临床研究 论 著

基于机器学习算法的子痫前期预测模型构建

郑江元[1]，祝锐[1]，颜永杰[1]，周洋[2]，罗亚玲[1*]

[1]重庆医科大学医学信息学院，重庆 400016；[2]重庆医科大学医学数据研究院，重庆 400016

[中图分类号] R714.244
[文献标志码] A

[DOI]
10.11855/j.issn.0577-7402.2022.08.0802

[声明]
本文所有作者声明无利益冲突

[引用本文]
郑江元, 祝锐, 颜永杰, 等. 基于机器学习算法的子痫前期预测模型构建[J]. 解放军医学杂志, 2022, 47(8): 802-808.

[收稿日期] 2021-08-11
[录用日期] 2021-10-12
[上线日期] 2022-06-10

[摘要] **目的** 筛选子痫前期的危险因素并构建基于机器学习算法的子痫前期预测模型。**方法** 收集重庆医科大学医学数据研究院大数据平台中2016年1月—2018年12月1609例住院孕妇的临床数据进行回顾性分析。依据住院期间是否发生子痫前期分为子痫前期组(n=291)与非子痫前期组(n=1318)。随机抽取70%患者的临床资料作为训练集(n=1126)构建预测模型，其余30%作为测试集(n=483)进行验证，并对测试集和训练集进行一致性检验。采用单因素分析及logistic回归分析筛选独立危险因素，利用5折交叉验证算法寻找LightGBM算法的最优参数，并基于LightGBM机器学习算法构建预测模型。**结果** 共收集了58项指标，排除缺失率≥30%的13项指标，最终共纳入45项指标。子痫前期组与非子痫前期组的谷氨酰转移酶、谷丙转氨酶、凝血酶时间、谷草转氨酶、尿比重等35项指标差异有统计学意义(P<0.05)。Logistic回归分析结果显示，尿比重、尿酸、平均红细胞血红蛋白浓度、球蛋白、血小板分布宽度、钾离子、就诊年龄、高血压家族史、收缩压、舒张压、脉搏和孕周≥34周是子痫前期的独立危险因素。经5折交叉验证，当num_leaves=5、max_depth=3、min_data_in_leaf=91、feature_fraction=0.8、bagging_fraction=0.6，bagging_freq=5时，LightGBM模型的效果达到最优，模型的曲线下面积(AUC)为0.964，敏感度为84.9%，特异度为92.7%。**结论** 基于LightGBM机器学习算法构建的子痫前期预测模型具有较好的预测效能，能够有效预测重庆地区孕妇子痫前期的发生，为临床医师提供决策参考。

[关键词] 子痫前期；机器学习；预测模型

Construction of prediction model of preeclampsia based on machine learning algorithm

Zheng Jiang-Yuan[1], Zhu Rui[1], Yan Yong-Jie[1], Zhou Yang[2], Luo Ya-Ling[1*]

[1]*College of Medical Informatics,* [2]*Medical Data Science Academy, Chongqing Medical University, Chongqing 400016, China*

[*]*Corresponding author, E-mail: gilo@163.com*

This work was supported by the National Philosophy and Social Science Foundation of China (15BGL191)

[**Abstract**] **Objective** To screen the risk factors of preeclampsia and construct the predictive model of preeclampsia based on machine learning algorithm. **Methods** A retrospective study was conducted to collect the clinical data of 1609 hospitalized pregnant women from January 2016 to December 2018 on the big data platform of Academy of Medical Data Science of Chongqing Medical University. The 1609 cases were divided into preeclampsia group (n=291) and non-preeclampsia group (n=1318) according to the occurrence of preeclampsia during hospitalization. The clinical data of 70% patients were randomly selected as the training set (n=1126) to construct the prediction model, and the remaining 30% were used as the test set (n=483) for verification, and a consistency check between training set and test set was performed. The independent risk factors were screened by univariate analysis and logistic regression analysis, and the optimal parameters of LightGBM algorithm were searched by 5-fold cross-validation algorithm, and the prediction model was constructed based on LightGBM machine learning algorithm. **Results** A total of 58 indicators were collected, 13 indicators with missing rate ⩾30% were excluded, and 45 indicators were finally included. Significant

[基金项目] 国家社会科学基金(15BGL191)
[作者简介] 郑江元，硕士研究生，主要从事电子病历数据挖掘方面的研究
[通信作者] 罗亚玲，E-mail：gilo@163.com

differences of 35 indicators existed between preeclampsia group and non-preeclampsia group (P<0.05) such as gamma-glutamyl transferase (GGT), alanine aminotrans ferase (ALT), thrombin time, aspartase transaminase (AST) and specific gravity of urine. Logistic regression analysis showed that specific gravity of urine, uric acid, hemoglobin concentration of erythrocyte, globulin, platelet distribution width, potassium ion, visiting age, family history of hypertension, systolic blood pressure, diastolic blood pressure, pulse and gestational age ⩾34 weeks were independent risk factors for preeclampsia. The results of 5-fold cross-validation showed that, when num_leaves=5, max_depth=3, min_data_in_leaf=91, feature_fraction=0.8, bagging_fraction=0.6, and bagging_freq=5, the LightGBM model achieved the best effect the area under the curve (AUC), sensitivity and specificity of LightGBM model were 0.964, 84.9% and 92.7%. **Conclusion** The prediction model of preeclampsia based on LightGBM machine learning algorithm has a higher prediction effect, which can effectively predict the occurrence of preeclampsia in pregnant women in Chongqing, and provide decisions for clinicians.

[**Key words**] preeclampsia; machine learning; prediction model

子痫前期是一种妊娠期高血压疾病，其特征是妊娠20周后出现的高血压和蛋白尿，由于其病因较多，发病机制较复杂，给孕产妇和围产儿带来了巨大的危害。文献报道，子痫前期是导致孕产妇及围产儿死亡的主要原因之一[1-2]。在过去的几十年里，尽管国内外在子痫前期相关领域取得了重大进展[3]，但是到目前为止，子痫前期的病因和发病机制仍未完全明确，且无有效的救治措施，及早发现并加强管理仍是主要的临床策略[4]。为了减少子痫前期带来的不良影响，有必要对孕妇进行子痫前期风险预测。随着智慧医学的发展，机器学习技术具有比传统统计学方法更好的优势，已被广泛应用于疾病的预测诊断中[5-8]。本研究从电子病历中收集数据，采用机器学习算法构建子痫前期风险预测模型并进行评价，以期为医护人员对子痫前期孕妇的评估和防治提供参考。

1 资料与方法

1.1 研究对象 本研究为回顾性分析，数据来源于重庆医科大学医学数据研究院大数据平台，该平台包含了重庆医科大学附属7家医疗机构的电子病历数据。收集该数据平台中2016年1月—2018年12月年龄为20~45岁的1609例住院孕妇的资料，其中子痫前期组291例，非子痫前期组1318例。在数据收集过程中，数据的提取和输入均经过检查，排除了临床资料严重缺失的病例，以及出院诊断中有糖尿病、慢性高血压病、肾脏疾病、心脏病等的病例。诊断标准：子痫前期根据中国《妊娠期高血压疾病诊治指南(2020)》[1]的标准进行诊断。本研究已通过重庆医科大学医学研究伦理委员会审批。

1.2 收集指标 收集患者的一般资料(年龄、高血压家族史、糖尿病家族史)、体征资料(收缩压、舒张压等)、妊娠情况(孕产次、妊娠期等)及实验室资料(血常规、肝功能、肾功能、电解质、凝血功能)等，排除缺失率⩾30%的指标。

1.3 指标分析 对子痫前期组与非子痫前期组患者的一般资料、体征资料、妊娠情况和实验室资料进行统计学描述及比较，分析子痫前期的影响因素。根据影响因素构建基于LightGBM机器学习算法的预测模型，并评估其效能。

1.4 统计学处理 采用SPSS 25.0软件进行统计分析，缺失率<30%的指标使用多重插补的方法填补。符合正态分布的计量资料以$\bar{x}\pm s$表示，组间比较采用t检验；不符合正态分布的计量资料以$M(Q_1, Q_3)$表示，组间比较采用Mann-Whitney U检验；计数资料以例(%)表示；对结局变量、高血压家族史、糖尿病家族史、孕周⩾34周和是否初产妇等分类变量进行赋值，采用χ^2检验进行比较。将两组间差异有统计学意义的指标纳入logistic回归分析，进一步筛选子痫前期的影响因素。P<0.05为差异有统计学意义。

1.5 机器学习模型构建 将子痫前期组与非子痫前期组按照7:3随机分为训练集(n=1126)和测试集(n=483)，并对训练集和测试集中的特征变量进行一致性检验。调用python3.7.0 lightgbm包中基于梯度提升决策树(Light Gradient Boosting Machine，LightGBM)的机器学习算法建立预测模型；采用5折交叉验证算法确定LightGBM模型的最优参数，包括num_leaves、max_depth、min_data_in_leaf、feature_fraction、bagging_fraction和bagging_freq。其中num_leaves用来提高模型的准确率，max_depth、min_data_in_leaf、feature_fraction、bagging_fraction和bagging_freq用来防止模型过度拟合。采用敏感度、特异度、准确度、曲线下面积(AUC)等指标评价测试集中预测模型的效能。

2 结 果

2.1 一般资料比较 纳入的1609例孕妇中，子痫前期291例，占18.1%，非子痫前期1318例，占81.9%。共收集了58项指标，排除缺失率⩾30%的13项指标，最终纳入45项指标。两组间谷氨酰转移酶(GGT)、谷丙转氨酶(ALT)、凝血酶时间(TT)、谷

草转氨酶(AST)、尿比重等35项指标差异有统计学意义(P<0.05)，而中性粒细胞计数、凝血酶原时间(PT)、平均红细胞体积(MCV)、淋巴细胞计数等10项指标差异无统计学意义(P>0.05，表1)。

表1 子痫前期组与非子痫前期组患者的基线资料比较

Tab.1 Comparison of baseline data between preeclampsia group and non-preeclampsia group

临床指标	子痫前期组(n=291)	非子痫前期组(n=1318)	χ^2/Z	P
GGT[U/L, $M(Q_1, Q_3)$]	16.00(9.93, 32.00)	11.20(8.00, 17.10)	−8.117	<0.001
ALT[U/L, $M(Q_1, Q_3)$]	15.00(10.00, 25.30)	11.50(8.50, 17.04)	−6.173	<0.001
中性粒细胞计数[$\times 10^9$/L, $M(Q_1, Q_3)$]	6.55(5.05, 8.19)	6.36(5.27, 7.86)	−0.610	0.542
PT[s, $M(Q_1, Q_3)$]	10.90(10.30, 11.60)	10.90(10.50, 11.40)	−0.037	0.971
TT[s, $M(Q_1, Q_3)$]	16.20(14.90, 17.60)	15.30(13.10, 16.60)	−6.854	<0.001
AST[U/L, $M(Q_1, Q_3)$]	21.70(16.80, 31.00)	18.10(15.10, 22.60)	−6.856	<0.001
尿比重[$M(Q_1, Q_3)$]	1.02(1.02, 1.03)	1.02(1.02, 1.02)	−6.085	<0.001
尿素[mmol/L, $M(Q_1, Q_3)$]	4.14(3.20, 5.11)	3.20(2.60, 3.93)	−10.776	<0.001
尿酸[μmol/L, $M(Q_1, Q_3)$]	419.00(342.00, 482.60)	316.95(270.80, 370.45)	−14.543	<0.001
总胆汁酸[μmol/L, $M(Q_1, Q_3)$]	3.70(2.00, 6.39)	2.83(1.80, 4.79)	−3.744	<0.001
TBIL[(μmol/L, $M(Q_1, Q_3)$]	7.30(5.60, 9.60)	8.80(6.70, 11.60)	−6.221	<0.001
总蛋白[g/L, $M(Q_1, Q_3)$]	60.80(55.74, 65.10)	64.59(59.68, 68.40)	−7.903	<0.001
MCV[fl, $M(Q_1, Q_3)$]	90.40(87.20, 94.00)	91.60(86.70, 95.63)	−1.899	0.058
MCHC[g/L, $M(Q_1, Q_3)$]	333.00(323.00, 341.00)	325.00(316.00, 333.00)	−8.610	<0.001
MCH[pg, $M(Q_1, Q_3)$]	30.40(28.80, 31.50)	29.80(27.80, 31.40)	−2.978	0.003
单核细胞计数[$\times 10^9$/L, $M(Q_1, Q_3)$]	0.41(0.32, 0.54)	0.43(0.31, 0.56)	−0.032	0.975
嗜碱性粒细胞计数[$\times 10^9$/L, $M(Q_1, Q_3)$]	0.02(0.01, 0.03)	0.02(0.01, 0.02)	−0.511	0.610
嗜酸性粒细胞计数[$\times 10^9$/L, $M(Q_1, Q_3)$]	0.05(0.02, 0.08)	0.05(0.03, 0.09)	−2.374	0.018
氯[mmol/L, $M(Q_1, Q_3)$]	105.50(102.60, 107.33)	105.00(103.00, 106.70)	−2.109	0.035
淋巴细胞计数[$\times 10^9$/L, $M(Q_1, Q_3)$]	1.56(1.23, 2.02)	1.54(1.27, 1.88)	−0.798	0.425
ALB[g/L, $M(Q_1, Q_3)$]	33.70(29.80, 37.00)	35.60(33.30, 38.00)	−6.922	<0.001
GLB[g/L, $M(Q_1, Q_3)$]	26.80(23.70, 30.40)	28.70(25.00, 32.00)	−4.871	<0.001
WBC[$\times 10^9$/L, $M(Q_1, Q_3)$]	8.70(7.28, 10.78)	8.50(7.21, 10.13)	−1.187	0.235
DBIL[μmol/L, $M(Q_1, Q_3)$]	1.70(1.30, 2.50)	2.20(1.70, 2.80)	−6.643	<0.001
ALP[U/L, $M(Q_1, Q_3)$]	156.00(123.70, 199.60)	172.00(134.25, 214.30)	−3.190	0.001
RBC[$\times 10^{12}$/L, $M(Q_1, Q_3)$]	4.04(3.80, 4.43)	3.96(3.70, 4.24)	−3.625	<0.001
HCT[%, $M(Q_1, Q_3)$]	36.70(34.00, 39.50)	35.90(33.50, 38.30)	−2.973	0.003
FIB[g/L, $M(Q_1, Q_3)$]	4.09(3.50, 4.76)	3.78(3.32, 4.44)	−5.039	<0.001
肌酐[μmol/L, $M(Q_1, Q_3)$]	54.60(45.40, 64.50)	46.40(41.30, 52.60)	−10.004	<0.001
葡萄糖[mmol/L, $M(Q_1, Q_3)$]	4.62(4.00, 5.40)	4.40(3.97, 5.01)	−3.127	0.002
PDW[%, $M(Q_1, Q_3)$]	16.60(15.70, 17.90)	16.50(15.90, 17.00)	−2.596	0.009
PCT[%, $M(Q_1, Q_3)$]	0.20(0.16, 0.24)	0.20(0.17, 0.23)	−0.249	0.803
PLT[$\times 10^9$/L, $M(Q_1, Q_3)$]	166.00(133.00, 215.00)	179.00(144.75, 222.00)	−3.082	0.002
HB[g/L, $M(Q_1, Q_3)$]	122.00(111.00, 133.00)	117.00(107.75, 126.00)	−5.391	<0.001
钠[mmol/L, $M(Q_1, Q_3)$]	137.90(136.34, 139.00)	138.20(137.00, 139.80)	−4.160	<0.001
钾[mmol/L, $M(Q_1, Q_3)$]	4.10(3.90, 4.31)	3.90(3.70, 4.10)	−9.410	<0.001
IBIL[(μmol/L, $M(Q_1, Q_3)$]	5.44(4.20, 7.32)	6.60(4.90, 8.90)	−5.860	<0.001
就诊年龄[岁, $M(Q_1, Q_3)$]	29(25, 33)	27(24, 30)	−5.094	<0.001
收缩压[mmHg, $M(Q_1, Q_3)$]	144(134, 158)	112(106, 120)	−24.215	<0.001
舒张压[mmHg, $M(Q_1, Q_3)$]	95(86, 102)	71(68, 80)	−23.023	<0.001
脉搏[次/min, $M(Q_1, Q_3)$]	86(80, 97)	80(80, 90)	−6.053	<0.001
高血压家族史[例(%)]	48(16.5)	56(4.2)	59.126	<0.001
糖尿病家族史[例(%)]	11(3.8)	27(2.0)	3.099	0.078
孕周≥34周[例(%)]	210(72.2)	1263(95.8)	172.476	<0.001
是否初产妇[例(%)]	167(57.4)	742(56.3)	0.115	0.734

GGT. 谷氨酰转移酶；ALT. 谷丙转氨酶；PT. 凝血酶原时间；TT. 凝血酶时间；AST. 谷草转氨酶；TBIL. 总胆红素；MCV. 平均红细胞体积；MCHC. 平均红细胞血红蛋白浓度；MCH. 平均红细胞血红蛋白含量；ALB. 白蛋白；GLB. 球蛋白；WBC. 白细胞计数；DBIL. 直接胆红素；ALP. 碱性磷酸酶；RBC. 红细胞计数；HCT. 血细胞比容；FIB. 纤维蛋白原浓度；PDW. 血小板分布宽度；PCT. 血小板压积；PLT. 血小板计数；HB. 血红蛋白；IBIL. 间接胆红素

2.2 子痫前期的影响因素分析 利用二元logistic回归分析对这35项指标进一步筛选，其中11项指标[尿比重、尿酸、平均红细胞血红蛋白浓度(MCHC)、球蛋白、血小板分布宽度(PDW)、钾离子、就诊年龄、收缩压、舒张压、脉搏和孕周≥34周]差异有统计学意义(P<0.05)，可作为子痫前期的独立危险因素；此外，高血压家族史虽然差异无统计学意义(P=0.063)，但通过咨询临床专家和查阅参考文献，最终也作为子痫前期的影响因素(表2)。

表2 子痫前期差异性指标logistic回归分析

Tab.2 Logistic regression analysis of difference index in preeclampsia

变量	β	SE	OR	95%CI	P
尿比重	43.036	17.973	4.899	2463.308~9.740	0.017
尿酸	0.008	0.001	1.008	1.005~1.011	<0.001
MCHC	0.016	0.007	1.016	1.003~1.029	0.019
GLB	−0.110	0.025	0.896	0.853~0.941	<0.001
PDW	0.116	0.044	1.123	1.03~1.225	0.009
钾	1.144	0.385	3.139	1.477~6.67	0.003
就诊年龄	0.064	0.026	1.066	1.012~1.123	0.016
高血压家族史	0.852	0.459	2.343	1.159~1.218	0.063
收缩压	0.172	0.013	1.188	1.159~1.218	<0.001
舒张压	0.006	0.003	1.006	1.001~1.011	0.019
脉搏	0.027	0.012	1.027	1.004~1.051	0.021
孕周≥34周	−1.738	0.441	0.176	0.074~0.417	<0.001

MCHC. 平均红细胞血红蛋白浓度；GLB. 球蛋白；PDW. 血小板分布宽度

2.3 机器学习模型 将上述12项独立危险因素作为预测模型的输入变量，孕妇是否发生子痫前期作为结局变量，并将子痫前期组与非子痫前期组按照7:3随机分为训练集和测试集，对训练集和测试集中的特征变量进行一致性检验，所有特征变量在训练集和测试集中均满足一致性检验(P>0.05，表3)。在训练过程中，采用5折交叉验证算法对LightGBM模型的参数进行优化，调优参数的范围为：num_leaves为5~100，max_depth为3~8，min_data_in_leaf为1~102，feature_fraction为0.6~1.0，bagging_fraction为0.6~1.0，bagging_freq为0~50。经过试验，参数设置为num_leaves=5、max_depth=3、min_data_in_leaf=91、feature_fraction=0.8、bagging_fraction=0.6、bagging_freq=5时，LightGBM模型的预测效果达到最优，模型的曲线下面积(AUC)为0.964，敏感度为84.9%，特异度为92.7%。

表3 特征变量在训练集和测试集中的一致性检验

Tab.3 Consistency test of characteristic variables in training set and test set

特征变量	KS值	P
尿比重	0.032	0.871
尿酸	0.033	0.862
MCHC	0.056	0.225
GLB	0.044	0.510
PDW	0.029	0.932
钾	0.023	0.992
就诊年龄	0.029	0.936
高血压家族史	0.007	1.000
收缩压	0.052	0.318
舒张压	0.026	0.972
脉搏	0.026	0.977
孕周≥34周	0.008	1.000

MCHC. 平均红细胞血红蛋白浓度；GLB. 球蛋白；PDW. 血小板分布宽度

2.4 模型效能检验 采用测试集对模型的效能进行验证，结果显示，LightGBM模型的敏感度和AUC均高于logistic回归模型，但logistic回归模型的特异度和准确度高于LightGBM模型(图1、表4)。

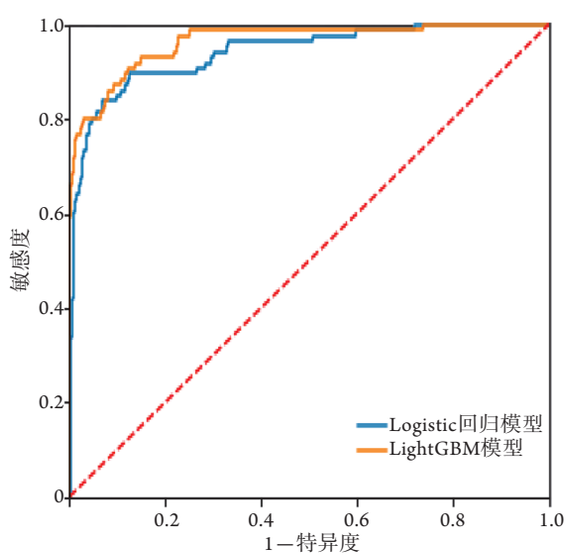

图1 Logistic回归模型与LightGBM模型在测试集中的ROC曲线图

Fig.1 ROC plot of logistic regression model and LightGBM model in test set

表4 Logistic回归模型与LightGBM模型的效能评价

Tab.4 Performance evaluation of logistic regression model and LightGBM model

模型	真阳性(例)	真阴性(例)	假阳性(例)	假阴性(例)	敏感度(%)	特异度(%)	准确度(%)	AUC
Logistic回归模型	59	381	16	27	68.6	96.0	78.7	0.936
LightGBM模型	73	368	29	13	84.9	92.7	71.6	0.964

AUC. 曲线下的面积

3 讨 论

子痫前期存在多因素、多机制、多通路发病的综合征性质[1]，唯一的治疗方法为中断妊娠，但是可能会增加母婴早产并发症的风险。虽然已有学者将检查指标用于子痫前期的诊断预测[9-11]，如可溶性血管内皮生长因子(soluble fms-like tyrosine kinase 1，sFlt-1)和胎盘生长因子(placental growth factor，PlGF)等，但其预测效果并不理想[12]，且在低收入和中等收入国家推广较为困难。本研究通过分析重庆医科大学医学数据研究院大数据平台中的1609例住院孕妇的临床数据构建了预测模型，以辅助初级临床医师和基层医疗机构评估子痫前期的发生风险。

本研究筛选出子痫前期的12项影响因素，其中，球蛋白和孕周≥34周两个指标为保护因素，尿比重、尿酸、平均红细胞血红蛋白浓度等指标为危险因素。目前，患者年龄、高血压家族史、收缩压和舒张压这4项指标对子痫前期发生风险的影响已被广泛报道[13-16]，而尿比重、尿酸、平均红细胞血红蛋白浓度、球蛋白、血小板分布宽度、钾离子、脉搏和孕周≥34周在子痫前期中的作用则少见报道。

有研究发现，与健康孕妇比较，子痫前期孕妇在妊娠期间更容易消耗血清免疫球蛋白，导致血清球蛋白浓度降低[17]。本研究发现，球蛋白为子痫前期的保护因素，球蛋白浓度越高，所消耗的球蛋白越少，患子痫前期的风险越小，与文献报道一致。临床上将孕周<34周定义为早发型子痫前期，孕周≥34周定义为迟发型子痫前期[18]，其中早发型子痫前期不良出生结局的发生率高于迟发型子痫前期，病情更危重，发生多器官功能损伤的风险更高[19]。因此，孕周越高孕妇发生子痫前期的风险越低。

子痫前期患者可能存在肾功能受损，由于肾脏灌注和肾小球滤过率下降，尿比重增高时，尿液浓缩，尿酸清除率下降，导致尿酸增多[20]。此外，近端小管对尿酸的重吸收增加和排泄减少，使子痫前期患者尿酸进一步增多[21]。有研究报道，子痫前期患者红细胞聚集能力增强，变形能力减弱，血浆扩容不足，导致血小板黏附于血管壁，红细胞膜破坏，红细胞血红蛋白浓度增加，血红蛋白/血细胞比容比值增高，血液黏度增高[22-23]。已有研究发现，与正常妊娠晚期比较，子痫前期患者血小板计数更低，原因为血小板平均容积、血小板分布宽度升高，使得血小板消耗增加，从而导致血小板计数减少[24]。此外，据文献报道，妊娠前中期高血钾水平与严重子痫前期的发展风险较高相关[25]。妊娠期间醛固酮和孕酮可影响孕妇的血钾水平，因此，血钾水平升高可能提示醛固酮和孕酮紊乱，而醛固酮和孕酮紊乱又可能与子痫前期的发生有关[26]。美国妇产科医师学会(American College of Obstetricians and Gynecologists，ACOG)提出，心率是子痫前期的一个预警指标[27]。子痫前期孕妇可能存在心功能损害，原因包括：(1)血管阻力增加，心脏后负荷加重；(2)肾素-血管紧张素-醛固酮系统平衡被破坏，造成水钠潴留，引起血液浓缩；(3)贫血、低蛋白血症导致血浆胶体渗透压降低；(4)血浆扩容不足[28]。而脉搏与心率相关，脉搏越快则提示心率越快，孕妇患子痫前期的风险越大[29]。

随着大数据时代的来临，机器学习在医疗卫生领域的应用越来越广泛，尤其是在疾病的预测和预后评估方面[30-31]。本研究根据筛选出来的子痫前期影响因素，构建了子痫前期预测模型，以预测孕妇子痫前期的发生风险。结果显示，LightGBM模型的效果达到最优时，其AUC为0.964，敏感度为0.849，特异度为0.927。本研究LightGBM模型的AUC高于Jhee等[26]的模型(敏感度=0.603，特异度=0.991，AUC=0.924)。但是，Jhee等[26]的模型由于病例组(n=474)与对照组(n=10058)例数不平衡，导致敏感度及特异度相差过大，而本研究的LightGBM模型在敏感度及特异度相差过大的问题上有所改善，综合性能较之前的预测模型有所提高。Logistic回归模型的可解释性非常好，从特征的权重可以解释不同特征对最后结果的影响，在医疗卫生领域可用于探索疾病的相关影响因素，但因为模型简单，容易出现欠拟合、模型总体效能不高等问题。LightGBM是一种快速的、分布式的、高性能的基于决策树算法的梯度提升框架[32]。LightGBM模型采用直方图算法对数据进行分割，通过离散化的统计量遍历寻找最优分割点，减小内存，提高训练速度[33]；采用有深度限制的按叶子生长策略，从当前叶子节点中找到增益值最大的节点进行分裂，并对树的深度进行限制，防止过度拟合，缩短寻找最优深度树的时间，降低了误差，提高了预测准确度[34]。

综上所述，本研究构建了基于机器学习算法的子痫前期预测模型，并利用敏感度、特异度、准确度和AUC等评价指标对构建的机器学习模型进行评价，一定程度上减少了单一评价指标带来的偏倚。同时，本研究结合了母体因素和常见的产前实验室检查指标，纳入的患者来自多个中心，样本量大且具有良好的代表性，可以有效地预测子痫前期的发生风险，对临床上孕妇子痫前期的早期识别有一定的辅助作用，具有潜在的临床价值。

本研究仍存在一些不足之处：(1)数据均来源于重庆地区，可能存在选择偏倚，需要进行外部验证以进一步评估模型的效能；(2)本文构建的预测模型综合效能较高，包含了12项指标，虽然均为易于获得的常规实验室检查指标，但指标数量较多，在临床推广应用有一定困难；(3)研究中部分指标的缺失率过大，如BMI是孕产妇产检的重要指标，但在本研究中由于该指标缺失率过大而未纳入模型中，重要指标的缺失可能会对模型的效能产生一些影响。因此，未来仍需进一步论证该指标的缺失是否会对预测结果有较大影响。

【参考文献】

[1] Hypertensive Disorders in Pregnancy Subgroup, Chinese Society of Obstetrics and Gynecology, Chinese Medical Association. Diagnosis and treatment of hypertension and pre-eclampsia in pregnancy: a clinical practice guideline in China (2020)[J]. Chin J Obstet Gynecol, 2020, 55(4): 227-238. [中华医学会妇产科学分会妊娠期高血压疾病学组. 妊娠期高血压疾病诊治指南(2020)[J]. 中华妇产科杂志, 2020, 55(4): 227-238.]

[2] Nobles CJ, Mendola P, Mumford SL, *et al*. Preconception blood pressure and its change into early pregnancy: early risk factors for preeclampsia and gestational hypertension[J]. Hypertension, 2020, 76(3): 922-929.

[3] Rana S, Lemoine E, Granger JP, *et al*. Preeclampsia: pathophysiology, challenges, and perspectives[J]. Circ Res, 2019, 124(7): 1094-1112.

[4] Phipps EA, Thadhani R, Benzing T, *et al*. Pre-eclampsia: pathogenesis, novel diagnostics and therapies[J]. Nat Rev Nephrol, 2019, 15(5): 275-289.

[5] Heo J, Yoon JG, Park H, *et al*. Machine learning-based model for prediction of outcomes in acute stroke[J]. Stroke, 2019, 50(5): 1263-1265.

[6] Bi Q, Goodman KE, Kaminsky J, *et al*. What is machine learning? A primer for the epidemiologist[J]. Am J Epidemiol, 2019, 188(12): 2222-2239.

[7] Shi QH, Zhang ZF, Hu B, *et al*. Recent advances in the use of deep learning and artificial intelligence in the diagnosis and treatment of cervical and lumbar spine degenerative diseases[J]. Med J Chin PLA, 2021, 46(10): 1034-1039. [施强慧, 张子凡, 胡博, 等. 深度学习与人工智能在颈腰椎退变性疾病诊断及治疗中的应用研究进展[J]. 解放军医学杂志, 2021, 46(10): 1034-1039.]

[8] Wu WT, Li YJ, Feng AZ, *et al*. Data mining in clinical big data: the frequently used databases, steps, and methodological models[J]. Mil Med Res, 2021, 8(4): 552-563.

[9] Correa PJ, Palmeiro Y, Soto MJ, *et al*. Etiopathogenesis, prediction, and prevention of preeclampsia[J]. Hypertens Pregnancy, 2016, 35(3): 280-294.

[10] Yang Z. Pay attention to the standardized diagnosis and treatment of hypertensive disorder complicating pregnancy[J]. J Pract Obstet Gynecol, 2020, 36(12): 881-885. [杨孜. 重视妊娠期高血压疾病的规范化诊断与处理[J]. 实用妇产科杂志, 2020, 36(12): 881-885.]

[11] Al-Rubaie Z, Askie LM, Ray JG, *et al*. The performance of risk prediction models for pre-eclampsia using routinely collected maternal characteristics and comparison with models that include specialised tests and with clinical guideline decision rules: a systematic review[J]. BJOG, 2016, 123(9): 1441-1452.

[12] Malik A, Jee B, Gupta SK. Preeclampsia: Disease biology and burden, its management strategies with reference to India[J]. Pregnancy Hypertens, 2019, 15: 23-31.

[13] Bartsch E, Medcalf KE, Park AL, *et al*. Clinical risk factors for pre-eclampsia determined in early pregnancy: systematic review and meta-analysis of large cohort studies[J]. BMJ, 2016, 353: i1753.

[14] Burton GJ, Redman CW, Roberts JM, *et al*. Pre-eclampsia: pathophysiology and clinical implications[J]. BMJ, 2019, 366: l2381.

[15] Phipps E, Prasanna D, Brima W, *et al*. Preeclampsia: updates in pathogenesis, definitions, and guidelines[J]. Clin J Am Soc Nephrol, 2016, 11(6): 1102-1113.

[16] Li K, Zhu DW, Chen JK, *et al*. Progress in the study for pathogenesis and clinical treatment of preeclampsia[J]. Med J Chin PLA, 2019, 44(5): 423-429. [李可, 朱大伟, 陈建昆, 等. 子痫前期发病机制与临床治疗研究进展[J]. 解放军医学杂志, 2019, 44(5): 423-429.]

[17] Dong CY. Relationship between platelet count, serum globulin and fibrinogen in patients with hypertensive disorder complicating pregnancy[J]. Matern Child Health Care Chin, 2013, 28(35): 5797-5798. [董春玉. HDP患者PLT计数、血清球蛋白与FIB的关系[J]. 中国妇幼保健, 2013, 28(35): 5797-5798.]

[18] Hung TH, Hsieh TT, Chen SF. Risk of abnormal fetal growth in women with early- and late-onset preeclampsia[J]. Pregnancy Hypertens, 2018, 12: 201-206.

[19] Wang MH, Tian WJ, Meng JL, *et al*. Interpretation of laboratory test results in early and late onset severe preeclampsia[J]. Chin J Lab Med, 2017, 40(3): 180-185. [王明辉, 田文君, 孟金来, 等. 早发型与晚发型重度子痫前期实验室检查结果及一般情况分析[J]. 中华检验医学杂志, 2017, 40(3): 180-185.]

[20] Khaliq OP, Konoshita T, Moodley J, *et al*. The role of uric acid in preeclampsia: is uric acid a causative factor or a sign of preeclampsia?[J]. Curr Hypertens Rep, 2018, 20(9): 1-9.

[21] Yang YK, Qi HB. Interpretation of the main points of ACOG guidelines for gestational hypertension and preeclampsia (2019)[J]. Chin J Pract Gynecol Obstet, 2019, 35(8): 895-899. [杨怡珂, 漆洪波. 美国妇产科医师学会(ACOG)“妊娠期高血压和子痫前期指南2019版”要点解读(第一部分)[J]. 中国实用妇科与产科杂志, 2019, 35(8): 895-899.]

[22] Wang C, Lin L, Su R, *et al*. Hemoglobin levels during the first trimester of pregnancy are associated with the risk of gestational diabetes mellitus, pre-eclampsia and preterm birth in Chinese women: a retrospective study[J]. BMC Pregnancy Childbirth, 2018, 18(1): 263.

[23] Li Y, Wu SW, Chen Y. Correlation analysis of adverse pregnancy outcomes of hypertensive disorder complicating pregnancy[J]. Chin J Fam Plan Gynecotokol, 2018, 10(9): 43-47. [李莹, 伍绍文, 陈奕. 妊娠期高血压疾病不良妊娠结局相关分析[J]. 中国计划生育和妇产科, 2018, 10(9): 43-47.]

[24] Xu C, Li YH, Zhang W, *et al*. Change of coagulation function in preeclampsia and its prevention and treatment[J]. J Pract Obstet Gynecol, 2019, 35(2): 113-116. [徐畅, 李昀晖, 张文, 等. 子痫前期患者凝血功能变化检测指标及其防治[J]. 实用妇产科

杂志, 2019, 35(2): 113-116.]

[25] Fotiou M, Michaelidou AM, Masoura S, *et al*. Second trimester amniotic fluid uric acid, potassium, and cysteine to methionine ratio levels as possible signs of early preeclampsia: a case report[J]. Taiwan J Obstet Gynecol, 2016, 55(6): 874-876.

[26] Jhee JH, Lee S, Park Y, *et al*. Prediction model development of late-onset preeclampsia using machine learning-based methods[J]. PLoS One, 2019, 14(8): e0221202.

[27] Bernstein PS, Martin JN Jr, Barton JR, *et al*. National partnership for maternal safety: consensus bundle on severe hypertension during pregnancy and the postpartum period[J]. Obstet Gynecol, 2017, 130(2): 347-357.

[28] Peng T, Li XT. Prevention and treatment of preeclampsia complicated with cardiac insufficiency[J]. Chin J Pract Gynecol Obstet, 2019, 35(11): 1213-1217. [彭婷, 李笑天. 子痫前期患者合并心脏功能不全防治[J]. 中国实用妇科与产科杂志, 2019, 35(11): 1213-1217.]

[29] Gu WR, Li XT. Intervention and management of preeclampsia[J]. Chin J Pract Gynecol Obstet, 2020, 36(2): 120-123. [顾蔚蓉, 李笑天. 子痫前期的干预与管理[J]. 中国实用妇科与产科杂志, 2020, 36(2): 120-123.]

[30] Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery[J]. Lancet Oncol, 2019, 20(5): e262-e273.

[31] Tan JT, Xu XM, He YX, *et al*. Construction of prediction model of cirrhosis-related hepatic encephalopathy based on machine learning algorithm[J]. Med J Chin PLA, 2021, 46(4): 354-360. [谈军涛, 许晓梅, 何雨芯, 等. 基于机器学习算法的肝硬化相关肝性脑病预测模型的构建[J]. 解放军医学杂志, 2021, 46(4): 354-360.]

[32] Zhou C, Wang Y, Ji MH, *et al*. Predicting peritoneal metastasis of gastric cancer patients based on machine learning[J]. Cancer Control, 2020, 27(1): 1073274820968900.

[33] Li N, Li BL, Zhu JH, *et al*. Transient stability assessment method considering sample imbalance and overlap[J]. AEPS, 2020, 44(21): 64-71. [李楠, 李保罗, 朱建华, 等. 计及样本不平衡与重叠的暂态稳定评估方法[J]. 电力系统自动化, 2020, 44(21): 64-71.]

[34] Xu GT, Shen YT. A malware detection method based on XGBoost and LightGBM two-layer model[J]. Netinfo Security, 2020, 20(12): 54-63. [徐国天, 沈耀童. 基于XGBoost和LightGBM双层模型的恶意软件检测方法[J]. 信息网络安全, 2020, 20(12): 54-63.]

(责任编辑：张小利)