基于两阶段自适应阈值集成学习算法的门诊患者及时取药预测模型研究Δ

范园园[1,2*]，王　丰[1]，曾攀科[1]，封卫毅[1#]（1.西安交通大学第一附属医院药学部，西安　710061；2.西北妇女儿童医院药剂科，西安　710061）

中图分类号　R952　　**文献标志码**　A　　**文章编号**　1001-0408（2025）24-3118-07
DOI　10.6039/j.issn.1001-0408.2025.24.18

摘　要　目的　构建门诊患者及时取药预测模型，精准识别延迟取药的高风险患者，为智慧药房差异化报到策略的制定与资源优化配置提供数据支撑。**方法**　基于西安交通大学第一附属医院2025年1—3月的680 568条门诊有效处方数据，先通过K均值聚类算法（K-means）与高斯混合模型（GMM）进行双聚类分析，结合轮廓系数择优确定取药时间差自适应阈值，以此划分“及时取药”与“延迟取药”，构建二元目标变量；通过多方法融合的策略筛选六大类特征；从区分度、整体性能与校准度3个维度对6种基学习器和4种集成学习模型进行性能评估，并开展模型解释性分析。**结果**　双聚类分析结果显示，GMM的轮廓系数优于K-means（0.702 4 vs. 0.698 8），最终确定的自适应阈值为49.82 min。纳入处方中，有74.99%的处方为及时取药，25.01%为延迟取药。10个候选模型中，堆叠集成（Stacking）模型表现最优，测试集曲线下面积为0.954 4、F1分数为0.942 4、准确率为0.911 5、Brier分数为0.066，区分度与校准度俱佳。模型解释性分析结果显示，模型的预测受患者历史行为、诊断相关特征等多因素共同驱动。**结论**　本研究构建了基于两阶段自适应阈值集成学习算法的门诊患者及时取药预测模型，其精准度与稳定性较高，可实现对患者取药行为的动态判定。

关键词　门诊处方；及时取药预测；聚类分析；机器学习；集成学习

Research on the timely medication retrieval prediction model for outpatients based on a two-stage adaptive threshold ensemble learning algorithm

FAN Yuanyuan[1,2], WANG Feng[1], ZENG Panke[1], FENG Weiyi[1]（1. Dept. of Pharmacy, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China; 2. Dept. of Pharmacy, Northwest Women's and Children's Hospital, Xi'an 710061, China）

ABSTRACT　OBJECTIVE To construct a predictive model for timely medication retrieval of outpatients, accurately identify high-risk patients with delayed medication retrieval, and provide data support for the development of differentiated registration strategies and resource optimization allocation in smart pharmacies. **METHODS** Based on 680 568 valid outpatient prescription records from January to March 2025 at the First Affiliated Hospital of Xi'an Jiaotong University, a dual-clustering analysis was conducted using K-means algorithm and Gaussian mixture model (GMM). An adaptive threshold for medication retrieval time difference was determined by combining contour coefficients, and "timely medication retrieval" and "delayed medication retrieval" were divided to construct binary objective variables; six types of features were screened through a multi-method fusion strategy; the performance of 6 kinds of base learners and 4 kinds of ensemble learning models were evaluated from three dimensions: discrimination, overall performance, and calibration, and explanatory analysis of the models were conducted. **RESULTS** The results of the dual-clustering analysis showed that the silhouette coefficient of GMM was better than K-means (0.702 4 vs. 0.698 8), and the final adaptive threshold was determined to be 49.82 min. Among the prescriptions included, 74.99% were for timely medication retrieval and 25.01% were for delayed medication retrieval. Among the 10 candidate models, the Stacking model performed the best, with an area under the test set curve of 0.954 4, F1 score of 0.942 4, accuracy of 0.911 5, Brier score of 0.066, and good discrimination and calibration. The explanatory analysis results of the model showed that its predictions were driven by multiple factors such as patient historical behavior, and diagnostic related characteristics. **CONCLUSION** This study constructed a timely medication retrieval prediction model for outpatients based on a two-stage adaptive threshold ensemble learning algorithm, which has

Δ**基金项目**　国家卫生健康委医院管理研究所医院药学高质量发展研究项目（No.NIHAYSZX2537）；西安交通大学第一附属医院基金项目（No.2024-RK-1）

＊**第一作者**　主管药师，硕士研究生。研究方向：医院药学。电话：029-85324174。E-mail：fanyy0029@stu.xjtu.edu.cn

\# **通信作者**　主任药师，博士生导师，博士。研究方向：肿瘤机制。电话：029-85323242。E-mail：fengweiyi@mail.xjtu.edu.cn

high accuracy and stability, and can achieve dynamic judgment of patient medication retrieval behavior.
KEYWORDS outpatient prescription; timely medication retrieval prediction; cluster analysis; machine learning; ensemble learning

在现代医院管理体系中，门诊药房是衔接医嘱执行与患者治疗的关键环节，其服务效率直接影响患者就医体验和医疗流程的连贯性。随着门诊量持续增长、处方结构日趋复杂以及患者取药行为差异扩大，门诊就医患者多、取药等候时间长，已成为长期普遍存在的问题[1]。医院普遍希望通过“报到”（即患者在取药前主动确认取药意愿并进入调剂队列）来辅助药房进行调剂管理，以减少资源浪费并提升流转效率。然而，在实际运行中，“是否需要报到”及“哪些患者应报到”成为门诊药房面临的突出难点[2—3]。

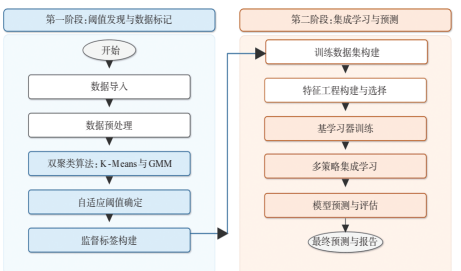
目前国内医院门诊药房主要存在两种取药模式：一是全报到模式，即所有患者均须在取药前进行报到，该模式能够减少部分患者因未及时取药造成的药品滞留和调剂资源浪费，但会增加患者重复排队和等待的时间成本，同时压缩药师可用于预调剂药品的时间窗口；二是不报到模式，即默认患者会及时前来取药，缴费后药师根据处方直接预调剂，无需患者确认，该模式尽管流程简化，但若患者未及时取药，就可能存在浪费药师调剂资源、增加待发药品管理成本、药品长期滞留药房的情况，这不仅造成资源闲置或浪费，还存在药品过期、冷藏药品室温暴露超时等安全隐患[4]。可见，如何精准识别具有延迟取药风险的患者，并对其实施“选择性报到”策略是在取药效率与患者就医体验之间取得平衡的关键。

近年来，机器学习技术在医疗行为预测方面展现出显著优势，其能够通过挖掘历史诊疗数据中的时序、交互与行为模式，对患者行为进行前瞻性预测[5—6]。然而，现有研究仍存在以下不足：（1）多数相关研究采用固定时间（如30、60 min）来界定是否“及时取药”[7—8]，忽略不同医院门诊流量、诊疗流程及取药模式的差异性，难以在不同医疗机构间泛化；（2）分析特征选择较单一，多集中于患者基础信息和药品属性，未充分考虑科室-诊断、科室-药品等跨维度交叉效应，限制了预测精度的提升。为弥补上述不足，本研究构建了基于两阶段自适应阈值集成学习算法的门诊患者及时取药预测模型，以精准识别延迟取药高风险患者，从而为取药患者差异化报到策略的制定与智慧药房资源优化配置提供数据支撑。

1 资料与方法

1.1 研究设计

本研究拟采用回顾性队列研究，基于西安交通大学第一附属医院（以下简称“本院”）信息系统的门诊处方与取药数据，构建基于两阶段自适应阈值集成学习算法的机器学习预测模型。该模型包括两个阶段（流程见图1）：第一阶段，阈值发现与数据标记；第二阶段，集成学习与预测。本研究已获得本院医学伦理委员会批准（伦理审批号XJTU1AF2025LSYY-401），并遵循《世界医学协会赫尔辛基宣言》。



K-means：K均值聚类算法；GMM：高斯混合模型。

图1 基于两阶段自适应阈值集成学习算法的门诊患者及时取药预测模型流程图

1.2 数据来源

收集本院2025年1—3月门诊患者的电子处方记录，纳入其中由门诊各科室开具且患者已完成缴费的处方，处方需同时包含以下4条关键信息：（1）患者信息，包括就诊科室、就诊号、处方号、患者姓名及主要诊断；（2）药品信息，包括药品名称及数量；（3）缴费信息，包括缴费方式与缴费时间；（4）取药信息，即发药时间。排除上述关键信息任一字段缺失的处方记录。

1.3 编程工具

本研究采用Python 3.9.12作为核心编程语言，依托相关科学计算库（NumPy、Pandas、Scikit-learn、Matplotlib）完成数据分析、模型构建与评估全过程，确保研究的可重复性与扩展性。中央处理器为Intel i9-13900K，图形处理器（graphics processing unit，GPU）为A6000 64 G，运行内存为64 GB，满足大规模数据集处理与复杂模型训练需求。

1.4 主要方法

1.4.1 数据预处理

计算所有纳入处方的取药时间差（time_diff），即从“缴费”到“发药”的时间（单位：min）。针对取药时间差数据的极端值与右偏态分布特征，采用对数变换[log1p（time_diff）]进行预处理，以提升聚类可分性。对取药时间差＞360 min（结合本院业务经验，超过360 min的取药行为多属于异常延迟）的数据进行极值上限缩尾法处理，以减少极端值对聚类结果的干扰。

1.4.2 基于双聚类算法的取药时间差自适应阈值确定

分别采用K-means和GMM对取药时间差进行双聚类分析，其中K-means通过识别距离聚类中心较近的簇的最大值来确定阈值，GMM通过均值较小的组件对应的聚类边界来确定阈值。为评价两种聚类方法的区分

能力，采用轮廓系数作为评价聚类效果的指标（取值范围－1～1，数值越接近1表示聚类越清晰），以轮廓系数最高的模型作为最优聚类方案。依据最优聚类方案确定最终的自适应阈值，从而将处方划分为“及时取药”与“延迟取药”两类。

1.4.3　二元目标变量的监督标签构建

将上述聚类得到的自适应阈值回代至所有纳入处方，进而明确定义二元目标变量。将取药时间差≤自适应阈值的处方定义为“及时取药”处方，记为1；将取药时间差＞自适应阈值的处方定义为“延迟取药”处方，记为0。该监督标签即为本模型的目标变量。

1.4.4　训练数据集构建

在完成数据预处理并生成监督标签后，得到包含所有特征变量与目标变量的数据集；将数据集按8∶2的比例进行分层抽样（目的是保持及时取药率在训练集和测试集中的一致性），划分为训练集和测试集，训练集用于模型训练和特征选择，测试集用于最终性能评估。

1.4.5　预测模型特征工程构建与选择

为构建高质量的模型输入变量集合，本研究采用多方法融合的特征选择策略，以确保最终特征子集在信息量、稳定性以及预测性能之间取得最佳平衡：首先，使用基于F统计量的SelectKBest特征算法筛选与目标变量关联性较强的特征，再结合递归特征消除与交叉验证法（recursive feature elimination with cross-validation，RFECV）逐步剔除冗余特征，根据树模型输出的特征重要性排序结果保留贡献度较高的特征，并通过5折交叉验证法评估不同特征子集的建模效果以确定最优特征组合。其次，在特征预处理方面，类别型特征根据具体模型需求采用目标编码或独热编码进行转换；数值型特征则进行标准化处理，以减少量纲差异对模型的影响，从而提高模型训练的稳定性与预测准确性。

基于数据内在属性与业务逻辑，将特征划分为六大类：（1）时间特征，包括缴费小时、缴费星期、缴费日、缴费月等原始时间特征，是否为工作时间、是否为周末等二分类特征，以及按时段划分的分类特征[上午（＞7∶00～12∶00）、下午（＞12∶00～18∶00）、晚上（＞18∶00～22∶00）、夜间（＞22∶00至次日7∶00）]。（2）患者行为特征，包括是否为本院职工、就诊次数等基础行为特征，患者历史处方数（指该患者过往处方记录的总行数，即所有处方明细条目的累计数量；当一张处方包含多种药品时，其记录行数会超过处方张数，可反映患者历史用药的复杂程度）、患者历史总处方数（指经过处方号去重后的处方张数，按就诊号聚合后对不同处方号计数，该特征不受包含多行记录的单张处方的影响）、患者历史及时取药次数、患者历史及时取药率等历史行为特征。（3）处方特征，包括处方药品种类数、处方总金额、同处方数量等处方特征，是否为复方药、药品频率（药品频率指药品名称出现的次数）、药品编码等药品特征，以及具体缴费方式等缴费特征。（4）科室特征，包括科室编码、科室名称（按业务属性分为内科、外科、妇儿科、专科等）等科室基础信息特征，以及科室及时取药率、科室处方数、科室工作量等科室行为统计特征。（5）诊断特征，包括诊断编码、诊断类型等分类特征，以及诊断类及时取药率、诊断类处方数等统计特征。（6）交叉特征，该类特征为捕捉多因素交互效应而设，采用目标编码构建组合特征，包括科室_诊断_及时取药率（即特定科室-诊断组合的历史及时取药率，以下类同）、科室_药品_及时取药率、时段_科室_及时取药率、科室_药品频率_及时取药率等。基于上述多方法融合的特征选择结果，对排名前20位的核心特征按重要性得分进行可视化呈现，直观识别影响患者及时取药行为的关键预测因子，为后续模型解释性分析与临床精准干预策略制定提供聚焦方向。

1.4.6　模型构建与训练

在确定的自适应阈值与监督标签基础上，本研究利用6种基学习器和4种集成学习来构建模型。其中，传统机器学习算法包括随机森林（Random Forest）、梯度提升决策树（Gradient Boosting Decision Tree，GBDT）与自适应提升（AdaBoost）；先进梯度提升算法包括采用GPU加速的极端梯度提升（XGBoost）、基于叶子节点优化的轻量级梯度提升（LightGBM）、专门针对类别型特征处理的类别提升（CatBoost）；多策略集成学习包括加权投票（Weighted Voting）、堆叠集成（Stacking）、混合集成（Blending）与简单平均集成（Simple Averaging）。

为提升模型性能，采用智能框架Optuna执行贝叶斯超参数优化：采用TPE（一种基于树结构的贝叶斯优化算法）采样器，配合Median Pruner剪枝器自动终止表现不佳的试验，每个模型均进行100次优化试验，并采用3折交叉验证超参数组合的泛化性能，优化目标为曲线下面积（area under the curve，AUC）。

1.4.7　模型性能评估

本研究从区分度、整体性能与校准度3个维度，对候选模型及其集成模型开展系统评价，旨在全面衡量模型预测能力的稳健性与可靠性：（1）区分度方面，采用受试者工作特征（receiver operating characteristic，ROC）的AUC作为核心指标，量化模型在不同决策阈值下对“及时取药”与“延迟取药”两类样本的区分效能。（2）整体性能方面，针对类别不平衡场景，采用“准确率、精确率、召回率、F1分数”的组合评价方法。其中，准确率反映模型整体预测正确率；精确率聚焦“预测为某类的样本中实际为该类的比例”，用于衡量预测结果的准确性；召回率聚焦“实际为某类的样本中被正确预测的比例”，用于衡量模型对目标类别的覆盖能力；F1分数则通过调和平均精确率与召回率用于综合平衡二者的表现，避免单一指标导致的评价偏差。（3）校准度方面，采用Brier分数量化预测概率与实际发生概率的偏差程度，同时借助校准曲线对模型在各概率区间的拟合一致性进行可视化。所

有指标均使用Python的Scikit-learn工具包进行计算，并在测试集上完成，确保评价结果的可靠性与可复现性。

1.4.8　模型解释性分析

在模型性能评估的基础上，本研究进一步开展模型解释性分析，以揭示预测结果的关键预测因子并验证模型决策逻辑的合理性。解释性分析主要包括两部分内容：模型整体区分能力的可视化呈现以及特征贡献度的量化解释。首先，通过混淆矩阵展示最佳模型在测试集上的分类结果，明确真阳性（true positive，TP）、真阴性（true negative，TN）、假阳性（false positive，FP）、假阴性（false negative，FN）的数量及比例，用于识别模型分类误差的主要来源。同时，绘制多模型ROC曲线，以比较其区分效能的差异，借助模型校准曲线的可视化分析，验证模型决策结果的可靠性。其次，采用基于博弈论原理的沙普利加性解释（Shapley additive explanations，SHAP）方法，分解各特征对预测结果的贡献度，既提供单样本的局部解释（如“历史及时取药率高”对“及时取药”预测的正向贡献），又通过汇总全局SHAP值揭示特征的正负向作用及强度。特征重要性摘要图（反映特征重要性排序及特征值与预测结果的正负相关性）以条形图形式展示特征贡献优先级，以突出关键预测因子。

2　结果

2.1　纳入数据基本情况

本研究纳入本院2025年1—3月门诊患者的电子处方数据680 568条，经数据清洗和预处理后，最终分析数据集包含680 568条有效记录，满足统计分析要求。

2.2　基于双聚类算法的取药时间差自适应阈值结果

K-means聚类分析结果显示，簇中心分别为9.26、175.53 min，代表两个簇内样本取药时间差的集中趋势（图2A）。GMM聚类分析结果显示，两个高斯分布成分可有效拟合样本数据的分布特征，二者的概率密度函数分别对应“及时取药”（取药时间差小）、“延迟取药”（取药时间差呈长尾分布）的特征（图2B）。聚类效果对比结果显示，GMM聚类的轮廓系数（0.702 4）优于K-means（0.698 8），因此，选择GMM聚类确定取药时间差的自适应阈值。对纳入处方的患者取药时间差进行数据分布分析，结果（图3）显示，数据呈显著右偏态分布，中位数（13.08 min）远低于平均值（69.36 min），标准差为117.96 min，提示大部分患者很快取药，但少数患者极晚取药把分布拖出长尾，因此均值被少数极端值抬高，均值不能代表“典型”行为。由GMM聚类分析结果可知，模型识别出的自适应阈值为49.82 min、中位数为13.08 min（图3），该阈值代表了两个核心用户群体的自然分界点，轮廓系数为0.702 4，表明聚类结果具有较高的内部一致性和簇间分离度。

2.3　二元目标变量的监督标签构建结果

将自适应阈值49.82 min回代至所有纳入处方后，有510 380条（74.99%）处方的取药行为属于“及时取药”

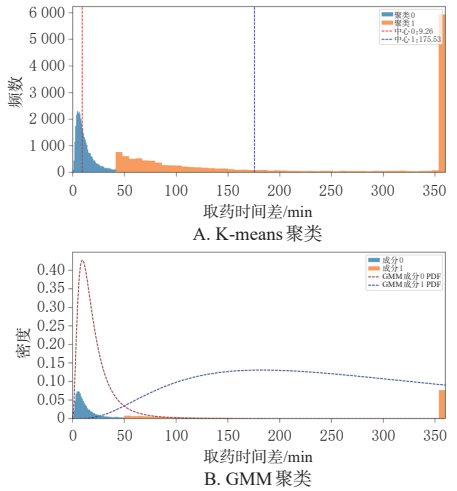


图2　双聚类分析结果

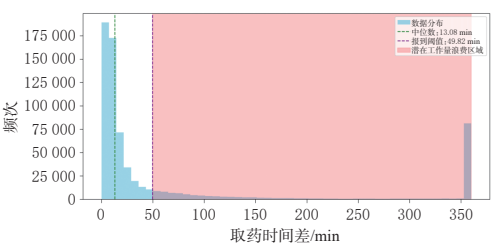


图3　取药时间差的数据分布图

（取药时间差≤49.82 min），记为1；有170 188条（25.01%）处方的取药行为属于“延迟取药”（取药时间差＞49.82 min），记为0。

2.4　训练数据集构建结果

将数据集分为训练集（544 454条）和测试集（136 114条），确保两组数据在目标变量（及时取药/延迟取药）的分布上保持一致。

2.5　特征工程选择结果

本研究通过特征工程生成28个初始特征，并挖掘特征交互关系构建交叉特征后生成35个衍生特征，最终得到63个特征。采用RFECV进行特征筛选，结果（图4）表明，重要性得分排序前5位的特征依次为患者历史及时取药率（0.258 2分）、科室_诊断_及时取药率（0.094 4分）、患者历史及时取药次数（0.085 7分）、科室_药品_及时取药率（0.081 4分）、科室_药品频率_及时取药率（0.062 1分）。可见，患者历史行为特征占预测主导地位，重要性得分排名前3位的特征有2个均与此相关；4个交叉特征（科室-诊断组合、科室-药品组合、科室-药品频率组合、时段-科室组合）进入排名前5位，充分体现了医疗场景中多维度因素交互作用的复杂性及其关键作用；缴费日、缴费小时等时间特征虽未占据核心位置，但仍展现出一定的预测价值；此外，多个科室特征（科室及时取药率等）表现突出，也反映出不同医疗服务差异所带来的重要影响。

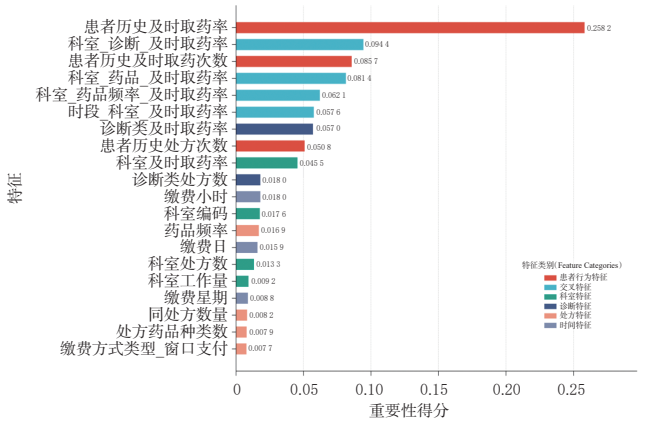


图4　门诊患者处方及时取药预测特征重要性得分可视化图（重要性得分排名前20位）

2.6　模型性能评估结果

在独立测试集上，基于Optuna进行超参数优化的6种基学习器及4种集成学习算法均获得了较好的预测效果（表1），其中Stacking模型表现最佳，在区分度与校准度方面均取得最优结果，其AUC为0.954 4，准确率为0.911 5，F1分数为0.942 4，明显优于其他模型。在预测概率的校准度方面，Stacking模型的Brier分数为0.066，是所有模型中最低的，表明其概率输出最为稳定、可信。综上，Stacking模型在本研究任务中展现出最强的预测性能，可作为门诊处方及时取药预测的首选模型。

表1　10种模型在测试集上的性能评估结果

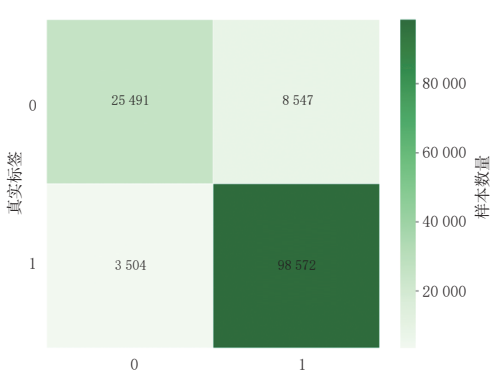
模型名称	准确率	精确率	召回率	F1分数	AUC	Brier分数
Stacking	0.911 5	0.920 2	0.965 7	0.942 4	0.954 4	0.066
Random Forest	0.900 9	0.904 4	0.970 4	0.936 3	0.947 9	0.073
Blending	0.909 6	0.918 7	0.964 8	0.941 2	0.953 7	0.068
Weighted Voting	0.906 1	0.911 4	0.968 9	0.939 3	0.952 1	0.074
Simple Averaging	0.906 0	0.911 4	0.968 9	0.939 3	0.952 1	0.075
GBDT	0.896 1	0.904 4	0.963 2	0.932 9	0.943 4	0.076
XGBoost	0.905 9	0.913 7	0.965 7	0.939 0	0.951 5	0.069
LightGBM	0.903 6	0.911 9	0.964 6	0.937 5	0.949 7	0.070
CatBoost	0.909 1	0.916 5	0.967 0	0.941 0	0.952 8	0.067
AdaBoost	0.884 7	0.896 5	0.956 7	0.925 6	0.931 2	0.224

2.7　模型解释性评估结果

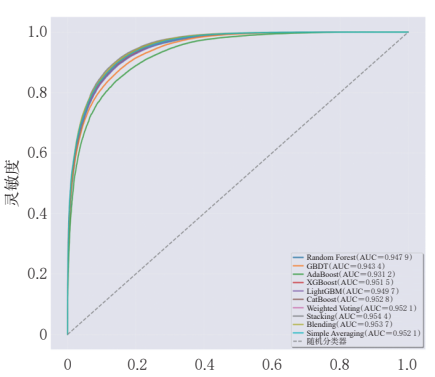
通过构建Stacking模型在测试集（136 114条）上的混淆矩阵，结果显示，TN为25 491条，FP为8 547条，FN为3 504条，TP为98 572条，最终计算出灵敏度为96.57%，特异度为74.89%，表明该模型具有较高的可靠性。结果见图5A。

模型ROC曲线比较结果显示，Stacking模型的ROC曲线最接近左上角，AUC达到0.954 4，其他集成模型的ROC曲线也表现出一定优异的区分能力，模型之间的性能差异在ROC曲线上清晰可见。结果见图5B。

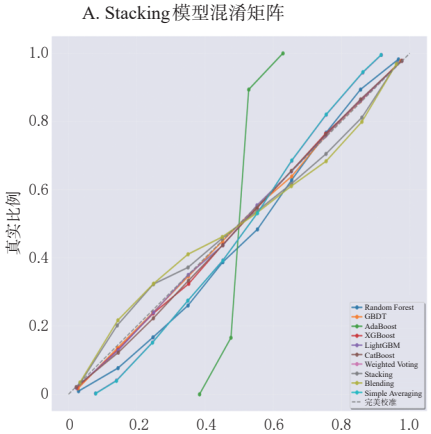
模型校准曲线结果显示，Stacking模型概率均值为0.749 5，方差为0.121 9，分布两端更分离，信息量更大，与最佳区分度一致，表明模型整体校准良好，预测概率与实际发生概率最接近对角线。结果见图5C。



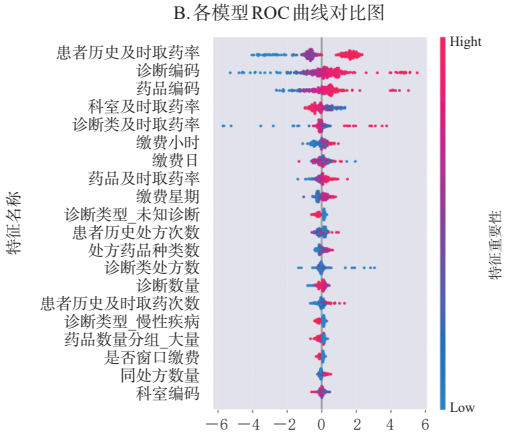
A. Stacking模型混淆矩阵



B. 各模型ROC曲线对比图



C. 各模型校准曲线



D. Stacking模型的SHAP特征重要性摘要图

注：D图中，红色代表高特征值，蓝色代表低特征值，SHAP值＞0表示正向作用，SHAP值＜0表示负向作用。

图5　模型解释性评估结果

Stacking模型SHAP特征重要性摘要图（图5D）结果显示：（1）整体而言，模型的预测主要受患者历史行为、诊断相关特征以及时间等多因素共同驱动，但历史行为变量占据主导地位，“患者历史及时取药率”贡献度最高，数值高时更易预测为“及时取药”、数值低时则更易预测为“延迟取药”，印证历史取药习惯是核心预测依据，“患者历史及时取药次数”“患者历史处方次数”也能说明历史数据的预测价值；（2）各类诊断特征（如“诊断编码”“诊断及时取药率”“诊断类型及时取药率”等）整体重要性位居前列，当诊断维度的及时取药率较高时，提示与特定疾病类别相关的既往行为模式会影响患者本次的取药决策；（3）“药品编码”“药品及时取药率”等药品相关特征同样具有较高贡献，尤其是药品及时取药率，数值高时说明患者取药行为与特定的药品类别存在关联；（4）“缴费小时”“处方药品种类数”等特征的重要性相对居中，但也有一定规律（比如部分时段缴费的患者，取药可能更易延迟）。综上，模型主要依据患者过去的取药行为进行判断，同时也会考虑就诊科室环境和就诊时间等因素，这与实际临床情况相符，表明模型可信度良好。

3 讨论

本研究构建的两阶段自适应阈值集成学习算法模型在门诊患者及时取药预测任务中表现出显著优势，为门诊药房患者报到的智能化管理提供了科学依据。通过预测患者是否会及时取药，可实现差异化的药房服务策略：对预测为“延迟取药”的患者要求报到后再进行预摆药，从而减少调剂资源的浪费；对预测为“及时取药”的患者，可启动提前预摆药与优先调剂机制，进一步缩短患者取药等候时长，提升门诊药房的服务效率与患者就医体验。

3.1 算法优势

本研究提出两阶段自适应阈值集成学习算法，有效克服了传统机器学习模型在医疗非平衡数据处理中常见的阈值固定化的局限。该算法通过“双聚类-阈值自适应”两阶段设计，在第一阶段通过双聚类分析构建客观的行为分界点，确立自适应阈值；在第二阶段，则引入自适应阈值调整机制动态优化分类边界，使模型能更精准地捕捉患者门诊取药的行为特征。相较于“先平衡后分类”理念，本研究进一步通过动态阈值强化了对个体行为差异的适应性[9]。樊丽娟等[10]采用Random Forest模型预测处方合理性虽获得0.90的特异度，但因固定阈值设置导致敏感度受到限制。本研究基于680 568条三甲医院真实处方数据，采用Stacking模型实现了高精度预测（AUC=0.954 4，F1分数=0.942 4），并界定了49.82 min的取药时间差自适应阈值，为智慧药学服务提供了更加可靠的行为判定依据。这一算法创新性将集成学习由传统的静态风险预测扩展至动态行为决策，通过对实时临床环境与患者行为变化的敏感响应，使得对患者取药行为的判定更贴合医院运营实际，避免了固定阈值带来的临床误判与资源错配。

3.2 双聚类对比分析的优势

本研究基于K-means与GMM双聚类对比，采用轮廓系数进行客观择优（GMM聚类和K-means聚类的轮廓系数分别为0.702 4、0.698 8），最终确定49.82 min为及时/延迟取药的自然分界点。相较于既往研究中常用的30、60 min的人工设定阈值，双聚类方法充分考虑了门诊不同时段的患者量波动、科室及药品类型差异，可有效减少固定阈值带来的错误分类[8]。GMM聚类通过概率分布拟合人群的内在行为结构确定阈值，形成74.99%（510 380条）的及时取药率与25.01%（170 188条）的延迟取药率分布，与三甲医院门诊运营特性高度一致[11]，为目标变量构建提供了科学、客观的依据。这一阈值确定方式让及时/延迟取药的界定更具临床适配性，避免了“一刀切”标准导致的患者体验下降或资源浪费，为后续差异化策略落地奠定了客观基础。

3.3 集成学习的优势

研究结果显示，Stacking模型在所有关键指标上均表现最佳（准确率为0.911 5，F1分数为0.942 4，AUC为0.954 4），优于单一机器学习模型（如XGBoost的AUC为0.951 5）及其他集成学习模型（如Blending的AUC为0.953 7）。Stacking模型的两层结构通过融合多种基学习器（如Random Forest的稳健性、XGBoost的非线性拟合能力等）有效降低了模型的偏差与方差[12—13]，进而提高了对不同类型患者（特别是首次就诊者）的预测稳定性；Optuna又基于TPE采样与Median Pruner剪枝器的超参数优化进一步提升模型泛化能力，使Stacking模型表现出更优的概率校准性（Brier分数为0.066）。本研究中，延迟取药率为25.01%，意味着传统预摆药模式中约1/4的调剂工作可能无效。利用预测模型实施智能化报到策略，可减少无效调剂、人力浪费及药品反复入库，改善高峰期药房负荷，同时以高召回率精准识别高风险患者，在效率与患者体验之间实现优化平衡[14]。该集成学习框架突破了单一模型在复杂医疗场景中易受噪声与类别不平衡影响的性能瓶颈，且实现了预测模型与门诊药房实际流程的无缝连接，其高精准度与高稳定性确保了技术方案在临床真实运行环境中的可落地性，直接助力药房运营效率提升与管理成本降低。

3.4 特征体系的优势

本研究构建了涵盖患者时间、行为、处方、科室、诊断以及交叉特征的多维特征体系，又与SHAP分析深度联动，不仅精准揭示了门诊取药行为的多维度驱动机制，更为药事管理提供了可直接落地的精细化干预靶点。从特征重要性来看，患者历史行为特征占绝对主导地位，这与药学领域“患者依从性惯性理论”高度契合[15]，既往取药及时的患者更易形成“缴费后立即取药”的行为习惯，而反复延迟的患者多存在行为偏差（如依

托职工健康管理中心集中取药）或流程特殊性（如部分检查预约当天统一取药），对于这类患者需要针对性地提供引导服务。此外，相关研究证实，患者历史处方数对及时取药率存在显著正向影响，处方数较多的患者延迟风险更低，这为医院设计“患者专属引导服务”（如长期用药患者的取药路径简化）提供了数据支撑[16]。值得注意的是，有4个交叉特征进入了特征重要性得分前10名（如科室_诊断_及时取药率的特征重要性得分为0.094 4，科室_药品_及时取药率的特征重要性得分为0.081 4），有效填补了既往研究仅关注单一维度、忽略多因素交互效应的空白，为按场景定制取药患者报到策略奠定了基础。

基于上述特征机制揭示的关键差异，医疗机构可进一步制定精准的分层药事服务优化策略，并通过系统化落地实现研究结果向临床实践转化。在科室层面，可依据不同科室就诊患者延迟取药率的显著差异实施差异化管理：对高风险科室（如医学影像科门诊、麻醉科门诊），建议对所有就诊患者实施强制报到策略，避免药品调剂后因患者延迟取药导致滞留；对中风险科室，可依托模型实时预测结果实施选择性报到，在控制风险的同时减少不必要的流程干预；对低风险科室，则可维持传统预摆药模式，缩短患者等候时间，保障取药效率。在药品管理层面，可针对不同类型药品的特性优化流程：造影剂（碘克沙醇、碘佛醇等）和检查用药是延迟取药的高发品类，这类药品因使用时间与检查预约绑定，患者对取药时间的控制能力弱且药品成本较高，建议实施强制报到策略，即患者需在检查前通过报到机确认后，药师再启动调剂流程；而慢性病常用药物（如治疗冠心病的阿司匹林、治疗睡眠障碍的氯硝西泮等）因及时取药率为89%～92%，可维持预摆药模式，但需结合患者历史及时取药率等行为特征进行个性化调整。在实施路径上，可将预测模型无缝集成至报到机系统或处方分配系统，构建“患者缴费时自动触发实时预测-依据预测结果智能分流至对应取药流程-结合实际取药行为持续优化模型参数”的闭环管理模式。同时，同步优化报到机工作流程：高风险患者遵循“缴费→报到→调剂→取药”流程，低风险患者可实行“缴费→调剂→取药”简化流程，中风险患者则根据实时预测结果动态分配，最终实现药事服务资源的精准配置与患者取药体验的双重提升[17]。

本研究也存在一定局限性：本研究数据来源于单中心三甲医院，虽然具有一定代表性，但仍可能存在地域、患者结构和科室设置等方面的偏倚，未来需开展多中心验证研究，以进一步评估模型的普适性和泛化能力。此外，本研究未纳入患者主观因素（如取药意愿、时间敏感性等），这些因素可能对取药及时性产生重要影响，后续可结合患者问卷纳入患者行为与心理特征，以提升模型的完整性和解释力。

参考文献

[1] 王哲. 基于闭环管理改造模式下的门诊药房管理系统对提高患者取药高峰时间段发药质量与缩短等候时间的影响[J]. 抗感染药学，2021，18(5)：752-755.

[2] 冀杨，张丁元，王晨，等. 北京市属医院门诊窗口服务现状、问题及对策研究[J]. 中国医院，2025，29(8)：6-8.

[3] 白晶. 智慧门诊建设流程优化之智能药房二次报到的效果分析[J]. 科技资讯，2020，18(19)：36-38.

[4] 叶微微，尉雯雯，李桂祥. 药房自动化系统中门诊流程的优化和应用[J]. 中国数字医学，2017，12(12)：85-87.

[5] 吕建祎，王纯熙，刘思成，等. 人工智能在生物医学研究中的应用[J]. 中国比较医学杂志，2025，35(7)：169-176.

[6] 杨晓雯. 人工智能视域下的医护科研人员知识服务[J]. 图书馆论坛，2024，44(7)：101-109.

[7] RHUDY C，JOHNSON J，PERRY C，et al. Machine learning approaches to predicting medication nonadherence：a scoping review[J]. Int J Med Inform，2025，204：106082.

[8] LI X N，CHAO T，MA P，et al. An improved variational adaptive CKF based on the Gaussian mixture model for abnormal observation and modeling uncertainty[J]. Meas Sci Technol，2025，36(8)：086102.

[9] 孟元，张轶哲，张功萱，等. 基于特征类内紧凑性的不平衡医学图像分类方法[J]. 南京大学学报（自然科学），2023，59(4)：580-589.

[10] 樊丽娟，张智琪，程晓军，等. 机器学习辅助处方合理性预测模型在围手术期合理用药管理中的应用[J]. 药物流行病学杂志，2024，33(11)：1219-1228.

[11] 齐睿娟，王应楷，杨婧，等. 基于SARIMA-LSTM组合模型构建抗过敏滴眼液消耗量预测模型[J]. 医学信息学杂志，2025，46(7)：59-65.

[12] ALOTAIBI A. Ensemble deep learning approaches in health care：a review[J]. Comput Mater Continua，2025，82(3)：3741-3771.

[13] GU Y Q，ZALKIKAR A，LIU M M，et al. Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data[J]. Sci Rep，2021，11(1)：18961.

[14] 房春丽，赖沛伦，彭怡，等. 品管圈在缩短门诊药房高峰期患者取药等候时间的应用和效果评价[J]. 中国药物滥用防治杂志，2022，28(2)：245-250.

[15] FÖRSTEL M，HAAS O，FÖRSTEL S，et al. A systematic review of features forecasting patient arrival numbers[J]. Comput Inform Nurs，2025，43(1)：e01197.

[16] 姜文彬，姜永梅，高玉芳，等. 基于KANO模型的门诊患者全方位全流程服务管理可行性分析[J]. 中国医院管理，2024，44(5)：5-9.

[17] 郑露华. 信息化管理在门诊预配排队取药系统流程改造中的应用[J]. 中医药管理杂志，2021，29(16)：174-175.

（收稿日期：2025-09-14　修回日期：2025-12-16）

（编辑：舒安琴）