

DOI 10.19656/j.cnki.1002-2406.20230804

信息工程

# 不同机器算法在乳腺癌发病风险预测模型中使用效果的评估

周亮<sup>1</sup>, 王月<sup>1</sup>✉, 田凌嘉<sup>2</sup>, 刘丽芳<sup>1</sup>, 张琴<sup>1</sup>, 梁黎昕<sup>1</sup>

(1. 湖南中医药大学第一附属医院, 湖南 长沙 410007; 2. 湖南中医药大学, 湖南 长沙 410208)

**【摘要】**目的:使用不同机器学习算法开发乳腺癌发病风险预测模型。方法:采用湖南中医药大学第一附属医院乳腺科患者数据库作为数据来源;根据乳腺癌相关风险因素,选取数据库中的初潮时间、流产次数、生育情况、月经及母乳喂养情况、乳腺癌家族史、作息时间、饮食习惯及中医证候特征等作为建模候选变量,提取其人口学特征、生命体征、病理检查等数据;使用6种机器算法开发模型,并对不同机器学习算法在预测模型中使用效果进行评估。结果:综合两种计算方法对两种建模算法的特征重要度预测结果,可以得出年龄、燥热、流产次数、是否曾患乳腺炎、生活中是否经常锻炼、第一次月经时间等变量对乳腺癌的预测可能有重要作用。随机森林算法的预测结果最好,准确率为0.86,AUC值达0.89,XGboost和GBDT算法的准确率都为0.85,AUC值也同为0.85,其次是逻辑回归算法的准确率和AUC值都为0.84,SVC和DT算法的预测准确率分别为0.83、0.79,AUC值分别为0.82、0.71。从各算法的建模结果可以看出,随机森林算法由于集成学习的特性,本身的精度比一般单个算法的要好,预测结果的准确性也高。结论:基于森林算法的乳腺癌患者发病风险预测模型对辅助临床医生指导患者乳腺癌发病风险预防有重要意义。

**【关键词】**机器算法;逻辑回归;随机森林;DT;GBDT;XGboost;SVC;乳腺癌;风险预测

## 【引用格式】

周亮,王月,田凌嘉,等.不同机器算法在乳腺癌发病风险预测模型中使用效果的评估[J].中医药信息,2023,40(8):23-28.

ZHOU L, WANG Y, TIAN L J, et al. Effect evaluation of different machine algorithms in breast cancer risk prediction model[J]. Information on TCM, 2023, 40(8):23-28.

乳腺癌作为女性中最常见的恶性肿瘤,发病率居女性中首位<sup>[1]</sup>。近十年来我国乳腺癌的发病率呈直线上升趋势,年增长率达3%~4%,高于全球增长速度<sup>[2]</sup>。现阶段,乳腺癌的规范化治疗包括手术切除、放、化疗,内分泌及分子靶向辅助中医药治疗的综合治疗模式。规范化治疗有效改善了患者生存时长,但手术及放、化疗对患者的身心造成了很大的创伤,严重影响生活质量。

当前,中国女性乳腺癌早期发现的比率低,中晚期乳腺癌比率明显高于欧美发达国家。随着我国经济及

医学技术发展,乳腺癌的研究重点已经从规范化治疗转变为早期风险预测和定期筛查,早诊、早治成为了提高乳腺癌治愈率的最佳途径,而预防是避免罹患乳腺癌的最有效方法<sup>[3]</sup>。因此,乳腺相关风险因素的预测对乳腺癌的基因检测及临床管理至关重要。FU等<sup>[4]</sup>连续从8 067例中国女性乳腺癌患者和13 129例中国无癌女性对照中筛选出15个既定或潜在的乳腺癌易感基因的致病变体,结果表明携带TP53、BRCA1、BRCA2和PALB2致病变异型的患者更常见早发型乳腺癌、双侧乳腺癌、有乳腺癌家族史和/或任何癌症。WIGGS等<sup>[5]</sup>

基金项目:湖南省卫生健康委科研项目(202112070480);湖南省中医药科研计划项目(C2022034);湖南省临床医疗技术创新引导项目(2021SK51417)

第一作者简介:周亮(1978-),女,主任医师,主要研究方向:乳腺疾病防治。

✉通信作者简介:王月(1985-),女,主治医师,主要研究方向:乳腺疾病的研究。

对43项乳腺癌进行风险研究总结,纳入饮食和运动干预的研究表明,更好地控制有害的雌激素形式和水平,可能是预防绝经后妇女乳腺癌发展的最佳策略。

目前,针对乳腺癌患病独立危险因素分析以及建立相关风险预测模型报道尚不够全面,面对海量临床数据,标准化、信息化的数据分析引起广泛关注。本研究选取湖南中医药大学第一附属医院827例乳腺癌患者,通过对临床相关数据进行6种算法的回顾性分析,构建乳腺癌患者患病风险的数据模型,最终为乳腺癌患病风险提供相关预测方法,降低发病率。

## 1 资料与方法

### 1.1 一般资料

选择2018年7月1日—2021年6月30日于湖南中医药大学第一附属医院乳腺科就诊的乳腺癌女性患者,年龄( $45.0 \pm 4.3$ )岁,共827例。

### 1.2 诊断标准

符合《中国抗癌协会乳腺癌诊治指南与规范(2015版)》<sup>[6]</sup>的诊断标准,并经病理学确诊。

### 1.3 纳入标准

①符合纳入标准;②经影像学以及术后病理组织学检查确诊为乳腺浸润性导管癌者;③患者性别为女性;④临床资料完整,有彩色超声影像结果、病理组织标本等。

### 1.4 排除标准

①合并肺部、胃部等其他恶性肿瘤;②具有哮喘、高血压、心脏病等基础疾病。

### 1.5 研究结果与预测变量

本研究是将我院就诊的827例乳腺癌患者调查问卷的数据纳入数据库中。研究结果为患者确诊前是否具有相关危险因素。研究特征:患者年龄、性别、职业、学历、就诊时间等。初潮时间、流产次数、生育情况、月经及母乳喂养情况、乳腺癌家族史、作息时间、饮食习惯等作为建模候选变量。

### 1.6 数据预处理

删除空值,修改填充异常值(众数,均值),对分类变量数值化。①异常值。更正处理:部分记录错误的数据,若可更正的,取均值或边界值;若不可以更正的,予以删除处理,删除处理的内容为缺失值。②缺失值。补全处理:部分缺失值经医学专家确认后在原数据补全,部分缺失值采用阴性值补全,部分缺失值不予补全,决定不采用插补法或多重插补法予以填补。删除规则:分析中对缺失太多的变量特征,经与医学专家讨论确认进行特征删除,其他缺失在分析中均采用个案删除的方式。③极端值处理。数据中心若出现极端数值,均不做

处理。④特征标准化。连续特征采用标准分数( $z$ -score)标准化方法转换,经转换后的数据满足均值为0且标准差为1,离散特征做独热编码 one-hot 变换,实现离散特征的“标准化”。

### 1.7 统计分析与建模

采用Python 2.7对数据进行建模分析。定量资料若符合正态分布,以 $\bar{x} \pm s$ 表示,用Student's  $t$  test做组间比较。不符合正态分布的定量资料以中位数表示,用Wilcoxon秩和检验做组间比较。定性资料以[例(%)]表示,用卡方检验做组间比较。以 $P < 0.05$ 表示差异具有统计学意义。使用逻辑回归,随机森林、DT、GBDT、XGboost、SVC 6种算法对BRCA数据进行建模处理。

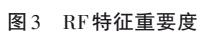
## 2 结果

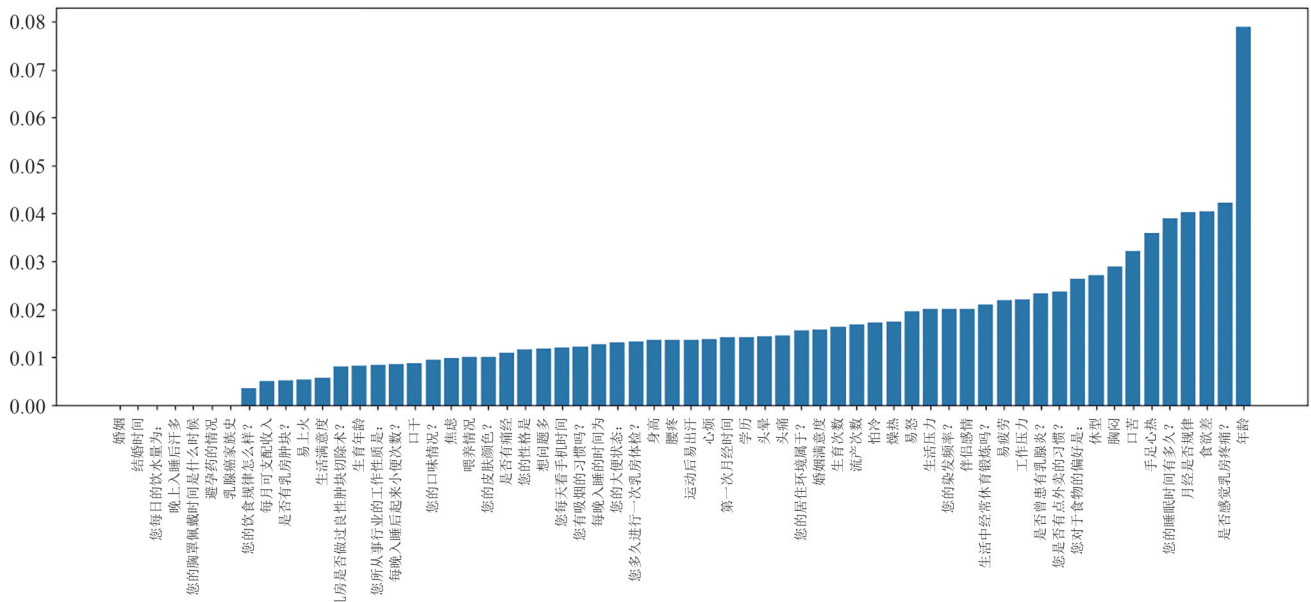
### 2.1 两种预测方法结果

本研究采用了逻辑回归(LR),随机森林(RF)、DT、GBDT、XGboost、SVC机器学习算法对BRCA数据进行建模分析,数据集随机划分为训练集和测试集,占比为训练集:测试集=7:3。建模分别采用了scikit-learn库内置特征变量重要度计算方法和基于博弈论中的Shapley值对特征变量重要度预测方法。两种方法的分析结果见图1~图4所示。图1和图2为基于Shapley值方法的RF和XBG算法特征变量重要度结果,图3和图4显示了Scikit-learn库内置特征变量重要度计算方法计算RF和XGB算法的特征重要度。综合两种计算方法对两种建模算法的特征重要度预测结果,可以得出年龄、燥热、流产次数、是否曾患乳腺炎、生活中是否经常锻炼、第一次月经时间等变量对乳腺癌的预测可能有重要作用。

### 2.2 6种机器学习建模算法结果

本研究选取了Accuracy、Precision、Recall、f1、AUC等评价指标对6种机器学习建模算法的结果进行评估,具体实验结果见表1,各算法的ROC曲线见图5。从表1的结果可以看出,针对本研究的实验数据集,随机森林算法的预测结果最好,准确率为0.86,AUC值达0.89,XGboost和GBDT算法的准确率都为0.85,AUC值也同为0.85,其次是逻辑回归算法的准确率和AUC值都为0.84,SVC和DT算法的预测准确率分别为0.83、0.79,AUC值分别为0.82、0.71。从各算法的建模结果可以看出,随机森林算法由于集成学习的特性,本身的精度比一般单个算法的要好,预测结果的准确性也高。DT算法相对XGBoost和GBDT等梯度提升树算法,由于本身的限制,在本实验数据集上的结果并不理想。考虑到数据本身的局限性对模型性能的影响,后续可以收集更多的BRCA数据以优化预测模型。





XGB Feature Importance

图4 XGB特征重要度

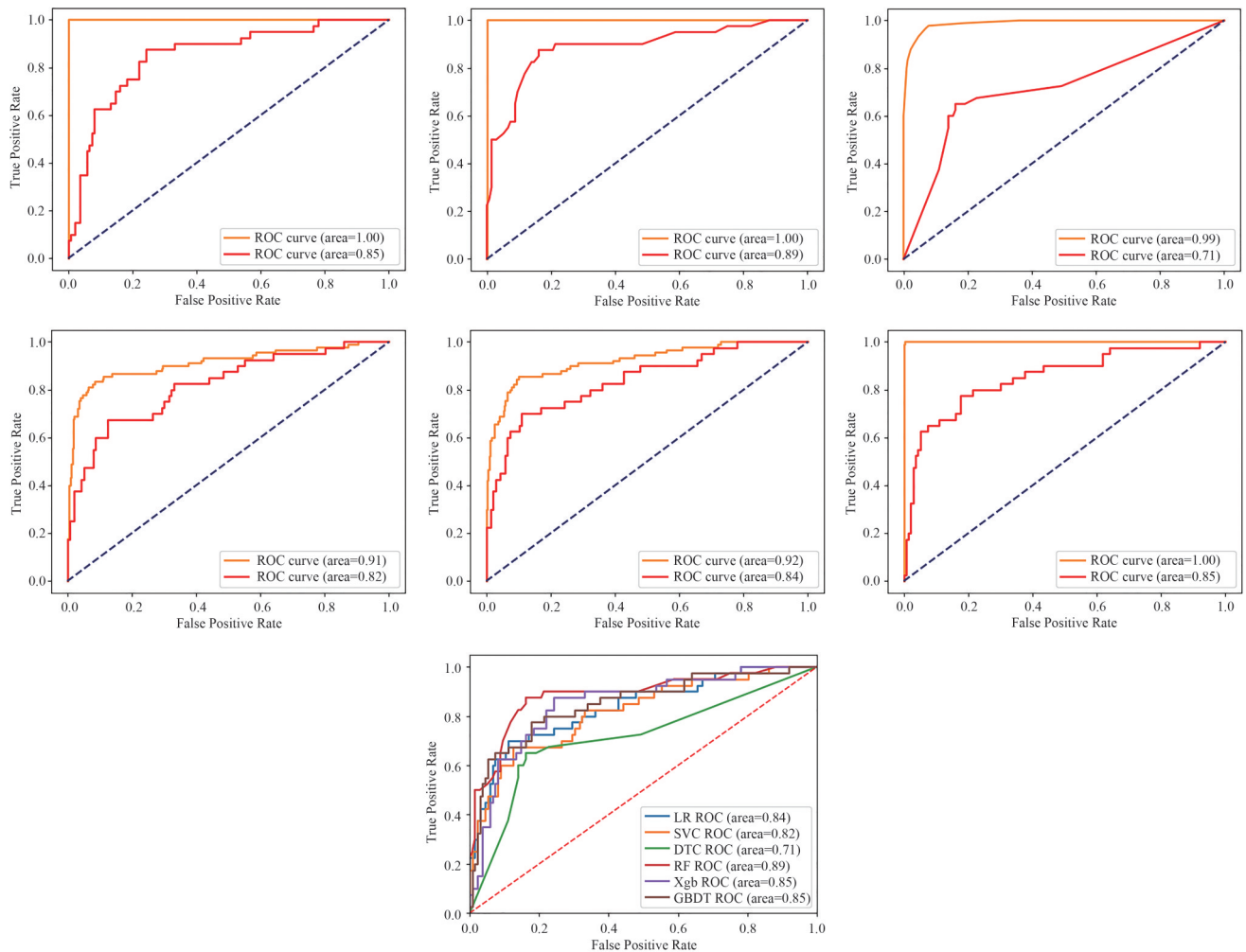


图5 各算法ROC曲线图



表 1 6 种算法评价指标得分

| 算法      | Accuracy | Precision | Recall | F1   | AUC  |
|---------|----------|-----------|--------|------|------|
| 逻辑回归    | 0.91     | 0.88      | 0.66   | 0.75 | 0.92 |
|         | 0.84     | 0.70      | 0.48   | 0.57 | 0.84 |
| SVC     | 0.92     | 0.87      | 0.72   | 0.79 | 0.91 |
|         | 0.83     | 0.63      | 0.60   | 0.62 | 0.82 |
| DT      | 0.97     | 0.92      | 0.88   | 0.90 | 0.99 |
|         | 0.79     | 0.53      | 0.60   | 0.57 | 0.71 |
| 随机森林    | 1.00     | 1.00      | 1.00   | 1.00 | 1.00 |
|         | 0.86     | 0.90      | 0.45   | 0.60 | 0.89 |
| XGboost | 1.00     | 1.00      | 1.00   | 1.00 | 1.00 |
|         | 0.85     | 0.69      | 0.63   | 0.66 | 0.85 |
| GBDT    | 0.99     | 1.00      | 0.96   | 0.98 | 1.00 |
|         | 0.85     | 0.67      | 0.65   | 0.66 | 0.85 |

3 讨论

乳腺癌目前已经超过肺癌成为全球发病率首位的疾病。随着诊疗技术的提高,风险预防及诊疗趋于个体化、精准化,而大数据分析技术作为分析工具在乳腺癌精准风险模型分析中不容忽视<sup>[7]</sup>。本文提出了一种基于不同机器算法的乳腺癌风险预测模型。从预测的准确度方面来看,在现有的数据上搭建风险预测模型是可行的,并且有较高的预测精度和性能。目前机器学习在计算机视觉、自然语言处理、搜索引擎与推荐系统等领域正发挥着不可替代的作用,相关应用涉及互联网、安防、医疗、交通和金融等众多行业,对社会发展起到了促进作用<sup>[8]</sup>。本研究利用基于不同机器学习算法的乳腺癌患者发病风险预测模型,结果表明随机森林算法的预测结果最好,其次是逻辑回归、SVC 和 DT 算法。

RF 排名前 10 位的重要发病特征分别为年龄、锻炼情况、流产次数、燥热、曾患乳腺炎、夜尿次数、婚姻情况、生育次数、学历、初潮时间。XGB 排名前 10 位的重要发病特征分别为年龄、曾患有乳腺炎、燥热、学历、流产次数、初潮时间、锻炼、乳房体检、易疲劳、外卖习惯。综合两种算法,排名第 1 位的重要发病特征均为年龄因素。另外乳腺癌发病与饮食情况也密切相关。有研究比较世界不同地区因各种健康和生活方式危险因素导致的 BC 死亡率和残疾调整寿命年(DALYs)的变化趋势,结果表明,1990—2017 年,发达和欠发达地区 BC 妇女的年龄标准化病死率和 DALYs 正在增加。危险因素酒精使用和吸烟的贡献显著。一项研究关于油炸食品摄入与女性乳腺癌相关风险相关性的 Meta 分析结果提示,高油炸食品摄入是乳腺癌的危险因素<sup>[9]</sup>。STECK 等<sup>[10]</sup>研究烧烤肉类摄入量与乳腺癌危险性的关系发现,绝经后组烧烤肉类摄入量增加与乳腺癌风险呈正相关,而绝经前组的相关性并不显著。

乳腺癌的发展是一个涉及多种细胞转化的复杂过

程,在世界范围内其预防治疗仍然具有很大的挑战性。早期诊断是预防乳腺癌发生发展的最佳方法之一。在一些发达国家采用早期预防措施直接影响乳腺癌患者的 5 年相对生存率,并维持在 80% 以上。近十年来,人们对乳腺癌预防方法的认识产生了巨大变化。其中乳腺癌干细胞的发现揭示了乳腺癌的发病和耐药机制,发现了许多与乳腺癌相关的基因。目前,人们有更多的药物选择用于乳腺癌的化学预防,而生物预防最近被开发用于改善患者的生活质量<sup>[11]</sup>。乳腺癌的发病机制多与雌激素水平、免疫密切相关。近年来,随着研究深入,多种与乳腺癌相关发病机制正在被逐一挖掘出来,乳腺癌发病也与多种因素相关,如肠道菌群、情志等多种诱因。通过操控肠道细菌达到肠道菌群的相对稳定来改善乳腺癌治疗效果,如改善饮食、控制抗生素的应用、益生菌的设计、粪菌移植<sup>[12]</sup>,极有可能成为乳腺癌的新型疗法。

两种算法中,燥热均为重要影响因素。《素问·阴阳应象大论》中指出“燥胜则干”,典型的表现为烦热口渴、多饮、多尿、便干、舌红苔黄而燥等属于燥热之象。乳腺癌患者多有燥热的表现,冷热均不耐受,既怕冷又怕热,具有寒热之特点,急躁易怒且有气机郁滞特点<sup>[13]</sup>。

本研究对调查问卷的设计将乳腺癌西医发病机制及中医病因病机充分结合,通过不同算法分析结果给出了乳腺癌发病因素重要指标,为乳腺癌的相关预防给出了有意义的预测分析。实验结果表明,与其他机器学习算法相比较,随机森林(RF)算法具有精准度高、模型易懂、训练高效等优越性。

4 展望

医疗大数据是一项多学科、多领域的研究,存在一定难度。由于医疗数据获取的困难性,使得此类分析的研究难以进行。尽管我国在大数据利用方面已积累了一定经验,但数据浪费、利用率低、可使用性低等问题仍很突出。本次研究中,RF 算法和 XGB 算法结果显示,排名前 10 位的重要发病因素中并无乳腺癌家族史,然而在临床中,通过病史采集及基因检测,部分乳腺癌患者相关家族史对其发病密切相关。有研究表明,乳腺癌具有非常明显的家族倾向,乳腺癌家族史可增加女性患乳腺癌的风险,且有一级亲属越多女性乳腺癌患病风险越高<sup>[14]</sup>。本研究可能是数据本身的局限性对模型性能的影响。另外,本次研究为回顾性分析,病例资料均由患者提供,变量因素以调查问卷的形式收集,数据可靠性方面存疑,可能存在结果偏倚。受纳入数据量所限,后续还应收集更多的 BRCA 数据,进一步优化乳腺癌发病风险预测模型。

## 【参考文献】

- [1] LIANG Y, ZHANG H, SONG X, et al. Metastatic heterogeneity of breast cancer: molecular mechanism and potential therapeutic targets[J]. *Seminars in Cancer Biology*, 2020, 60: 14–27.
- [2] 郭科, 孔学军. 多模态超声成像联合 PRL、sB7-H3、ELR 对乳腺癌的诊断价值[J]. *实用癌症杂志*, 2021, 36(12): 2011–2014.
- [3] 肖怀远, 郝希山. 乳腺癌早诊早治关键技术的研究与应用[J]. *科技导报*, 2021, 39(7): 75–78.
- [4] FU F F, ZHANG D, HU L, et al. Association between 15 known or potential breast cancer susceptibility genes and breast cancer risks in Chinese women[J]. *Cancer Biology & Medicine*, 2022, 19(2): 253–262.
- [5] WIGGS A G, CHANDLER J K, AKTAS A, et al. The effects of diet and exercise on endogenous estrogens and subsequent breast cancer risk in postmenopausal women[J]. *Frontiers in Endocrinology*, 2021, 12: 732255.
- [6] 中国抗癌协会乳腺癌专业委员会. 中国抗癌协会乳腺癌诊治指南与规范(2015版)[J]. *中国癌症杂志*, 2015, 25(9): 692–754.
- [7] 江泽飞, 许凤锐. 乳腺癌精准治疗: 20年探索历程[J]. *中国实用外科杂志*, 2020, 40(1): 83–88.
- [8] 李家宁, 熊睿彬, 兰艳艳, 等. 因果机器学习的前沿进展综述[J]. *计算机研究与发展*, 2023, 60(1): 59–84.
- [9] 付攸缘, 黄丽梅, 王慧欣, 等. 油炸食品摄入与女性乳腺癌风险相关性的 meta 分析[J]. *实用预防医学*, 2021, 28(12): 1437–1442.
- [10] STECK S E, GAUDET M M, ENG S M, et al. Cooked meat and risk of breast cancer – lifetime versus recent dietary intake[J]. *Epidemiology (Cambridge, Mass.)*, 2007, 18(3): 373–382.
- [11] SUN Y S, ZHAO Z, YANG Z N, et al. Risk factors and preventions of breast cancer[J]. *International Journal of Biological Sciences*, 2017, 13(11): 1387–1397.
- [12] TENG N M Y, PRICE C A, MCKEE A M, et al. Exploring the impact of gut microbiota and diet on breast cancer risk and progression[J]. *International Journal of Cancer*, 2021, 149(3): 494–504.
- [13] 孔柄坛, 朱晓云, 黄毅君, 等. 论“邪伏阴分”思想在内伤病诊疗中的应用[J]. *中国中医基础医学杂志*, 2021, 27(8): 1304–1306.
- [14] COLLABORATIVE GROUP ON HORMONAL FACTORS IN BREAST CANCER, et al. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58209 women with breast cancer and 101986 women without the disease[J]. *Lancet (London, England)*, 2001, 358(9291): 1389–1399.

(收稿日期: 2022–11–12)

## Effect Evaluation of Different Machine Algorithms in Breast Cancer Risk Prediction Model

ZHOU Liang<sup>1</sup>, WANG Yue<sup>1✉</sup>, TIAN Lingjia<sup>2</sup>, LIU Lifang<sup>1</sup>, ZHANG Qin<sup>1</sup>, LIANG Lixin<sup>1</sup>

(1. The First Affiliated Hospital of Hunan University of Chinese Medicine, Changsha 410007, China;

2. Hunan University of Chinese Medicine, Changsha 410208, China)

**【Abstract】** Objective: To develop a prediction model for breast cancer risk by using different machine learning algorithms. Methods: The database of the patients with breast cancer in the First Affiliated Hospital of Hunan University of Chinese Medicine was used as the data source. According to the risk factors of breast cancer, the menarche time, abortion frequency, fertility condition, menstrual and breastfeeding, family history of breast cancer, rest time, eating habits and TCM syndrome feature were selected as modeling variables, and demographic characteristics, vital signs and disease examination were extracted. Six machine learning algorithms were used to develop the model, and the effects of different machine learning algorithms were evaluated in the prediction model. Results: It was found that the variables of age, dryness–heat, abortion times, mammitis, exercises and first period time were of significance to predict breast cancer by synthesizing two kinds of algorithm and predicting the importance of the two algorithm models. Algorithm of random forest had the best prediction, with the accuracy rate of 0.86, the AUC value of 0.89; the accuracy rate of XGboost and GBDT were both 0.85, and the AUC value was 0.85 as well; the accuracy rate of logistic regression and the AUC value were both 0.84; the accuracy rates of SVC and DT were 0.83 and 0.79, respectively; the AUC values were 0.82 and 0.71, respectively. The results of each algorithm showed that the accuracy and prediction result of random forest algorithm were better than those of any other single algorithm due to its feature in ensemble learning. Conclusion: The prediction model of breast cancer risk based on random forest algorithm is important for assisting clinicians to know the risk of breast cancer.

**【Key words】** Machine algorithm; Logistic regression; Random forest; DT; GBDT; XGboost; SVC; Breast cancer; Risk prediction