

基于XGBoost算法的出血性脑卒中患者 血肿扩张风险预测

钟玮琦¹, 曾海涛²

¹上海理工大学环境与建筑学院, 上海

²宜宾市第二人民医院康复医学中心, 四川 宜宾

收稿日期: 2024年6月12日; 录用日期: 2024年7月5日; 发布日期: 2024年7月12日

摘 要

出血性脑卒中血肿扩张事件的发生在医学上是一个重要的治疗时间点, 如果发现及时可以减少脑卒中患者发生不良预后结果的概率。本文使用XGBoost算法对出血性脑卒中患者发生血肿扩张的风险进行预测, 模型训练完成后, 使用混淆矩阵、ROC曲线、AUC值、准确率、精确率、召回率和F1_score等指标进行模型的性能评估。通过对130名患者的数据进行研究后发现, 基于XGBoost算法的血肿扩张预测模型的预测准确率为80%。其中特征向量数为32时, 预测模型的表现最佳, 各项性能指标均优于其他特征向量数。本研究纳入模型建立的患者样本数量较少, 后续可考虑增加样本量从而进一步提高模型的预测准确率。综上所述, 基于XGBoost算法的血肿扩张预测模型可为医疗人员在临床医治时提供参考, 具有一定的临床应用价值。

关键词

XGBoost模型, 出血性脑卒中, 血肿扩张

Risk Prediction of Hematoma Expansion in Hemorrhagic Stroke Patients Based on XGBoost Algorithm

WeiQi Zhong¹, Haitao Zeng²

¹School of Environment and Architecture, University of Shanghai for Science and Technology, Shanghai

²Center for Rehabilitation Medicine, Yibin Second People's Hospital, Yibin Sichuan

Received: Jun. 12th, 2024; accepted: Jul. 5th, 2024; published: Jul. 12th, 2024

文章引用: 钟玮琦, 曾海涛. 基于 XGBoost 算法的出血性脑卒中患者血肿扩张风险预测[J]. 建模与仿真, 2024, 13(4): 4271-4278. DOI: 10.12677/mos.2024.134387

Abstract

The occurrence of hemorrhagic stroke hematoma dilatation events is an important therapeutic time point in medicine, and if detected in a timely manner can reduce the probability of adverse prognostic outcomes in stroke patients. In this paper, the risk of hematoma dilatation in hemorrhagic stroke patients was predicted using the XGBoost algorithm, and after the model training was completed, the performance of the model was evaluated using metrics such as confusion matrix, ROC curves, AUC values, accuracy, precision, recall, and F1_score. After studying the data of 130 patients, it was found that the prediction accuracy of hematoma expansion prediction model based on XGBoost algorithm was 80%. Among them, the best performance of the prediction model was achieved when the number of feature vectors was 32, and all the performance metrics were better than the other number of feature vectors. The number of patient samples included in this study for modeling is small, and increasing the sample size can be considered to further improve the prediction accuracy of the model. In conclusion, the hematoma expansion prediction model based on the XGBoost algorithm can provide reference for medical personnel in clinical treatment and has certain clinical application value.

Keywords

XGBoost Model, Hemorrhagic Stroke, Hematoma Expansion

Copyright © 2024 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

脑卒中(Stroke)是一种严重危害人类健康和生活质量的疾病,其类型主要分为缺血性脑卒中和出血性脑卒中两类,而出血性脑卒中相较于缺血性脑卒中更为致命。出血性脑卒中是指由于脑脊动脉破裂导致脑实质内、蛛网膜下腔和脑室内出血,患者一旦发生出血性脑卒中大约会有 75% 概率死亡,每年约有 280 万人因为这种疾病死亡[1]。已有研究表明,血肿扩张会显著导致出血性脑卒中患者的预后情况不良[2]。因此,血肿扩张事件的发生在医学上已经成为了重要的治疗时间点[3]。

如果可以及时预测脑卒中患者发生血肿扩张的情况,就可以减少脑卒中患者发生不良预后结果的概率。当前,针对出血性脑卒中患者发生血肿扩张事件的预测研究已经成为了研究热点,许多研究将与血肿扩张事件发生的高相关因素作为预测特征,使用建模预测方法来预测血肿扩张事件发生的概率。Chan [4]等人对 257 例出血性脑卒中患者进行研究,收集了患者的 23 项临床和影像学特征,经过特征选择后选取 2 项特征纳入训练,利用多变量 logistic 回归方法建立血肿扩张事件预测模型,研究结果显示模型预测结果的 AUC 值为 0.67。Miyahara [5]等人收集了 622 名出血性脑卒中患者的 23 项临床特征,选取 4 项特征作为模型的预测特征,预测模型使用多变量 logistic 回归方法建立,结果显示该模型的 AUC 值为 0.80。袁伟[6]等人使用机器学习算法中的支持向量机、逻辑回归和极限学习机 3 种模型与临床上常用的 BRAIN 评分进行对比,比较机器学习方法和 BRAIN 评分在预测血肿扩张概率上的差异,研究结果发现机器学习方法可以用于预测血肿扩张概率,其的预测准确性要大于 BRAIN 评分。刘雨安[7]等人对机器学习模型在疾病预测方面的应用进行了研究,结果表明机器学习模型在疾病预测和辅助临床决策方面具有明显的优势,适用于预测疾病的发生风险、预后风险分层、慢性病进展以及治疗效果的预测。

XGBoost 算法是属于集成学习中的一种算法, 其由陈天奇于 2014 年提出, 该模型具有较强的泛化能力和较高的拓展性, 应用场景广泛且训练速度快, 预测准确性高[8]。由于 XGBoost 算法具备优异性能, 目前已在许多领域被广泛应用, 例如临床上针对疾病的诊断、疾病发生风险预测、安全用药剂量预测等方面[9]。Torlay [10]等人将 XGBoost 算法应用于影像学诊断领域, 使用该算法对癫痫患者 MRI 数据进行分析, 从而识别患者的癫痫病灶区。Nishio [11]等人利用 XGBoost 算法对患者的肺结节进行辅助诊断, 并将该算法与支持向量机算法的预测准确性进行对比, 结果发现 XGBoost 算法的预测性能要优于支持向量机算法。Ye [12]等人将 823,627 名患者的数据纳入 XGBoost 算法中, 建立了一个预测未来 1 年内发生原发性高血压风险的预测模型, 研究结果表明预测准确率可达 87%。Liu [13]等人使用 XGBoost 算法对骨关节患者历史病历数据进行数据挖掘, 建立镇痛药对骨关节炎患者副作用的预测模型。采集了 2213 名疫苗受试者的相关信息, 将疫苗类型、受试者信息、受试者副作用等相关数据纳入 XGBoost 模型中构建疫苗副作用预测模型, 研究结果发现 XGBoost 疫苗副作用预测模型的预测准确率为 79%, 可用于提前预判接种该疫苗会产生严重不良反应的受试者。综上所述, 目前 XGBoost 算法在医药领域中的相关研究与应用, 表明了该算法具备很好的预测性能与应用潜力, 适用于临床诊疗或其他领域, 辅助医务人员更好地为患者提供个性化精准医疗。

当前, 较少有研究将 XGBoost 算法用于预测出血性脑卒中患者发生血肿扩张事件的概率。而 XGBoost 算法具有的诸多特性, 使其具备应用于预测出血性脑卒中患者发生血肿扩张事件的潜力, 因此本文使用真实世界的出血性脑卒中患者数据集来训练 XGBoost 算法, 构建出血性脑卒中患者血肿扩张预测模型, 预测出血性脑卒中患者发病 48 小时内是否发生血肿扩张的概率, 从而为医疗人员进行临床医治时提供一定参考。

2. 研究对象与方法

2.1. 研究对象

本研究所使用的数据集来自于 130 名出血性脑卒中患者的真实数据, 数据集中包含每个患者的个人信息、疾病史、发病情况、治疗相关信息、多次 CT 平扫结果及影像学检查数据。影像学检查数据包括: 患者在各个时间点血肿/水肿的体积、位置、形状特征及灰度分布等信息。

2.2. 基于 XGBoost 算法的血肿扩张概率预测模型

2.2.1. XGBoost 算法概述

集成学习思想由 Dasarathy 和 Sheela 于 1979 年提出, 其思想类似于现实社会中的人们在面临一项复杂任务时, 通过多方讨论后达成共识, 最后进行决策的过程。例如, 现实世界中公民通过投票来决定官员的选举和法律的制定; 进行重大手术前需要咨询多名医生进行会诊。不同类型的集成学习算法之间的区别主要在于基学习器的训练方法与组合策略不同。集成学习算法主要可以分为三类: Bagging、Boosting、Stacking。本研究使用 XGBoost 算法属于集成学习中的 Boosting 类型。Boosting 也被称为增强学习, 它是集成学习算法中的一种重要框架。

XGBoost (eXtreme Gradient Boosting) 是极端梯度上升的简称, 它由陈天奇在 2014 年基于梯度提升框架提出, 具有线性模型求解器和树学习算法的特点。该算法的核心在于它对梯度提升算法进行了优化, 通过引入正则化项来防止模型复杂度过高, 同时它还支持用户自定义优化目标和评价标准, 这使得 XGBoost 算法在各种不同的机器学习任务中都具有很高的灵活性和适用性, 不仅提高了模型的训练速度, 还减少了模型的过拟合风险。由于 XGBoost 算法具有的诸多优秀特点, 该算法已经被广泛应用于医疗领域, 例如疾病诊断、疾病风险预测、疾病预后预测、用药剂量预测、临床辅助决策等[9]。

XGBoost 算法基本原理为：将损失函数的二阶泰勒展开用来近似当前模型的负梯度，该负梯度随后被当成前一模型的残差项进行学习；为了增强模型对于上一轮迭代中预测准确度较低的样本的学习能力，这些样本在训练过程中会被赋予更高的权重；采用逐步迭代的串行方式进行训练，直到模型的误差得到有效修正，最终使损失函数的值达到预定的收敛标准[7]。

XGBoost 算法的目标函数为：

$$\hat{\Theta}_J \approx \arg \min_{\Theta_J} \left\{ \sum_{i=1}^N \left[L(y_i, \hat{y}_j^{(j-1)}) + v g_i^{(j)} f_j(x_i; \Theta_j) + \frac{1}{2} v^2 h_i^{(j)} f_j^2(x_i; \Theta_j) \right] + \gamma M_j + \frac{1}{2} \lambda \sum_{k=1}^{M_j} (w_k^{(j)})^2 \right\} \quad (1)$$

其中， N 表示模型训练样本量； $L(y, \hat{y}) = (\hat{y} - y)^2$ 表示为损失函数的平方； $w_k^{(j)}$ 是指第 j 棵回归树中，第 k 个叶子节点的权重； M_j 是指第 j 棵回归树上的叶子节点数量； γ 是指每分裂一个叶子节点时产生的最小损失量； λ 为正则化参数，用于调节叶子节点得分的复杂度。

公式(1)中的 $g_i^{(j)}$ 、 $h_i^{(j)}$ 见下公式(2)，(3)：

$$g_i^{(j)} = \frac{\partial L(y_i, \hat{y}_j^{(j-1)})}{\partial \hat{y}_j^{(j-1)}} = 2(\hat{y}_j^{(j-1)} - y_i) \quad (2)$$

$$h_i^{(j)} = \frac{\partial^2 L(y_i, \hat{y}_j^{(j-1)})}{(\partial \hat{y}_j^{(j-1)})^2} = 2 \quad (3)$$

2.2.2. 训练数据预处理

本文根据患者的临床治疗信息和血肿的体积及位置的影像信息，对 130 名患者的数据集进行预处理，判断患者发病后 48 小时内是否发生血肿扩张事件及血肿扩张发生时间。首先进行发病时间到首次影像时间和各后续影像时间的时间间隔统计，确认各患者位于 48 小时内的各影像检查，并记录首次影像和发病后 48 小时内的后续影像的血肿体积，计算绝对体积变化和相对体积变化。只要血肿体积数值的绝对体积超过首次检查体积的 6 ml 或者相对体积超过首次检查体积的 33%，即可认定为该患者发生了血肿扩张。由于患者的影像数据拍摄时间间隔较长，当根据影像数据计算结果发现血肿扩张事件发生时，患者可能已经在进行影像学检查前已出现血肿扩张情况。为了较为准确地计算患者血肿扩张实际发生时间，本文采用线性回归的方法进行拟合计算，首先根据首次影像时间和后续随访影像时间计算得到血肿扩张事件发生的时间点，然后拟合影像拍摄时间间隔与血肿体积的变化大小之间的函数关系，得到患者血肿体积随时间变化的方程。最后利用回归方程计算患者发生血肿扩张事件的准确时间点，再计算该时间点距离患者发病时刻的间隔，得到血肿扩张事件发生时刻距离患者发病时刻的时间间隔，具体流程见图 1。本研究训练集数据和测试集数据的划分比例为 8:2。首先将前 100 例患者的数据进行训练，模型训练完成后，使用后 30 例患者的数据进一步测试血肿扩张预测模型的性能。

2.2.3. 模型参数设置与特征选择

由于本研究使用的患者数据集中共有 103 项特征，包括患者的个人信息、疾病史、发病情况、治疗相关信息及影像学检查数据，但是并非所有的特征都与血肿扩张发生具有强相关性。因此需要对这些特征进行筛选，选出与患者发生血肿扩张最相关的特征，再纳入 XGBoost 模型中进行训练。本研究使用 SelectKBest 方法来筛选数据集中与血肿扩张发生有关的重要特征，以此减少模型的复杂度，降低模型发生过拟合的风险，从而提高模型的计算效率。SelectKBest 是一种基于统计学的特征选择方法，它通过对每个特征进行统计分析，计算出它们与结果之间的关系程度，并根据得分从高到低进行排序，从而选择前 K 个最佳特征。本文选择 SelectKBest 中适合分类问题的 F 检验分类(f_classif)进行特征筛选。

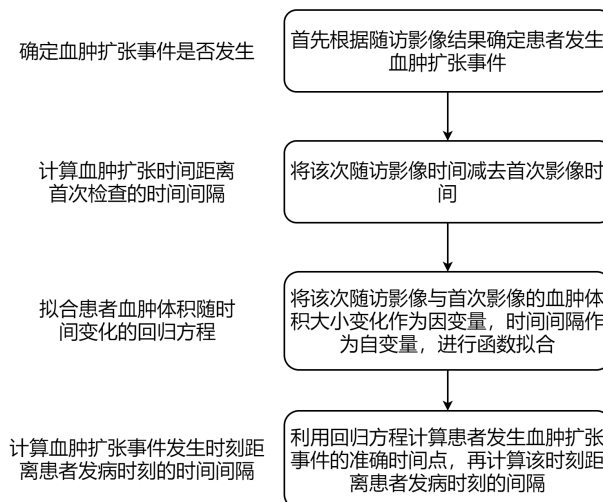


Figure 1. Procedure for calculating the time taken from the moment of the hematoma expansion event to the moment of the patient's onset of symptoms

图 1. 血肿扩张事件发生时刻距离患者发病时刻所用时间的计算流程

本文使用 GridSearchCV 方法自动搜索并设置 XGBoost 模型的超参数, 为模型选择出在验证集上表现最佳的参数配置, 从而避免人工调参的庞大工作量和错误设置。GridSearchCV 方法是一种广泛应用于机器学习领域的超参数优化技术, 它通过遍历给定参数网格的方式, 结合 k 折交叉验证来评估每一组参数的性能, 从而寻找最佳的超参数组合。该方法的核心在于其简单性和有效性, 尤其是在参数搜索空间较小且模型训练成本不高的情况下, 因此适合用于本研究模型的超参数设置。超参数寻找过程: 首先定义了一个参数网格, 该网格包含了所有需要优化的超参数及其可能的取值, 然后使用 GridSearchCV 方法遍历所有参数组合, 并对每一组参数使用交叉验证来评估模型的性能, 最后根据交叉验证得到的平均分数, 得到最优的参数组合。对于本研究的 XGBoost 血肿扩张概率预测模型的性能评估, 采用混淆矩阵、ROC 曲线、AUC 值、准确率、精确率、召回率和 F1_score 等指标进行评估。

3. 研究结果

首先使用 SelectKBest 方法对已经完成预处理的数据集进行特征筛选, 从 103 个特征中选择与血肿扩张发生最为相关的重要特征。对患者个人信息(年龄、性别等)、个人以往疾病史(高血压病史、房颤史等)、首次影像学数据等一共 103 个特征进行筛选。本研究分别选取 103、45、32 和 16 等 4 种特征数量, 对比不同特征数情况下 XGBoost 血肿扩张概率预测模型的预测性能, 从而确定最适宜的特征数量。再使用 GridSearchCV 方法分别对 4 种特征数量情况时的预测模型进行超参数设置。本文使用前 100 例患者的相关病历信息进行 XGBoost 血肿扩张概率预测模型的训练, 训练完成后再采用后 30 例患者的相关病历信息测试模型的预测性能。不同模型完成训练和验证后, 使用混淆矩阵、ROC 曲线、AUC 值、准确率、精确率、召回率和 F1_score 等指标进行预测性能评估, 比较结果见表 1、图 2、图 3。

从表 1 的结果中可以发现, 当特征参数为 32 个时, XGBoost 预测模型的性能最佳, 各项性能指标的数值均为最大值。表 2 为特征数量为 32 时, XGBoost 血肿扩张概率预测模型的最佳超参数值。图 2 和图 3 为不同特征数量下血肿扩张的预测模型 ROC 曲线和混淆矩阵。图 4 为与血肿扩张事件发生最相关的 32 个特征排序, 可以看出房颤史与是否发生血肿扩张事件具有很强的关联性, 之后发病到首次影像时间的间隔、患者饮酒史等。综上所述, 基于 32 个特征建立的 XGBoost 血肿扩张概率预测模型的预测性能是最佳的, 模型的预测准确性为 80%。

Table 1. Performance comparison of algorithms with different number of features
表 1. 不同特征数量的算法性能对比

模型	特征数量	准确率	AUC	精确率	召回率	F1_score
默认参数 XGBoost	103 项	0.80	0.37	0.80	1.0	0.89
	103 项	0.70	0.41	0.78	0.87	0.82
优化参数 XGBoost	45 项	0.80	0.66	0.80	1.0	0.89
	32 项	0.80	0.68	0.80	1.0	0.89
	16 项	0.80	0.59	0.80	1.0	0.89

Table 2. Optimal hyperparameter value of the model when the number of features is 32
表 2. 特征数量为 32 时模型最佳超参数值

参数	取值
learning_rate	0.07
n_estimators	200
gamma	0
reg_lambda	1
subsample	1
min_child_weight	1
max_depth	3
max_delta_step	0

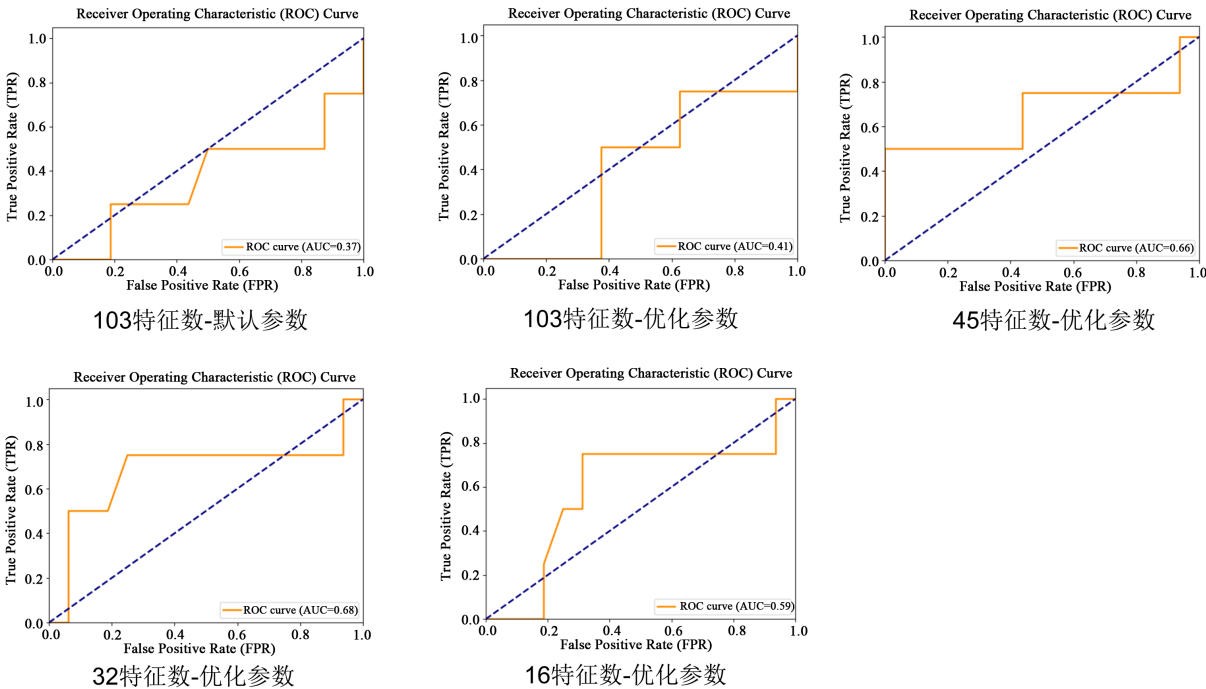


Figure 2. ROC curves for hematoma expansion prediction models
图 2. 血肿扩张预测模型的 ROC 曲线

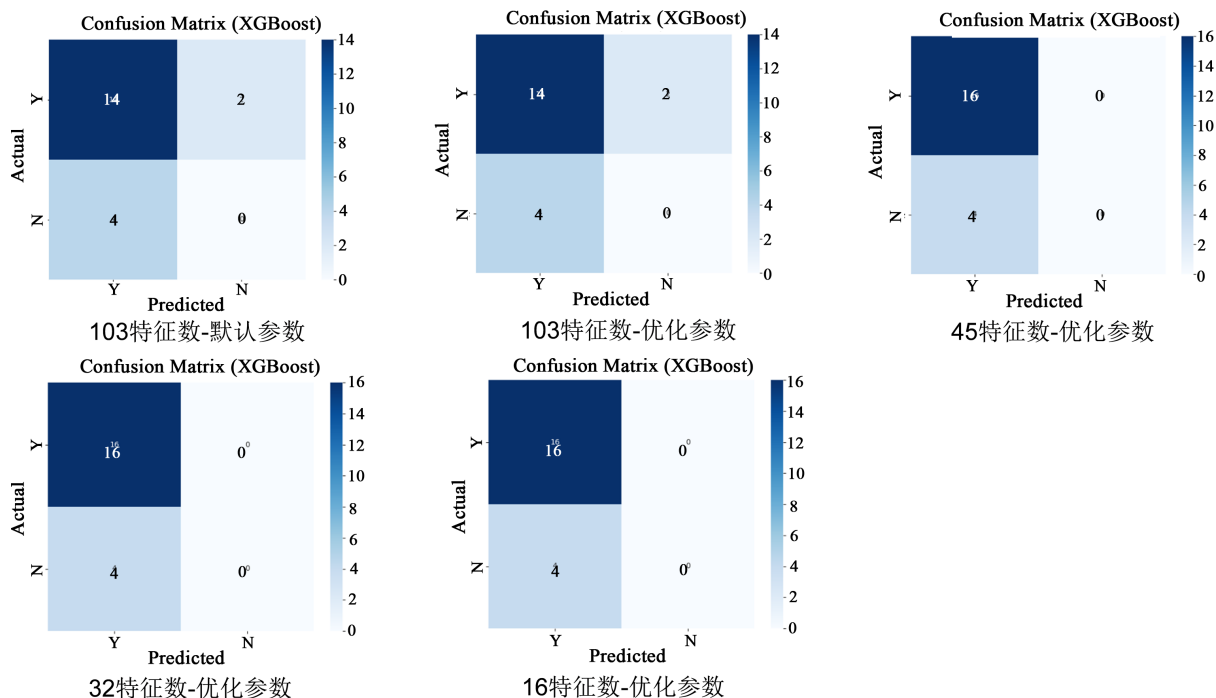


Figure 3. Confusion matrix for predictive modeling of hematoma expansion

图 3. 血肿扩张预测模型的混淆矩阵

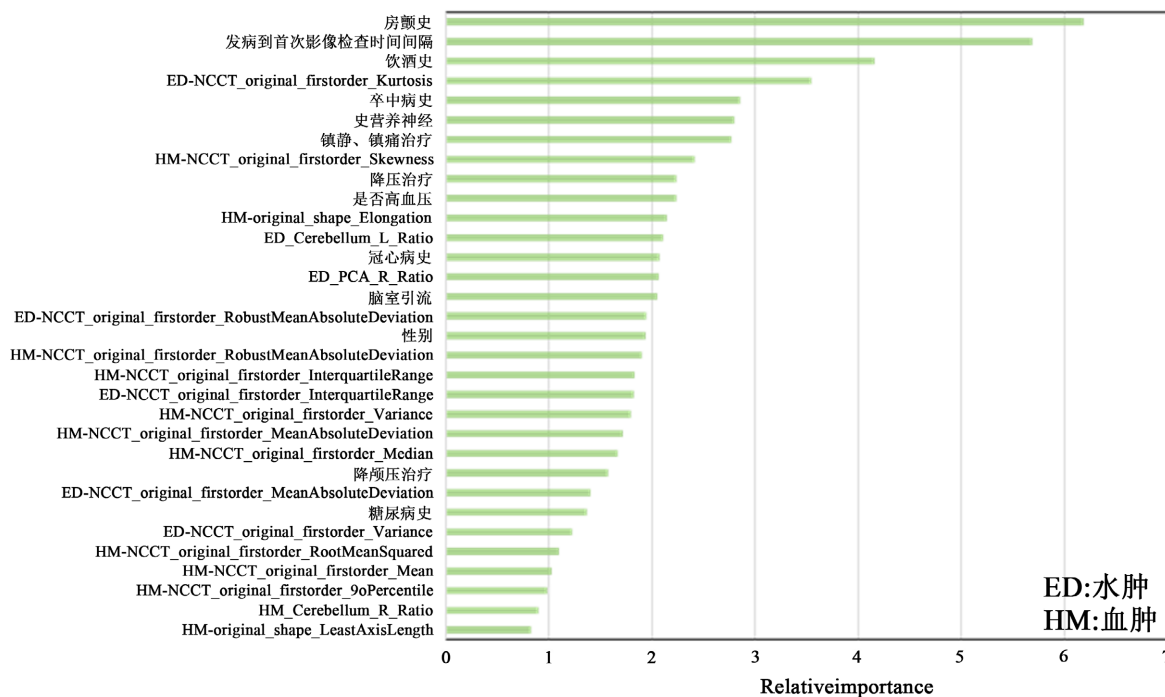


Figure 4. Ranking of features highly correlated with the occurrence of hematoma expansion

图 4. 与血肿扩张发生高度相关的特征排序

4. 结论

研究结果发现, 基于 XGBoost 算法建立的出血性脑卒中患者血肿扩张预测模型具备较好的预测性

能, 在特征向量数为 32 时预测模型的表现最佳, 使用测试集数据进行验证后, 得到模型的预测准确性为 80%。根据本研究的特征分析结果, 房颤史与是否发生血肿扩张事件具有很强的关联性, 其次为发病到首次影像时间的间隔、患者饮酒史等特征。此外, 由于本研究中纳入的患者数量仅为 130 名, 可能会对模型的预测准确性具有一定的影响, 在后续研究中增加患者的训练样本数量可有助于模型预测性能的进一步提升。

参考文献

- [1] Virani, S.S., Alonso, A., Benjamin, E.J., *et al.* (2020) Heart Disease and Stroke Statistics-2020 Update: A Report from the American Heart Association. *Circulation*, **141**, e139-e596.
- [2] Dowlathshahi, D., Demchuk, A.M., Flaherty, M.L., Ali, M., Lyden, P.L. and Smith, E.E. (2011) Defining Hematoma Expansion in Intracerebral Hemorrhage. *Neurology*, **76**, 1238-1244. <https://doi.org/10.1212/wnl.0b013e3182143317>
- [3] 陈亦豪. 基于人工智能技术的脑出血预后研究[D]: [博士学位论文]. 北京: 北京协和医学院, 2024.
- [4] Chan, S., Conell, C., Veerina, K.T., Rao, V.A. and Flint, A.C. (2015) Prediction of Intracerebral Haemorrhage Expansion with Clinical, Laboratory, Pharmacologic, and Noncontrast Radiographic Variables. *International Journal of Stroke*, **10**, 1057-1061. <https://doi.org/10.1111/ijis.12507>
- [5] Miyahara, M., Noda, R., Yamaguchi, S., Tamai, Y., Inoue, M., Okamoto, K., *et al.* (2018) New Prediction Score for Hematoma Expansion and Neurological Deterioration after Spontaneous Intracerebral Hemorrhage: A Hospital-Based Retrospective Cohort Study. *Journal of Stroke and Cerebrovascular Diseases*, **27**, 2543-2550. <https://doi.org/10.1016/j.jstrokecerebrovasdis.2018.05.018>
- [6] 袁伟, 侯文仲, 王倩, 等. 基于机器学习预测自发性脑出血血肿扩张研究[J]. 广东药科大学学报, 2022, 38(3): 76-84.
- [7] 刘雨安, 杨小文, 李乐之. 机器学习在疾病预测的应用研究进展[J]. 护理学报, 2021, 28(7): 30-34.
- [8] 翟晓风. 基于 XGBoost 算法的中国上市公司财务舞弊预测模型研究[D]: [硕士学位论文]. 北京: 中国财政科学研究院, 2022.
- [9] 齐巧娜, 刘艳, 陈霁晖, 等. 机器学习 XGBoost 算法在医学领域的应用研究进展[J]. 分子影像学杂志, 2021, 44(5): 856-862.
- [10] Torlay, L., Perrone-Bertolotti, M., Thomas, E. and Bacia, M. (2017) Machine Learning-XGBoost Analysis of Language Networks to Classify Patients with Epilepsy. *Brain Informatics*, **4**, 159-169. <https://doi.org/10.1007/s40708-017-0065-7>
- [11] Nishio, M., Nishizawa, M., Sugiyama, O., Kojima, R., Yakami, M., Kuroda, T., *et al.* (2018) Computer-Aided Diagnosis of Lung Nodule Using Gradient Tree Boosting and Bayesian Optimization. *PLOS ONE*, **13**, e0195875. <https://doi.org/10.1371/journal.pone.0195875>
- [12] Ye, C., Fu, T., Hao, S., Zhang, Y., Wang, O., Jin, B., *et al.* (2018) Prediction of Incident Hypertension within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. *Journal of Medical Internet Research*, **20**, e22. <https://doi.org/10.2196/jmir.9268>
- [13] Liu, L., Yu, Y., Fei, Z., Li, M., Wu, F., Li, H., *et al.* (2018) An Interpretable Boosting Model to Predict Side Effects of Analgesics for Osteoarthritis. *BMC Systems Biology*, **12**, Article No. 105. <https://doi.org/10.1186/s12918-018-0624-4>