患者会遗留严重的神经功能障碍。在出血性脑卒中后，血肿的快速扩张和血肿周围的水肿发生是影响患者预后的关键因素，它们可能导致颅内压急剧增加，加重脑部损伤，从而进一步恶化患者的神经功能[1]。

当前，利用先进的医学影像技术和人工智能算法对出血性脑卒中的影像资料进行深入分析和研究，以早期识别和预测血肿扩张和周围水肿的发生及发展，具有重要的临床意义。这不仅有助于精准判断患者的病情风险，还可以为临床决策提供科学依据，从而实现个性化的治疗和管理，提高患者的生活质量和生存率[2]。

2. 数据预处理

2.1. 数据说明

本数据旨在研究出血性脑卒中患者的预后、血肿扩张风险以及血肿周围水肿的演变规律。

数据来源，数据集包括 160 例出血性脑卒中患者的信息，分为两个部分：训练数据集(100 例)和两个测试数据集(测试数据集 1 和测试数据集 2，各 30 例)。这些数据来自真实的临床情境，具有高度的现实性。患者信息，每个患者在数据集中都有唯一的标识号(ID)，用于区分不同患者。患者信息包括患者的年龄、性别以及一些基本的健康史信息。发病和治疗相关特征，数据集还包括了与发病时相关的特征，如血压(收缩压和舒张压)以及发病后首次影像检查的时间间隔。影像特征，数据集中包含了大量与血肿和水肿相关的影像特征。这些特征包括血肿和水肿的体积信息，位置信息，形状信息，以及灰度分布等。这些数据反映了出血性脑卒中患者的脑部病变情况，并提供了关键的影像学数据。

2.2. 缺失值、重复值、异常值处理

缺失值处理：对于连续变量(如血压)，按照性别、年龄段或是否治疗等分组分别计算均值，以避免数据被整体平均稀释。基于模型的插补：通过相似样本预测缺失值，KNN 插补特别适用于样本量不大的临床数据，能更好保留数据的结构特征。缺失值出现在具有时间先后顺序的变量中，可使用线性插值、前向填充等时序方法进行插补。

异常值处理：对于血压字段的异常字符，除了处理“/”外，应设定合理范围以筛除生理上不可能的

值。对于类别变量统一格式前先做 value_counts()检查可能的隐形异常。

特征工程的改进与深入：体积信息标准化，考虑将血肿、水肿体积分别按患者颅内容积或脑区体积进行归一化，使其具有更强的可比性。

3. 建模求解

3.1. 判断是否发生血肿扩张事件

患者 sub001 至 sub100 大致可以分为 2 类，第一类是在发病后 48 小时内进行多次的检查，第二类是在发病后 48 小时内只进行了一次检查[3]。

使用两个时间点的 HB 体积数据来确定恒定的增长率[4]。

$$\text{增长率} = \frac{\text{第二次检查的 HB 体积} - \text{入院首次检查的 HB 体积}}{\text{两次检查之间的时间差}} \tag{1}$$

使用计算得到的增长率来计算发病后 48 小时的 HB 体积，具体公式如下：

$$\text{HB 体积} = \text{入院首次检查的 HB 体积} + \text{增长率} \times \left(48\text{小时} - \text{发病到首次影像检查的时间}\right) \tag{2}$$

用计算出的 HB 体积与入院首次检查时的 HB 体积之间的百分比变化是否大于 33%来判断这 4 名患者是否发生了血肿扩张。具体公式如下[5]：

$$\text{变化率} = \frac{\text{HB 体积} - \text{入院首次检查的 HB 体积}}{\text{入院首次检查的 HB 体积}} \times 100 \tag{3}$$

$$\begin{cases} \text{变化率} < 33\% & \text{未发生血肿扩张} \\ \text{变化率} \geq 33\% & \text{发生了血肿扩张} \end{cases} \tag{4}$$

运用 Python 求解上述模型，结果如下表 1 所示。

Table 1. Model calculation results table
表 1. 模型计算结果表

病号	增长率	发病 48 小时后 HB 体积	变化率	是否发生血肿扩张
sub001	899.3932	110636.4	58.70039	发生了血肿扩张
……	……	……	……	……
sub098	658.1519	218079.9	15.91733	未发生血肿扩张
sub100	44.31641	16764.03	14.35998	未发生血肿扩张

由计算结果我们可以得知，第一类患者中在发病后的 48 小时进行过 2 次检查的患者中一共有 22 名患者出现了血肿扩张，剩余的 55 名患者在 48 小时内没有出现血肿扩张的现象。

为了计算出发病后多久出现了血肿扩张，我们设出现血肿扩张的时间为 t，只要求解出当 HB 体积比入院首次检查 HB 体积增加了 33%，即可认为这个时间就是发病血肿扩张出现的时间。方程如下所示[6]。发生血肿扩张的时间如下表 2 所示。

$$\text{入院首次检查 HB 体积} + \text{增长率} \times t = \text{入院首次检查 HB 体积} \times 1.33 \tag{5}$$

上述我们已经求解了患者发病后只进行 2 次检查的患者是否有血肿扩张，还有部分患者在 48 小时内进行了 3 次的检查，则会有 3 个 HB 的体积，之前假设的 HB 体积与时间成正比关系，就无法捕捉现在 HB 体积与时间之间复杂的关系。综合考虑下我们采用分段线性拟合的方式。

Table 2. Time of hematoma dilation
表 2. 发生血肿扩张的时间

患者 ID	*t*	患者 ID	*t*
sub001	25.6	sub054	17.3
sub017	12.3	sub077	9.6
sub038	12.2	sub081	22.0
sub048	11.8	sub095	2.2

部分患者的分段线性拟合图 1 如下：

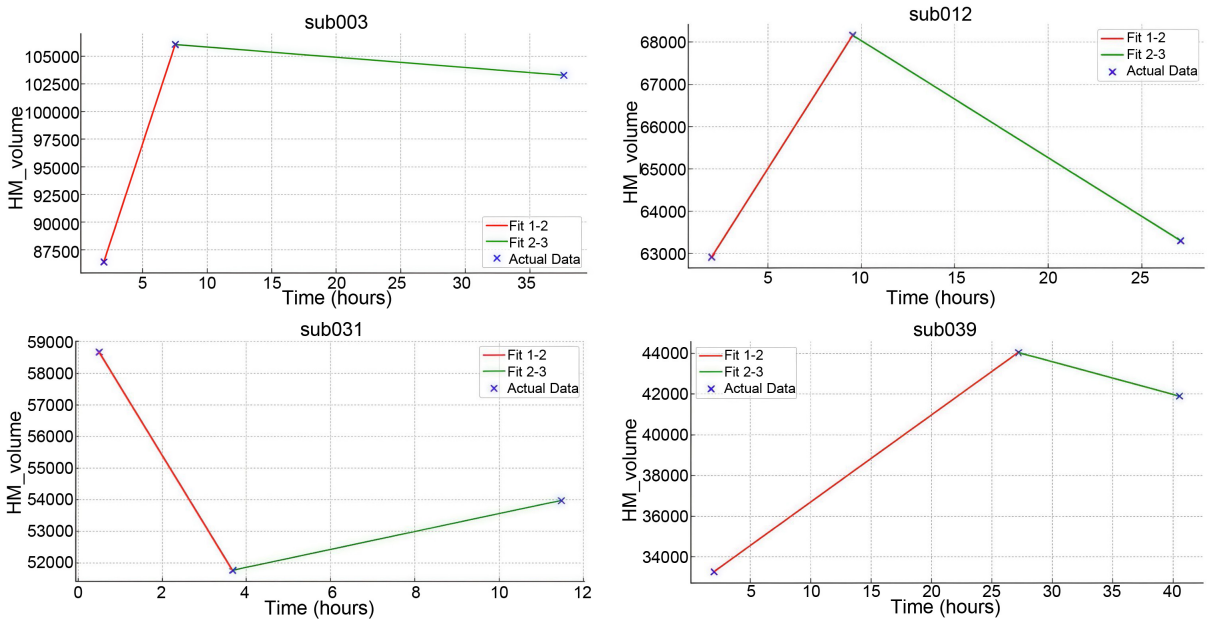

Figure 1. Segmented linear fitting plot of some patients
图 1. 部分患者的分段线性拟合图

如果 $k_{2-3}>0$ 并且第二次检查 HB 体积入院 ≤ 首次检查的 HB 体积 + 增加体积 ≤ 第三次检查 HB 体积，则有血肿扩张[7]。此时，扩张时间为：

$$\text{扩张时间} = \text{发病至随访 1 时间点} + \frac{\text{增加体积}}{k_{2-3}} \tag{6}$$

1) 在第三次检查之后的血肿扩张的判断

如果在第一次和第二次检查之间没有检测到血肿扩张，并且在第二次和第三次检查之间也没有检测到，则可以考虑在第三次检查之后的可能的血肿扩张，我们使用 k_{2-3} 来预测。剩余时间为：

$$\text{剩余时间} = 48 \text{ 小时} - \text{随访 2 时间点至发病的时间} \tag{7}$$

如果 $k_{2-3}>0$ 并且第三次检查 HB 体积 + k_{2-3} × 剩余时间 − 第一次检查 HB 体积 ≥ 增加体积，则有血肿扩张。此时，扩张时间为：

$$\text{扩张时间} = \text{发病至随访 2 时间点} + \frac{\text{增加体积}}{k_{2-3}} \tag{8}$$

通过运用 Python 求解以上模型，得出结果如下表 3 所示。

Table 3. Calculation results
表 3. 计算结果

患者 ID	是否发生血肿扩张(1 是，0 否)	血肿扩张时间
sub061	1	28.3
sub063	1	21.7
sub084	0	-
sub099	1	9.9

运用 Python 我们计算得出病号为 sub052 的病人在发病后约 7.3 小时内发生了血肿扩张。如下图 2 所示。

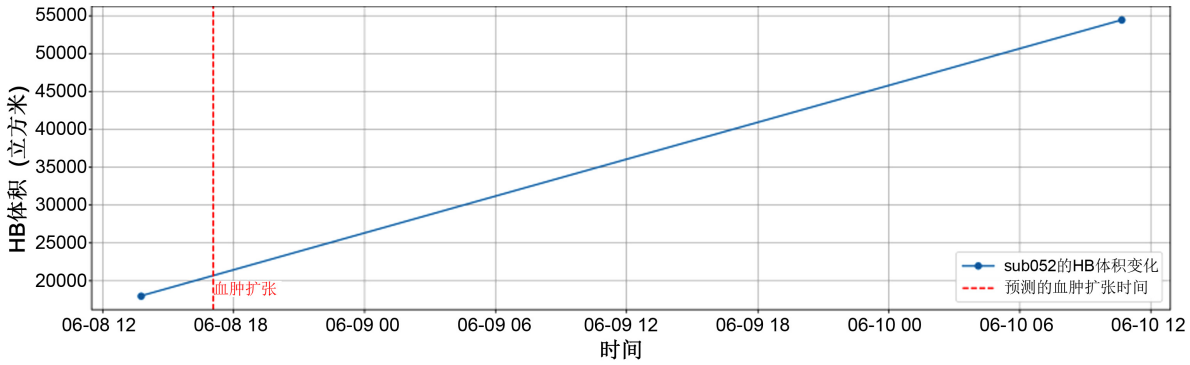

Figure 2. Time of hematoma expansion in patient sub052
图 2. 患者 sub052 血肿扩张的时间

下图 3 是这 5 人的血肿体积随时间变化的正比图，并在图中标识了发病 48 小时后的 HB 体积。

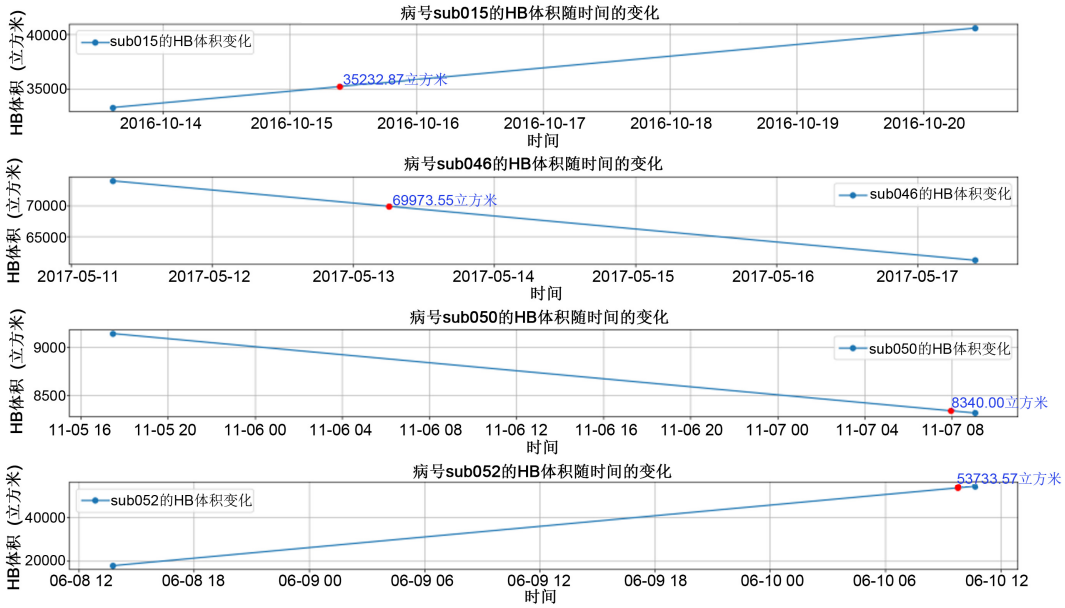

Figure 3. The rate of change in HB volume over time for four patients
图 3. 4 名患者 HB 体积随时间的变化率

3.2. 血肿扩张预测概率

对每一个患者 sub001 至 sub160 进行血肿扩张的概率计算，可以用分类预测模型计算每名患者发生血肿扩张的概率，分类预测模型中单一的模型的精度较为差，通常通过交叉验证用网格优化进行超参数的寻优，本文为了精准地进行预测采用 Stacking 融合模型进行概率的计算，并对每个基模型的参数通过交叉验证用网格优化进行超参数的寻优。

3.3. 模型的建立与优化

使用了 SMOTE (Synthetic Minority Over-sampling Technique)技术对训练数据进行过采样，旨在解决分类任务中的类不平衡问题。具体插值过程如下图 4 所示：

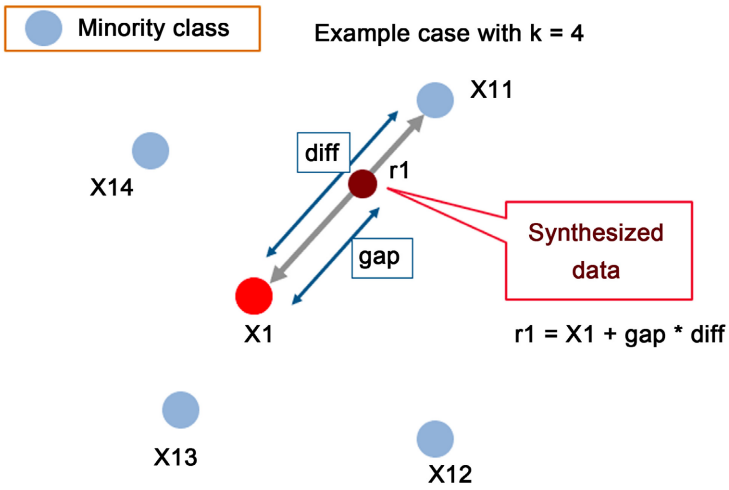

Figure 4. SMOTE synthesis process
图 4. SMOTE 合成过程

对于数据集中的每一个少数类样本 x。从少数类样本中随机选择一个最近邻样本 x_{nn}。生成一个随机数 λ，范围在 0 到 1 之间。

生成新的少数类样本 x_{new} 为：

$$x_{new} = x + \lambda \times \left(x_{nn} - x \right) \tag{9}$$

其中，$x_{nn} - x$ 是原始样本和其最近邻之间的差异，通过乘以 λ 我们得到了这个差异的一个随机部分。

定义了六个不同的分类器(SVM、随机森林、梯度提升树、XGBoost、LightGBM 和额外的树)作为基学习器。这些基学习器是初步的预测模型，它们各自独立地在原始数据上进行学习。

SVM (Support Vector Machine)：SVM 是一种用于分类和回归的监督学习方法。其目标是找到一个超平面，使得两个类别之间的边界最大化。

对于线性可分的数据，其基本形式为：

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad \text{subject to } y_i \left(w \cdot x_i + b \right) \geq 1, \forall i \tag{10}$$

其中，w 是超平面的法向量，b 是偏置项。

对于非线性数据，引入核技巧(kernel trick)将数据映射到更高维度的空间，使其线性可分。常见的核包括线性核、多项式核、径向基函数(RBF)核等。原理图 5 如下所示。

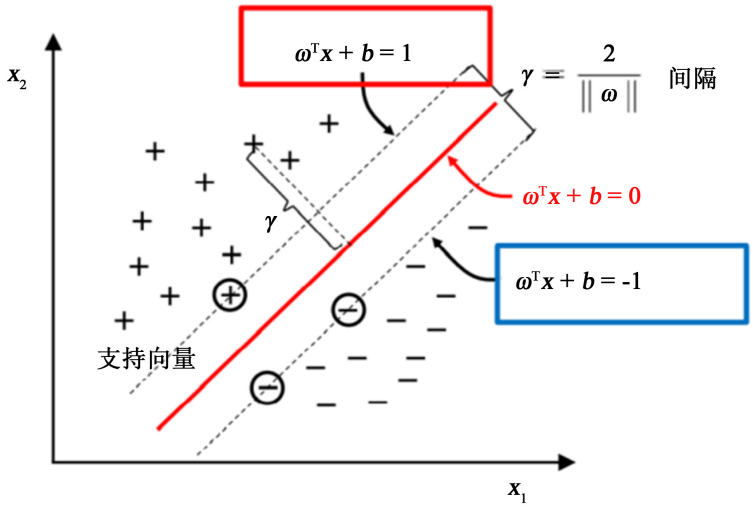

Figure 5. The principle of SVM
图 5. SVM 原理

随机森林(Random Forest)：随机森林是一个集成学习方法，它通过结合多个决策树的预测来提高整体模型的准确性并控制过拟合。梯度提升树(Gradient Boosting Machines, GBM)：GBM 是一个迭代的决策树算法，它通过拟合残差(真实值与当前预测的差)来改进模型。

XGBoost：XGBoost 是梯度提升算法的一个优化实现。除了标准 GBM 的策略外，它还考虑了正则化，并在特征分类上使用了更高级的策略，如最大深度、子样本比率等。

损失函数为：

$$\sum_k Loss\left(y_i, \hat{y}_i\right) + \sum_k \Omega\left(f_k\right) \tag{11}$$

其中，$\Omega\left(f_k\right)$ 是正则化项。

LightGBM (Light Gradient Boosting Machine)：LightGBM 是 GBM 的一个轻量级实现，特别适用于大数据。它使用了基于直方图的算法，可以快速找到最佳分裂点，并支持类别特征。

初始化堆叠分类器：使用 StackingClassifier 来构建堆叠模型。相关原理如下：设有 M 个基学习器。对于一个给定的输入样本 x，每个基学习器 m_i 会产生一个预测 p_i。因此，对于所有的基学习器，我们会有一个预测向量：

$$P = \left[p_1, p_2, \cdots, p_M\right] \tag{12}$$

这个预测向量 P 就是元学习器的输入特征。元学习器(在本文中是逻辑回归)会对这个预测向量进行学习，并输出最终的预测结果 y。如果元学习器是逻辑回归，其形式可以表示为：

$$y = \sigma\left(w_1 p_1 + w_2 p_2 + \cdots + w_M p_M + b\right) \tag{13}$$

其中，σ 是 Sigmoid 函数，它将任意实数值映射到 0 和 1 之间，用于二分类问题。w_i 是权重，反映了每个基学习器预测在最终预测中的重要性，而 b 是偏置项。

在 StackingClassifier 中，基学习器的预测首先通过交叉验证得到，然后这些交叉验证的预测被堆叠起来形成新的特征，这些新特征再被提供给元学习器进行学习。这确保了在训练元学习器时，它从未看到基学习器在这些特定数据上的预测，从而避免了过拟合。

定义参数网格：为了优化基学习器的性能，我们需要确定它们的参数。

假设我们有 n 个参数，每个参数有 m_i 个可能的值，其中 $i=1,2,\cdots,n$ 。那么，总的参数组合数量为：

$$C = m_1 \times m_2 \times \cdots \times m_n \tag{14}$$

在定义参数网格时，我们为每个基学习器定义了一系列参数值。

启动网格搜索时，它会自动生成并遍历所有这些参数的组合，为我们找到每个模型的最佳参数值。

本文的 Satcking 融合模型框架如下图 6 所示。

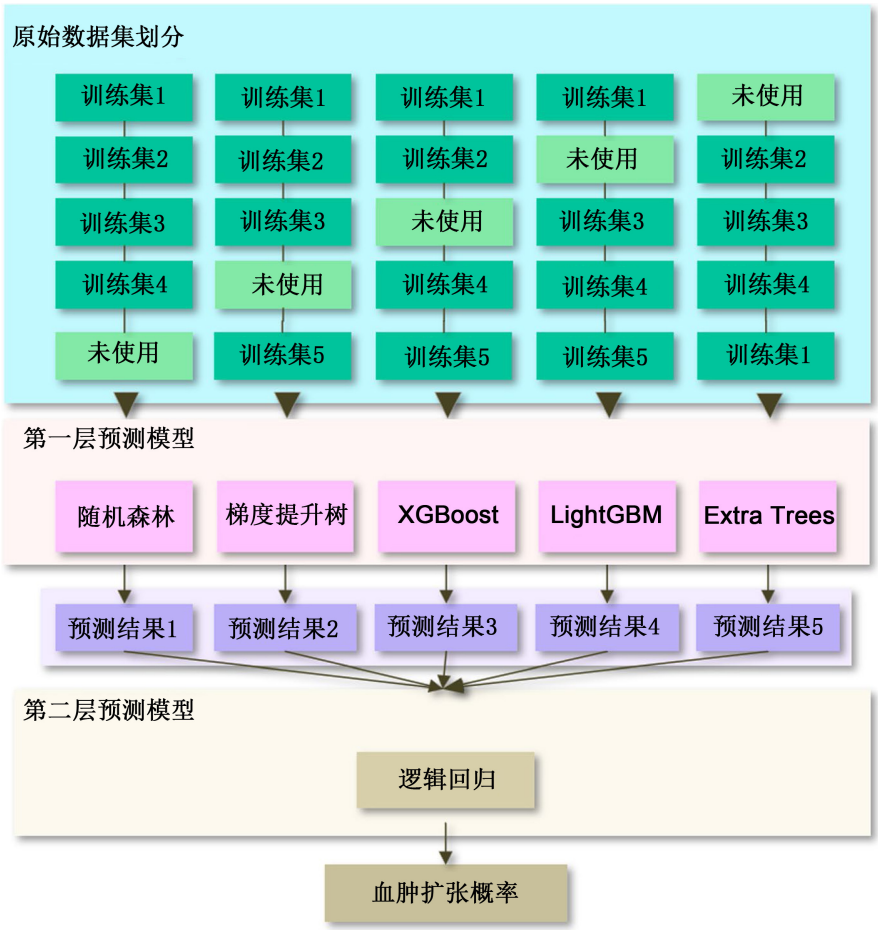

Figure 6. Stacking fusion model framework
图 6. Stacking 融合模型框架

3.4. 预测 sub001 至 sub100 患者发生血肿扩张的概率

当我们使用分类模型进行预测时，除了直接的类别预测之外，我们还可以得到数据点属于每个类别的概率。这些概率为我们提供了模型对其预测的置信度，并可以用于多种分析和决策应用中。

1) 预测概率

在许多分类算法中，对于给定的输入数据点 x，模型首先会计算它属于每个类别的概率。这些概率是基于模型内部的数学计算和函数得到的。

对于逻辑回归，给定数据点 x 和模型参数 θ，预测的概率 p 为：

$$p = \frac{1}{1+\mathrm{e}^{-\theta^{\mathrm{T}}x}} \tag{15}$$

其中 e 是自然对数的底。

在本文中，我们使用了 predict_proba 方法，这是 scikit-learn 中大多数分类器都有的一个方法，它返回一个数组，数组中的每行对应于输入数据的一个数据点，每列对应于一个类别的概率。

2) 获取目标变量为 1 的概率

从 probabilities_all 中，我们只选择了第二列的概率。这是因为在二分类问题中，predict_proba 方法返回两列，第一列是数据点属于类别 0 的概率，第二列是数据点属于类别 1 的概率。因此，我们通过索引[:, 1]提取了所有数据点属于类别 1 的概率。

3.5. 模型的评估

1) 模型评估函数

评估模型的性能对于理解模型在未见过的数据上的表现是至关重要的。评价指标为我们提供了模型性能的量化表示，可以帮助我们在不同模型之间做出选择或调整模型的参数。定义了一个 evaluate_model 函数，它接受一个模型、测试数据 X_test 和真实标签 y_test。这个函数首先对测试数据进行预测，得到类别预测 y_test_pred 和属于类别 1 的概率 y_test_prob。然后，函数计算了五个关键的评价指标：准确率、精确率、召回率、F1 分数和 AUC。

2) 对每个基模型和堆叠模型进行评估

对于每个基模型，我们首先使用过采样的训练数据进行训练。使用预先定义的 evaluate_model 函数来评估其在测试数据上的性能。通过比较所有模型的评价指标，我们发现 Stacking 融合模型的准确度最高，我们得出的发病概率较为准确。

3) 可视化工具：ROC 曲线和精确度-召回率曲线

为了更好地了解模型的性能，分别绘制了 ROC 曲线和精确度-召回率曲线。定义了一个 plot_roc 函数，该函数为给定的模型绘制 ROC 曲线。首先使用 predict_proba 方法获取模型预测的概率，然后使用 roc_curve 函数计算真正例率和假正例率，并绘制 ROC 曲线。定义了一个 plot_precision_recall 函数，该函数为给定的模型绘制精确度－召回率曲线。使用 precision_recall_curve 函数计算精确度和召回率，并绘制曲线。如下图 7 所示。

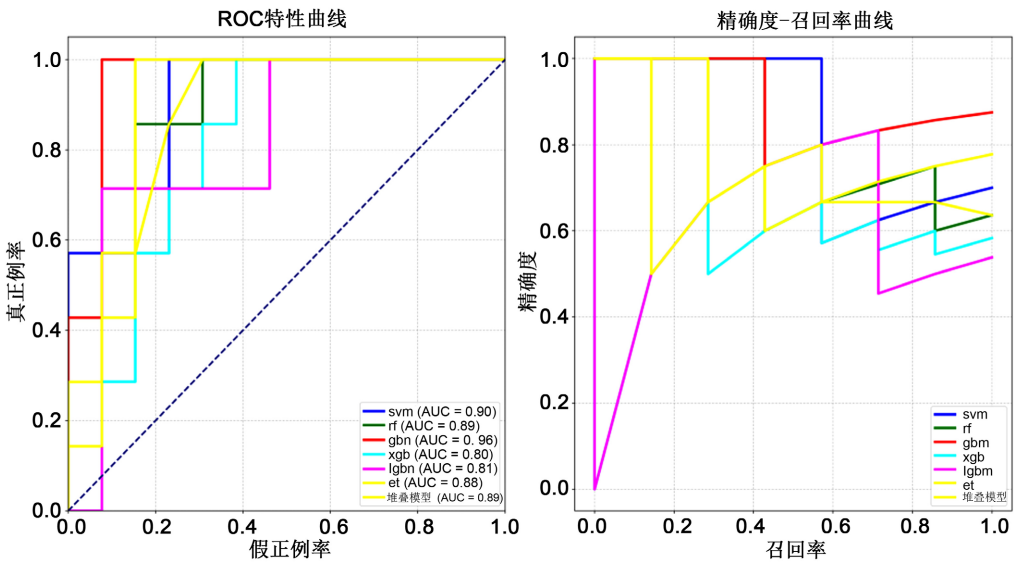

Figure 7. Relevant performance chart
图 7. 相关性能图

3.6. 预测 sub101 至 sub160 患者发生血肿扩张的概率

在数据整合的环节，我们还得到了“新分类预测.xlsx”其中包含了 sub101 至 sub160 患者首次检查的信息，但是没有“是否发生血肿扩张”这一目标变量，我们可以用上面训练好的模型进行发病率的预测。

4. 结果讨论

4.1. 临床意义

本研究通过构建融合临床、影像等多源数据的智能诊疗模型，有效实现了对出血性脑卒中患者血肿扩张风险预测及 90 天 mRS 评分评估的目标，具有如下重要临床价值：辅助早期干预，模型可在患者发病初期结合影像与临床指标对血肿扩张风险进行预警，辅助医生制定更具前瞻性的治疗策略，如及时调整降压方案、评估是否需要脑室引流等干预措施。精准化康复管理，mRS 评分预测为康复期治疗和护理分级提供重要参考，有助于评估患者功能预后，为个体化康复路径规划提供决策依据。提升医疗资源利用效率，通过智能模型对重症风险进行分层，有助于优化 ICU 床位分配、提前布局多学科会诊资源，从而提升临床资源使用效率。促进智能诊疗系统发展，本研究探索了 Stacking 融合算法在小样本临床数据中的应用策略，为类似疾病智能辅助诊断系统的建设提供了可推广的技术路径和实现框架。

4.2. 局限性分析

尽管本研究取得了一定成果，但仍存在以下几个局限性，需在后续工作中进一步优化：样本量相对较小，本研究基于 160 例患者数据，训练集与测试集数量有限，可能对模型泛化能力造成一定影响。后续应在更大样本、多中心数据上进一步验证模型稳定性。特征来源单一、部分数据缺失，虽然整合了多个特征维度，但部分变量仍存在缺失，且生化、基因组等深层次数据暂未纳入，限制了对病理机制的更全面建模。影像特征维度高但解释性差，部分自动提取的影像特征虽然能提升模型性能，但在临床解释性上仍存在“黑箱”问题，有待引入可解释 AI 增强模型透明度。

4.3. 展望

推进模型在真实临床流程中的集成与测试，结合电子病历系统实现实时预测与反馈。探索多模态建模以增强诊疗智能化水平；持续扩大样本量并开展多中心验证研究，推动模型的标准化和可推广应用。

参考文献

[1] 梁奕, 柳柏玉, 陶静雄, 等. 基于 CT 图像的深度学习预测高血压脑出血早期血肿扩大的应用研究[J]. 临床放射学杂志, 2023, 42(9): 1388-1392

[2] 薛光, 白咏, 常小娜, 等. 基于 CT 影像特征的脑出血病人术后早期脑积水发生风险预测模型的效能评估[J]. 中西医结合心脑血管病杂志, 2023, 21(18): 3458-3462.

[3] 乔光念, 肖遥, 戴大鹏, 等. 脑出血患者血肿扩大相关因素的研究进展[J]. 中国现代医生, 2023, 61(24): 133-137.

[4] 李翔, 闻海兵, 李升, 等. 脑膜中动脉阻断 + 颞肌贴敷术治疗机化型慢性硬膜下血肿的疗效[J]. 中国临床神经外科杂志, 2023, 28(8): 490-492, 496.

[5] 沈锦明. 过度通气与呋塞米联用对重症颅脑损伤患者脑水肿、神经功能及预后的影响[J]. 医学理论与实践, 2023, 36(18): 3106-3109.

[6] 徐华伟, 赵施竹. 脑脊液容量定量评估大脑半球大面积梗死病人脑水肿程度的临床价值[J]. 中西医结合心脑血管病杂志, 2023, 21(17): 3269-3272.

[7] 王强, 王萃, 赵燕, 等. 高血压脑出血术后脑水肿患者血清 CC 类趋化因子配体 5 和纤维蛋白原样蛋白 2 水平变化及意义[J]. 陕西医学杂志, 2023, 52(9): 1181-1185.