http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

基于电子病历的心血管疾病合并认知障碍可解释识别模型

刘一鹏[1] 陈先来[2] 谭伟[2]

（[1]中南大学国家卓越工程师学院 长沙 410083，[2]中南大学计算机学院 长沙 410083）

〔摘要〕 目的/意义 构建基于多源电子病历的心血管疾病合并认知障碍（CVD-CI）可解释识别模型，评价多源电子病历数据治理和可解释机器学习在 CVD-CI 回顾性识别中的应用价值。方法/过程 回顾性收集长沙市某 3 家三甲医院 2021—2025 年间就诊记录 10 438 次，经数据治理和特征工程后，保留 8 414 次用于建模分析。用 XGBoost 等 9 种方法分别建立 CVD-CI 模型并进行比较，采用患者分组的 5 折交叉验证方法评价模型性能，结合 SHAP 方法对模型进行可解释性分析。结果/结论 XGBoost 在训练集中的折外 AUROC 为 0.889，在测试集中的 AUROC、AUPRC 和 Brier 评分分别为 0.868、0.667 和 0.103；测试集 CVD-CI 阳性记录占比为 19.9%。SHAP 分析显示，年龄、脑卒中史、呼吸频率、心肌肌钙蛋白 I 和脑室扩大等是模型识别 CVD-CI 共病状态的重要特征。结果表明，在本研究抽样设定下，规范治理后的多源电子病历数据可用于构建 CVD-CI 回顾性研究队列，并支持共病状态可解释识别模型的开发。
〔关键词〕 电子病历；心血管疾病；认知障碍；机器学习；可解释识别

An Explainable Model for Identifying Cardiovascular Disease With Comorbid Cognitive Impairment Based on Electronic Medical Records

LIU Yipeng[1], CHEN Xianlai[2], TAN Wei[2]

[1] National Graduate College for Elite Engineers, Central South University, Changsha 410083, China; [2] School of Computer Science and Engineering, Central South University, Changsha 410083, China

〔**Abstract**〕 **Purpose/Significance** To develop an explainable model for identifying comorbid cardiovascular disease and cognitive impairment (CVD-CI) using multisource electronic medical records (EMRs), and to evaluate the utility of EMR data governance and explainable machine learning in retrospective CVD-CI identification. **Methods/Process** A total of 10,438 medical encounter records from three tertiary hospitals in Changsha between 2021 and 2025 were retrospectively collected, of which 8,414 records were retained for model development and evaluation after data governance and feature engineering. Nine machine learning algorithms, including XGBoost, were developed and compared. Model performance was evaluated using patient-grouped five-fold cross-validation, and SHapley Additive exPlanations (SHAP) was employed to interpret model predictions. **Results/Conclusions** XGBoost achieved an out-of-fold area under the receiver operating characteristic curve (AUROC) of 0.889 in the training set. In the test set, the AUROC, area under the precision-recall curve (AUPRC), and Brier score were 0.868, 0.667, and 0.103, respectively, with a CVD-CI label-positive rate of 19.9%. SHAP analysis identified age, history of stroke, respiratory rate, cardiac troponin I, and ventricular enlargement as important features for distinguishing CVD-CI comorbidity. These findings suggest that, under the sampling scheme used in this study, appropriately governed multisource EMR data can support the construction of retrospective CVD-CI research cohorts and the development of explainable models for identifying comorbid disease status.
〔**Keywords**〕 electronic medical records; cardiovascular disease; cognitive impairment; machine learning; explainable identification

10.12201/bmr.202611.00063V1

1 引言

心血管疾病（cardiovascular disease，CVD）与认知障碍（cognitive impairment，CI）均是全球重要的公共卫生问题。全球疾病负担研究显示，2019 年全球 CVD 患者约 5.23 亿[1]；我国现有 CVD 患者约 3.30 亿[2]。此外，WHO 数据显示，2021 年全球约有 5700 万痴呆患者[3]；我国全国性调查估计，60 岁及以上人群中约有 1507 万痴呆患者和 3877 万轻度认知障碍患者[4]。随着人口老龄化进程加快，CVD 与 CI 所造成的疾病负担可能进一步叠加。值得注意的是，CVD 与 CI 并非相互独立的两类疾病，其发生发展具有共同的危险因素和病理生理基础。高血压、糖尿病、血脂异常、动脉粥样硬化、慢性炎症及年龄增长等因素既参与 CVD 的发生发展，也可能通过脑血管损伤、脑卒中、持续性脑血流改变和全身性血管病变影响认

http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

知功能。因此，CVD 患者是认知障碍筛查与早期识别的重点人群[5-9]。及时识别心血管疾病合并认知障碍（CVD-CI）患者，对开展疾病预防和改善预后具有重要意义[5-7]。

然而，CVD 与 CI 目前多由不同专科分别诊治，CVD 患者中的认知障碍容易被忽视。CI 早期症状较为隐匿，MMSE 和 MoCA 等量表虽可用于认知功能评价，但其实施受时间、患者配合程度、受教育水平及评估环境等因素影响，在常规诊疗中开展大范围筛查仍存在一定限制[10-12]。因此，如何利用常规诊疗过程中已经积累的数据辅助识别 CVD-CI 患者，是值得关注的问题。

电子病历（electronic medical record，EMR）包含诊断、实验室检验、生命体征、既往史及影像报告等多维临床信息，为 CVD-CI 的回顾性识别和模型构建提供了数据基础[12-14]。但 EMR 同时存在多源异构、术语不统一、非结构化文本和缺失特征等问题[12,14]。近年来，自然语言处理和大语言模型为非结构化病历信息抽取提供了新的技术路径[15-16]；机器学习方法也逐渐用于 EMR 中的疾病识别和临床预测[12-13,17-18]。其中，梯度提升树等方法适用于复杂非线性临床变量，SHAP 等解释方法可量化特征对模型输出的贡献[17-18]尽管如此，现有医学人工智能研究仍面临数据质量、模型可解释性和验证规范性不足等问题[11-12,19-20]。

基于此，本研究整合多源 EMR 数据，通过数据治理与特征工程构建 CVD-CI 回顾性研究队列，比较不同机器学习模型对既有 CVD-CI 共病状态的识别性能，并结合 SHAP 方法分析主要特征对模型输出的贡献。本研究旨在评估多源 EMR 数据支持 CVD-CI 回顾性病例识别的可行性，为相关真实世界研究提供方法参考。

2 资料与方法

2.1 数据来源

数据来源于长沙市 3 家三甲医院的电子病历系统，时间范围为 2021 年 1 月 2 日至 2025 年 3 月 8 日。提取人口学特征、诊断、入院记录、入院评估、首次病程、会诊记录、出院记录、实验室检验、辅助检查及三测单等数据。

2.2 纳入标准

结合临床指南、专家共识及项目医学专家意见，分别制定 CVD 与 CI 的操作性识别标准。其中，CVD 包括高血压、冠心病和心力衰竭；CI 的识别依据认知筛查量表记录及既往明确诊断，具体标准如下：

（1）高血压

参照《国家基层高血压防治管理指南 2025 版》，符合以下任意一项者判定为高血压[21]：

①诊室血压诊断标准：规范测量的诊室血压≥140/90 mmHg；

②家庭血压诊断标准：家庭血压平均值≥135/85 mmHg；

③动态血压诊断标准：24 小时动态血压平均值≥130/80 mmHg，或白天血压平均值≥135/85 mmHg，或夜间血压平均值≥120/70 mmHg；

④医疗记录中已有高血压诊断。

（2）冠心病

参照《冠心病 CT 检查和诊断中国专家共识》，符合以下任意一项者判定为冠心病[22]：

①冠脉造影、冠脉 CTA 报告结论明确符合预设冠状动脉病变标准；

②医疗记录中已有冠心病诊断。

（3）心力衰竭

参照《中国心力衰竭诊断和治疗指南 2024》并结合项目医学专家意见，心力衰竭纳入标准为医疗记录中存在明确的心力衰竭诊断[23]。

（4）认知障碍

参照《中国简短认知测试在痴呆诊断中的应用指南》，结合认知筛查量表记录与既往诊断，符合以下任意一项者判定为认知障碍[10]：

①痴呆诊断：MMSE＜24 分，提示存在明显认知功能损害；

②轻度认知障碍诊断：MoCA 18～25 分，或 MMSE 24～29 分，提示存在不同程度的认知功能减退；

③医疗记录中已有痴呆、轻度认知障碍诊断。

2.3 队列构建

10.12201/bmr.202611.00063V1

http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

同时满足 CVD 纳入标准和 CI 纳入标准的就诊记录定义为 CVD-CI 共病组；仅满足 CVD 或 CI 标准的就诊记录即为单病种组。同一患者的多次就诊按独立记录处理，但在数据集划分时以患者为分组单位，保证同一患者的全部记录仅进入一个数据分区。

鉴于不同单病种记录数量差异较大，从互斥单病种记录中按预设病种配额进行无放回随机抽样。

2.4 数据治理

EMR 治理包括术语标准化、单位统一、异常值处理、去重及时间字段核验。检验项目按名称、缩写和别名映射至标准名称，统一单位后按专家设定范围处理异常值；对同次住院重复测量构建首次值、末次值、均值、中位数、极值、标准差、极差和变化量等特征。就诊日期优先采用入院时间，无明确入院时间时采用首次病程记录时间作为参考。

2.5 特征工程与非结构化文本抽取

既往史等非结构化文本先由正则规则识别否定表达和术语，再由 DeepSeek-V3 提取疾病实体并编码为“肯定”“否定”或“未提及”[15-16,24]；随机抽取 100 例独立文本，以 Kappa、精确率、召回率和 F1 值评价抽取质量。

为避免直接标签信息进入模型，特征构建时未纳入痴呆、认知障碍等直接诊断字段，亦未纳入 MMSE、MoCA 评分及原始入院、出院诊断字段。用于标签赋值的疾病清单成员标记和结局变量仅用于队列划分及监督学习，不作为预测特征；患者标识、就诊编码及数据集划分日期亦不作为模型输入。最终模型纳入 216 项特征，包括人口学特征 6 项、病史与生活方式特征 12 项、实验室检验特征 153 项、生命体征特征 36 项及影像与脑电特征 9 项。标签定义信息、未纳入字段及完整特征清单见文末附录。

经核查，直接用于标签赋值的字段与最终模型输入字段不存在重叠。即模型保留了从既往史中提取的高血压史和冠心病史特征，而这两类信息与 CVD 标签定义中的部分疾病存在概念重叠。因此，字段层面的分离并不意味着模型输入与标签所反映的疾病信息在语义层面完全独立。

2.6 数据集划分与模型构建

本研究采用患者级时间留出法划分数据集。对每名患者，取其全部记录中最晚的就诊日期作为参考日期，使同一患者的全部记录始终位于同一数据分区，以避免患者重复信息跨分区造成数据泄漏[19-20]。设第 i 名患者共有 n_i 次就诊，第 j 次就诊日期为 t_{ij}，则其参考日期 T_i 为：

$$T_i = \max_{1 \le j \le n_i} (t_{ij}) \quad (1)$$

为兼顾模型开发样本量与较晚时间段独立测试所需样本量，研究设计阶段预先保留参考日期较晚的 20%患者作为测试对象，并以全部患者参考日期的第 80 百分位数确定时间切点：

$$T_{cut} = Q_{0.80}(T_1, T_2, \ldots, T_N) \quad (2)$$

本研究时间切点为 2023 年 7 月 8 日；参考日期早于该日的患者进入训练集，参考日期为该日及以后的患者进入测试集。

考虑到 EMR 数据具有变量数量多、缺失情况复杂及变量间可能存在非线性关系等特点，本研究选择 9 种具有不同建模特点的分类模型进行比较，包括 XGBoost、CatBoost、LightGBM、直方图梯度提升、随机森林、极端随机树、Transformer、弹性网络逻辑回归和高斯朴素贝叶斯。其中，前四种梯度提升树模型用于学习临床变量间的非线性关系和交互作用；随机森林与极端随机树作为袋装树集成模型；Transformer 用于探索注意力机制对变量关系的建模能力；弹性网络逻辑回归作为正则化的线性模型；高斯朴素贝叶斯作为低复杂度概率模型[13,17-18]。

训练集采用患者分组的分层 5 折交叉验证，同一患者全部记录位于同一折，并尽量保持各折 CVD-CI 比例接近；折外预测用于模型比较、概率校准和分类阈值确定。

所有数据依赖型预处理均在相应训练折内完成。梯度提升树模型保留原始缺失值，其余模型用训练折中位数填补；Transformer 和弹性网络逻辑回归进一步标准化，类别不平衡采用训练折内类别权重处理[14,19-20]。

各模型的关键参数详见表 1。

10.12201/bmr.202611.00063V1

http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

2.7 模型评价与可解释性分析

模型判别性能采用受试者工作特征曲线下面积（AUROC）和精确率-召回率曲线下面积（AUPRC）评价；分类性能指标包括 F1 值、阳性预测值（PPV）、阴性预测值（NPV）、敏感度和特异度[19-20]。

以训练集折外 AUROC 作为主要模型比较指标；折外预测用于拟合 Platt 校准器并确定 F1 最高的阈值。最终模型在全部训练集上重训后，将锁定的模型、校准器和阈值一次性应用于测试集。校准性能采用 Brier 评分、校准截距、校准斜率及校准曲线评价，并以决策曲线分析净获益[19-20]。

采用 SHAP 解释最终模型，以平均绝对 SHAP 值衡量单特征贡献，并以二阶 SHAP 交互值评价特征组合的联合贡献[17,25]。

2.8 统计学方法

采用 Python 3.13 完成数据整理、统计分析、模型构建及结果评价。连续变量根据数据分布特征采用均数±标准差或中位数（四分位数）进行描述，训练集与测试集间比较采用独立样本 t 检验或 Mann–Whitney U 检验；分类变量采用例数进行描述，组间比较采用 χ^2 检验，当理论频数不满足 χ^2 检验适用条件时采用 Fisher 确切概率法。所有统计检验均为双侧检验，以 $P<0.05$ 为差异有统计学意义。此外，采用标准化差异（standardized mean difference，SMD）描述训练集与测试集关键基线特征的分布差异。

模型由 XGBoost、LightGBM、CatBoost、scikit-learn 和 PyTorch 实现，SHAP 软件包用于模型解释。

表 1 模型的关键训练参数

模型	关键训练参数
XGBoost	树数量（n_estimators）=850；最大树深（max_depth）=3；学习率（learning_rate）=0.025；最小子节点权重（min_child_weight）=10；样本采样比例（subsample）=0.90；特征采样比例（colsample_bytree）=0.85；L1 正则化系数（reg_alpha）=0.20；L2 正则化系数（reg_lambda）=6.00
CatBoost	迭代次数（iterations）=700；树深（depth）=6；学习率（learning_rate）=0.035；L2 正则化系数（l2_leaf_reg）=5
LightGBM	树数量（n_estimators）=600；学习率（learning_rate）=0.035；叶节点数（num_leaves）=24；叶节点最小样本数（min_child_samples）=35；样本采样比例（subsample）=0.90；特征采样比例（colsample_bytree）=0.85；L2 正则化系数（reg_lambda）=3.00
直方图梯度提升	最大迭代次数（max_iter）=420；学习率（learning_rate）=0.045；最大叶节点数（max_leaf_nodes）=23；L2 正则化系数（l2_regularization）=2.50
随机森林	决策树数量（n_estimators）=600；叶节点最小样本数（min_samples_leaf）=3；每次分裂考虑的特征数（max_features）为总特征数的平方根
极端随机树	决策树数量（n_estimators）=650；叶节点最小样本数（min_samples_leaf）=3；每次分裂使用的特征比例（max_features）=0.70
Transformer	潜在标记数（latent_tokens）=8；嵌入维度（embedding_dim）=8；注意力头数（heads）=2；编码层数（layers）=1；丢弃率（dropout）=0.15；学习率=0.0007；批量大小（batch_size）=512；训练轮数（epochs）=15
弹性网络逻辑回归	正则化强度倒数（C）=0.35；L1 正则化比例（l1_ratio）=0.15
高斯朴素贝叶斯	方差平滑系数（var_smoothing）=1×10^{-8}

3 结果

3.1 研究对象及训练集与测试集特征

初始队列包括 10 438 次就诊记录（9 072 名患者），其中 CVD-CI 共病记录 2 438 次、按病种配额抽样的单病种记录 8 000 次。

经核验后，8 414 次具有可靠时间信息的记录进入主要分析，2 024 次因缺少可用于划分的数据日期未纳入。根据本研究采用的患者级时间留出设计，训练集 6 889 次就诊（5 899 名患者），测试集 1 525 次就诊（1 478 名患者），无患者交叉。此外，考虑到真实世界电子病历数据可能存在患者构成及数据记录模式的变化，本研究进一步采用 SMD 描述训练集与测试集关键特征的分布差异，而不以所有基线特征的组间比较均达到统计学不显著作为数据划分的目标。

训练集与测试集的研究对象特征见表 2。

表 2 研究对象及训练集与测试集的基线特征

10.12201/bmr.202611.00063V1

http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

特征	总体（n=8 414）	训练集（n=6 889）	测试集（n=1 525）	*P* 值	*SMD* 值
就诊记录（次）	8 414	6 889	1 525	—	—
患者（例）	7 377	5 899	1 478	—	—
CVD-CI 共病记录（次）	2 012	1 709	303	＜0.001	0.119
人口学特征					
年龄（岁）	61.8±18.4	61.2±19.2	65.1±13.4	＜0.001	0.235
男性（次）	4 750	4 053	697	＜0.001	0.265
主要诊断关键词构成					
高血压（次）	4 142	3 382	760	0.599	0.015
冠心病（次）	2 398	1 876	522	＜0.001	0.152
心力衰竭（次）	768	675	93	＜0.001	0.137
单纯 CI（次）	1 272	1 082	190	0.001	0.093
主要既往史文本抽取					
脑卒中史（次）	1 022	824	198	0.269	0.031
糖尿病史（次）	1 395	1 100	295	0.001	0.089
血脂异常史（次）	336	283	53	0.254	0.033
吸烟史（次）	1 656	1 343	313	0.360	0.026
部分关键数据缺失					
检验指标缺失（次）	1 693	1 153	540	＜0.001	0.435
既往史缺失（次）	1 951	1 573	378	0.102	0.046
生命体征特征缺失（次）	16	11	5	0.189	0.034
脑电特征缺失（次）	5 792	4 785	1 007	0.009	0.073
尿酸首次值缺失（次）	1 870	1 299	571	＜0.001	0.422

3.2 EMR 数据治理与文本抽取结果

经术语标准化、单位统一、异常值处理及质量核验后，形成规范化建模数据。

100 例文本复核中，大语言模型结合正则规则与人工标注的一致性 Kappa 为 0.758 6；高血压、糖尿病、冠心病、卒中史和痴呆实体抽取 F1 值分别为 0.865 4、0.634 9、0.717 0、0.883 7 和 0.888 9，具体抽取结果见表 3。

表 3 既往史文本疾病实体抽取结果

特征	真实实体数	正确提取	误报	漏报	精确率	召回率	F1 值
高血压	59	45	0	14	100.00%	76.27%	0.865 4
糖尿病	43	20	0	23	100.00%	46.51%	0.634 9
冠心病	34	19	0	15	100.00%	55.88%	0.717 0
卒中史	24	19	0	5	100.00%	79.17%	0.883 7
痴呆	15	12	0	3	100.00%	80.00%	0.888 9
痛风	2	2	0	0	100.00%	100.00%	1.000 0
血脂异常	2	1	0	1	100.00%	50.00%	0.666 7
慢阻肺	1	1	0	0	100.00%	100.00%	1.000 0
心力衰竭	1	1	0	0	100.00%	100.00%	1.000 0

研究发现糖尿病、冠心病存在较多的漏报情况，回溯核查发现，38 项漏报中，有 22 项（57.9%）在病历中的疾病表述过于口语化，模型未能准确提取；另 16 项（42.1%）虽表述明确，但因患者标识格式不一致，导致抽取结果关联失败、评估输出为空。

3.3 模型在训练集中的性能

患者分组 5 折交叉验证结果显示，各模型的折外预测性能存在一定差异，梯度提升类模型整体表现较优。XGBoost 的折外 AUROC 和 AUPRC 分别为 0.889 和 0.742，在 9 种候选模型中均最高，表现出较好的综合判别能力。CatBoost 的 AUROC 和 AUPRC 分别为 0.887 和 0.741，与 XGBoost 较为接近，表明不同梯度提升算法均能够较好地识别 CVD-CI 共病记录。

根据预先设定的主要模型比较指标，XGBoost 取得最高的折外 AUROC 点估计，同时 AUPRC 和 F1 值亦处于候选模型前列，且敏感度与特异度相对均衡，因此确定 XGBoost 为最终模型。

各模型的折外判别性能见图 1，校准与决策曲线见图 2，具体评价指标见表 4。

表 4 模型在训练集中的性能

10.12201/bmr.202611.00063V1

http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

模型	AUROC	AUPRC	F1 值	PPV	NPV	敏感度	特异度
XGBoost	0.889	0.742	0.687	0.679	0.898	0.694	0.892
CatBoost	0.887	0.741	0.679	0.622	0.911	0.747	0.851
LightGBM	0.884	0.727	0.678	0.630	0.907	0.735	0.858
直方图梯度提升	0.884	0.727	0.682	0.650	0.903	0.716	0.873
随机森林	0.849	0.675	0.635	0.574	0.896	0.710	0.826
极端随机树	0.846	0.664	0.631	0.560	0.899	0.722	0.813
Transformer	0.787	0.550	0.562	0.474	0.880	0.690	0.747
弹性网络逻辑回归	0.562	0.298	0.401	0.257	0.820	0.917	0.124
高斯朴素贝叶斯	0.520	0.255	0.405	0.257	0.858	0.957	0.085

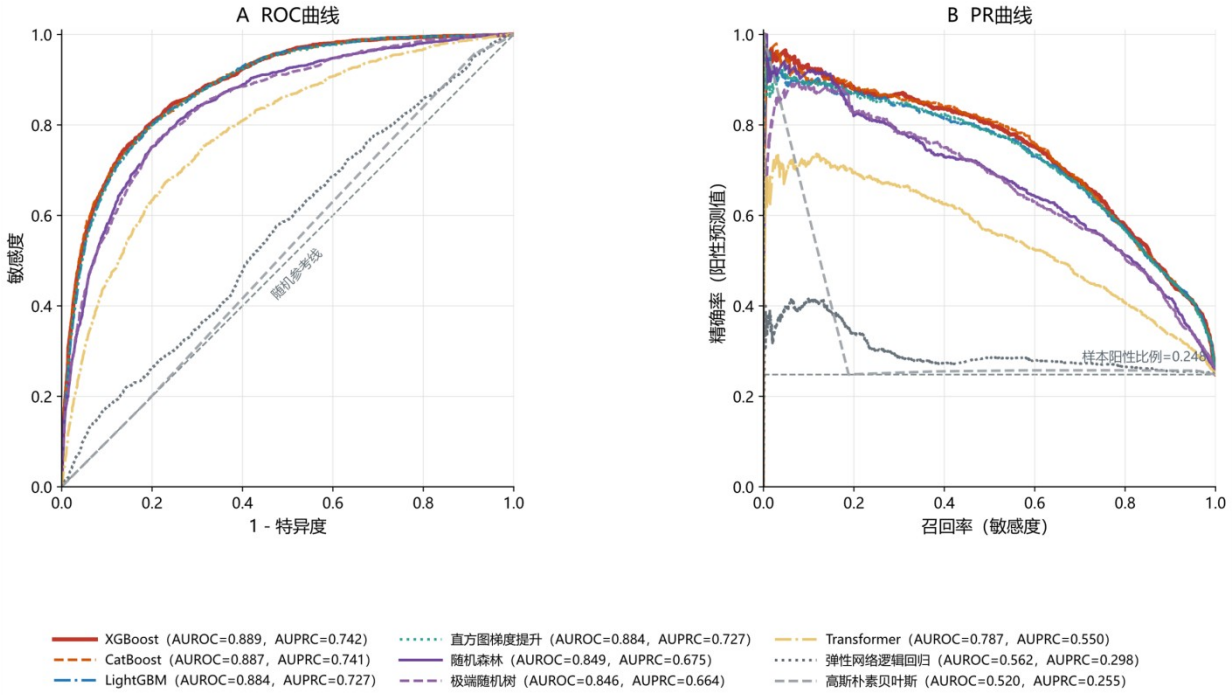

图 1 模型在训练集中的折外判别性能

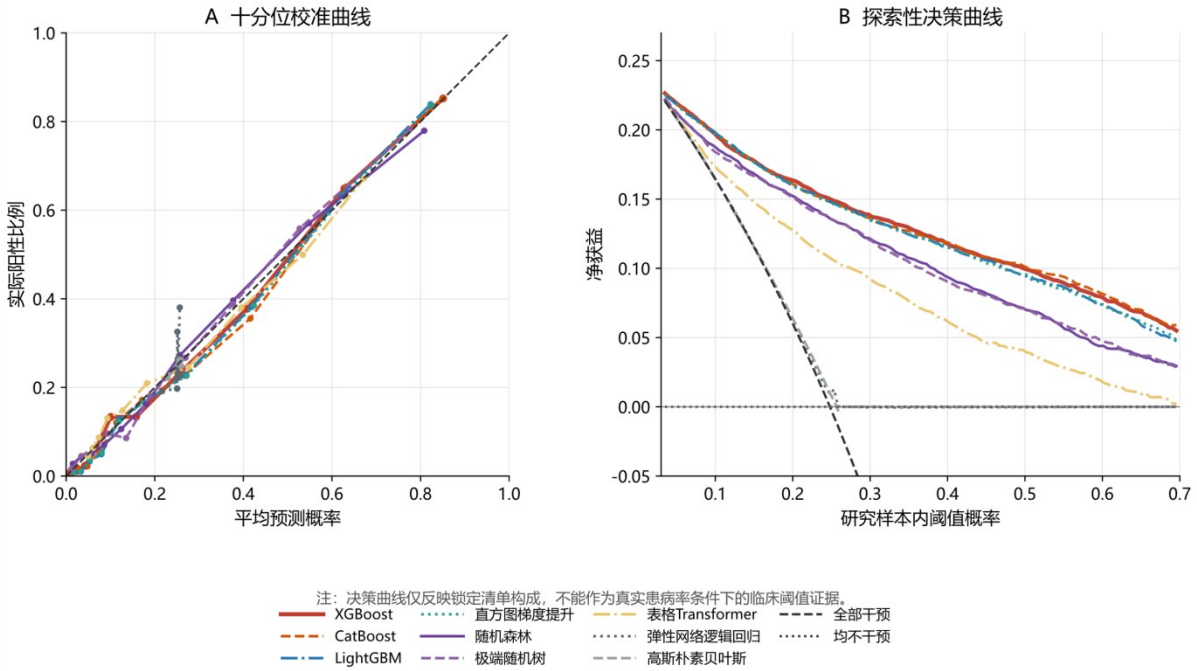

图 2 候选模型在训练集中的折外校准与决策表现

10.12201/bmr.202611.00063V1

http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

3.4 模型在测试集中的性能

测试集中，XGBoost 的 AUROC 为 0.868、AUPRC 为 0.667、F1 值为 0.620、Brier 评分为 0.103、校准斜率为 0.998，具有较好的判别与校准性能，其余模型性能详见表 5。

表 5 模型在测试集中的性能

模型	AUROC	AUPRC	F1 值	PPV	NPV	敏感度	特异度
XGBoost	0.868	0.667	0.620	0.613	0.907	0.627	0.902
CatBoost	0.857	0.646	0.602	0.602	0.896	0.601	0.876
LightGBM	0.845	0.615	0.620	0.569	0.913	0.681	0.843
直方图梯度提升	0.844	0.613	0.647	0.619	0.918	0.677	0.897
随机森林	0.807	0.505	0.519	0.510	0.882	0.528	0.874
极端随机树	0.827	0.531	0.544	0.506	0.893	0.587	0.858
Transformer	0.784	0.466	0.501	0.420	0.893	0.620	0.787
弹性网络逻辑回归	0.522	0.197	0.333	0.200	1.000	1.000	0.008
高斯朴素贝叶斯	0.516	0.204	0.338	0.204	0.932	0.987	0.045

3.5 SHAP 模型解释

SHAP 分析显示，患者年龄贡献最高，其次为脑卒中史、呼吸频率均值、心肌肌钙蛋白 I 末次值、脑室扩大和同型半胱氨酸等。较高年龄和脑卒中史阳性主要推动模型向 CVD-CI 共病组判定，详见图 3。

SHAP 交互分析中，入院年龄与呼吸频率均值的交互贡献最高，其次为年龄与尿酸标准差、年龄与脑卒中史等组合，详见图 4。

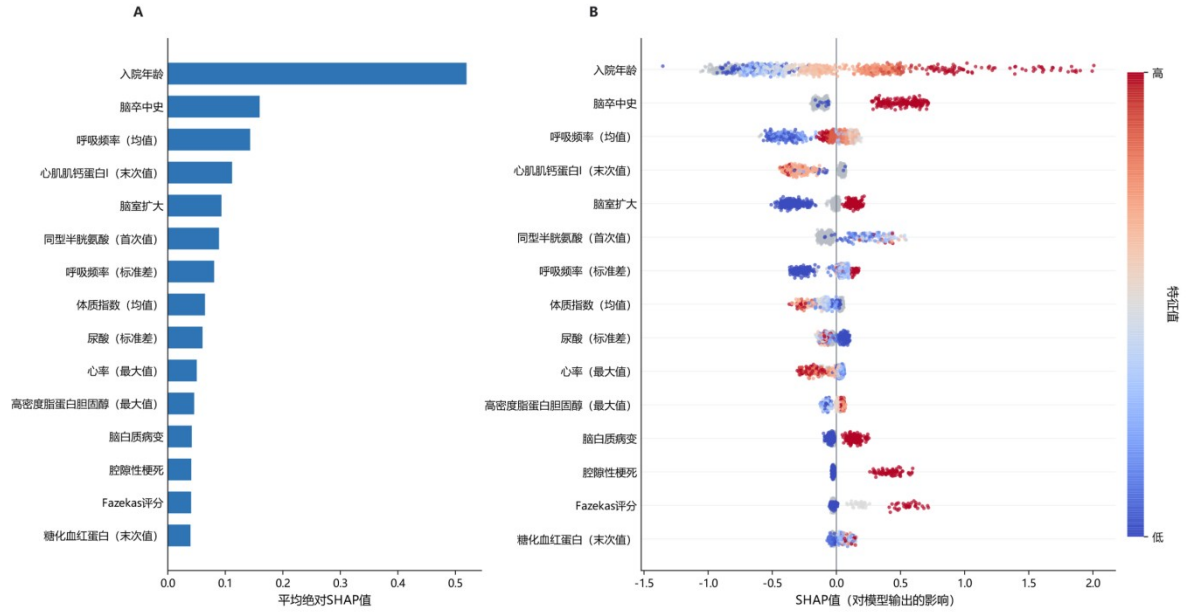

图 3 XGBoost 模型 SHAP 解释结果

10.12201/bmr.202611.00063V1

http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

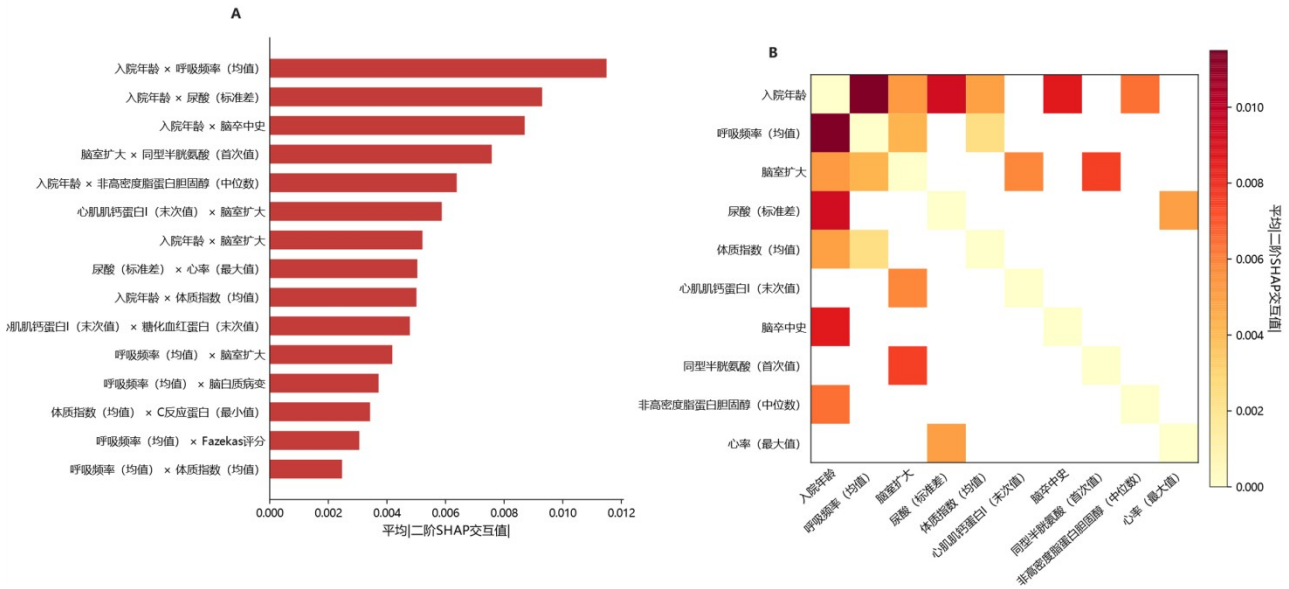

图 4 XGBoost 的二阶 SHAP 交互

4 讨论

4.1 主要研究发现

本研究整合 3 家医院的多源 EMR 数据，通过数据治理、非结构化文本抽取及特征工程构建 CVD-CI 回顾性识别模型。模型比较结果显示，梯度提升树模型整体表现较优，其中 XGBoost 在训练集中的折外 AUROC 为 0.889，在患者级时间留出测试集中的 AUROC 为 0.868，表明模型在不同时间分布的研究样本中保持了较好的判别能力。相比其他候选模型，XGBoost 对多维临床特征及其非线性关系具有一定的建模优势，适用于本研究中变量类型多样、缺失情况复杂的电子病历数据[13,17-18]。

既有研究已将机器学习用于心血管疾病患者的认知障碍识别。Wang 等基于冠心病患者的临床资料建立了轻度认知障碍识别模型[26]；Zhang 等结合英国生物样本库与中国医院数据开展模型开发和外部验证[27]。与上述研究相比，本研究侧重常规多源电子病历的数据治理与 CVD-CI 共病状态识别，整合了非结构化病史、实验室检验及辅助检查等信息，并采用患者级时间留出方式评价模型性能。

值得注意的是，模型性能不仅取决于所采用的算法，还受到病历数据质量和特征构建方式的影响。本研究通过术语标准化、单位统一及文本实体抽取整合不同来源的临床信息。既往史文本抽取与人工标注具有较好的一致性，但糖尿病和冠心病等实体的原始评价结果仍存在一定差异，提示非结构化信息的利用受到文本识别及数据关联质量的影响。因此，多源数据治理不仅是模型构建的前期处理环节，也是影响 EMR 识别结果的重要因素。

4.2 临床与信息学解释

SHAP 分析显示，年龄、脑卒中史和脑室扩大均为模型的重要判别特征。从临床角度看，三者均与脑血管损伤及脑结构改变相关。既往研究表明，年龄增长、血管危险因素和卒中后脑损伤与认知下降密切相关，脑结构改变也与血管性认知障碍的发生发展有关[5-9,28]。

除脑血管相关特征外，心肌肌钙蛋白 I 反映心肌损伤，既往研究提示高敏肌钙蛋白与卒中后认知功能变化相关[29-30]；同型半胱氨酸升高亦与卒中后认知障碍及轻度认知障碍向痴呆进展风险增加有关[29,31]。两者成为重要特征，提示心脑相互作用及血管代谢异常可能参与 CVD-CI 状态判别。

与上述特征相比，呼吸频率的贡献较具探索性。该指标可反映患者住院期间的循环和呼吸功能、急性生理应激，也是识别临床恶化的重要生命体征之一[32]。进一步的 SHAP 交互分析显示，年龄与呼吸频

10.12201/bmr.202611.00063V1

http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

率均值的交互贡献较高，提示模型在识别共病状态时不仅利用单项特征，还可能结合年龄与住院期间的生理状态进行综合判别。因此，其模型贡献可能既包含患者急性生理状态，也可能反映一定的临床监测与记录过程信息。

总体而言，年龄、脑卒中史、脑室扩大、心肌肌钙蛋白 I 和同型半胱氨酸共同指向脑血管损伤、心脑相互作用和血管代谢异常等 CVD-CI 相关病理过程；呼吸频率则提示，患者急性生理状态和诊疗过程信息同样参与了模型判别。

5 结语

本研究构建了基于多源 EMR 的 CVD-CI 可解释识别模型。结果表明，XGBoost 模型在测试集中保持了较好的判别与校准表现，经数据治理与特征工程后的 EMR 能够支持 CVD-CI 共病状态的病例识别。年龄、脑卒中史、脑室扩大及相关实验室和生命体征指标构成了 CVD-CI 重要判别特征。

本研究发现文本抽取性能评价不仅可能受到实体识别误差的影响，也可能受到疾病状态判定和数据关联错误的影响。若将关联失败的记录直接填充为阴性，可能导致疾病状态编码错误，并使抽取性能指标产生偏差。因此，后续需通过修正记录关联流程、开展独立人工标注及重新评价抽取性能，进一步明确文本处理误差的影响。

此外，本研究尚存在局限：一是数据来源集中于单一地区，缺乏独立外部验证；二是临床记录质量与数据缺失模式可能影响模型判别。现阶段，该模型仅适用于回顾性研究中的病例初筛与辅助识别，后续需在独立医疗机构开展外部验证，进一步优化病历信息抽取与缺失数据处理，并评估模型在真实临床工作流程中的应用价值。

作者贡献：刘一鹏负责文献调研、实验实施、论文撰写；谭伟负责指导与论文修订；陈先来负责提出选题、论文审核与修订。

利益声明：所有作者均声明不存在利益冲突。

10.12201/bmr.202611.00063V1

参考文献

[1] ROTH G A, MENSAH G A, JOHNSON C O, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study[J]. Journal of the American College of Cardiology, 2020, 76(25): 2982-3021.

[2] 刘明波，何新叶，杨晓红，等.《中国心血管健康与疾病报告 2023》概要（心血管疾病流行及介入诊疗状况）[J]. 中国介入心脏病学杂志，2024，32(10)：541-550.

[3] WORLD HEALTH ORGANIZATION. Dementia[EB/OL]. [2026-08-11]. https://www.who.int/news-room/fact-sheets/detail/dementia.

[4] JIA L, DU Y, CHU L, et al. Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: a cross-sectional study[J]. The Lancet Public Health, 2020, 5(12): e661-e671.

[5] LIVINGSTON G, HUNTLEY J, LIU K Y, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission[J]. The Lancet, 2024, 404(10452): 572-628.

[6] TESTAI F D, GORELICK P B, CHUANG P Y, et al. Cardiac contributions to brain health: a scientific statement from the American Heart Association[J]. Stroke, 2024, 55(12): e425-e438.

http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

[7] EL HUSSEINI N, KATZAN I L, ROST N S, et al. Cognitive impairment after ischemic and hemorrhagic stroke: a scientific statement from the American Heart Association/American Stroke Association[J]. Stroke, 2023, 54(6): e272-e291.
[8] SPEH A, KRAMBERGER M G, WINBLAD B, et al. Change in cardiovascular health and rate of cognitive decline in older adults: a 15-year population-based study[J]. BMC Geriatrics, 2024, 24(1): 263.
[9] SACHDEV P S, BENTVELZEN A C, GUSTAFSON D, et al. Vascular cognitive impairment and dementia: clinical features, neuropathology, and biomarkers[J]. Journal of the American College of Cardiology, 2026, 87(1): 52-76.
[10] 田金洲，解恒革，秦斌，等. 中国简短认知测试在痴呆诊断中的应用指南[J]. 中华医学杂志，2016，96(37)：2945-2959.
[11] 司唯，徐婷，林家玥，等. 人工智能技术在老年认知症筛查中的应用现状[J]. 医学信息学杂志，2025，46(2)：36-41，47.
[12] LU A, SRIKANTH V, ANDREW N, et al. Electronic health record-based prediction models for dementia detection: a systematic review of model performance and quality[J]. Journal of the American Medical Informatics Association, 2026, 33(6): 1210-1224.
[13] 王颖帅，王智飞，万艳丽，等. 基于电子病历的慢性病辅助诊断模型与特征边际效应分析[J]. 医学信息学杂志，2025，46(10)：17-24.
[14] 盖彦蓉，张云秋，张慧，等. 面向知识抽取的真实世界中文电子病历数据质量分析与治理对策研究[J]. 医学信息学杂志，2025，46(12)：47-53.
[15] 郭维嘉，郭少友. 中文电子病历数据元抽取方法[J]. 医学信息学杂志，2024，45(8)：78-83.
[16] 赵霞，赵尔康，程春雷，等. 大语言模型驱动的电子病历智能：应用、挑战与展望[J]. 医学信息学杂志，2026，47(1)：24-32.
[17] CATERSON J, LEWIN A, WILLIAMSON E. The application of explainable artificial intelligence (XAI) in electronic health record research: a scoping review[J]. Digital Health, 2024, 10: 20552076241272657.
[18] 田欣，陈红军，谢健. 基于 Transformer 与可解释人工智能的老年脓毒症预后预测模型：一项多中心研究[J]. 医学信息学杂志，2026，47(6)：73-80.
[19] COLLINS G S, MOONS K G M, DHIMAN P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods[J]. BMJ, 2024, 385: e078378.
[20] MOONS K G M, DAMEN J A A, KAUL T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods[J]. BMJ, 2025, 388: e082505.
[21] 国家心血管病中心国家基本公共卫生服务项目基层高血压管理办公室，国家基层高血压管理专家委员会. 国家基层高血压防治管理指南 2025 版[J]. 中华心血管病杂志，2025，53(9)：977-991.
[22] 中国医师协会放射医师分会. 冠心病 CT 检查和诊断中国专家共识[J]. 中华放射学杂志，2024，58(2)：135-149.
[23] 中华医学会心血管病学分会，中国医师协会心血管内科医师分会，中国医师协会心力衰竭专业委员会，等. 中国心力衰竭诊断和治疗指南 2024[J]. 中华心血管病杂志，2024，52(3)：235-275.
[24] DEEPSEEK-AI. DeepSeek-V3 technical report[EB/OL]. [2026-08-11]. https://arxiv.org/abs/2412.19437.
[25] LUNDBERG S M, ERION G, CHEN H, et al. From local explanations to global understanding with explainable AI for trees[J]. Nature Machine Intelligence, 2020, 2(1): 56-67.

10.12201/bmr.202611.00063V1

http://www.biomedrxiv.org.cn/article/doi/10.12201/bmr.202611.00063
此预印本（未经同行评审）的作者拥有该文稿的版权，biomedRxiv拥有永久保存权。任何人未经允许不得重复使用。

[26] WANG Q, LIU Y F, XU S H, et al. Development and validation of the eMCI-CHD tool: A multivariable prediction model for the risk of mild cognitive impairment in patients with coronary heart disease[J]. Journal of Evidence-Based Medicine, 2024, 17(3): 535-549.

[27] ZHANG J M, CHEN Z Q, ZHONG D L, et al. Interpretable machine learning model for classification of mild cognitive impairment in patients with coronary artery disease and hypertension: development and external validation using UK Biobank and Chinese hospital data[J]. BMC Medical Informatics and Decision Making, 2026. DOI: 10.1186/s12911-026-03741-5.

[28] MA Y, YANG Y, WANG X, et al. Prevalence and risk factors of poststroke cognitive impairment: a systematic review and meta-analysis[J]. Public Health Nursing, 2025, 42(2): 1047-1059.

[29] MA Y, CHEN Y, YANG T, et al. Blood biomarkers for post-stroke cognitive impairment: a systematic review and meta-analysis[J]. Journal of Stroke and Cerebrovascular Diseases, 2024, 33(8): 107632.

[30] VON RENNENBERG R, NOLTE C H, LIMAN T G, et al. High-sensitivity cardiac troponin T and cognitive function over 12 months after stroke-results of the DEMDAS study[J]. Journal of the American Heart Association, 2024, 13(6): e033439.

[31] ZULIANI G, BROMBO G, POLASTRI M, et al. High plasma homocysteine levels predict the progression from mild cognitive impairment to dementia[J]. Neurochemistry International, 2024, 177: 105763.

[32] CONSIDINE J, CASEY P, OMONAIYE O, et al. Importance of specific vital signs in nurses' recognition and response to deteriorating patients: a scoping review[J]. Journal of Clinical Nursing, 2024, 33(7): 2544-2561.

10.12201/bmr.202611.00063V1