论著·临床研究

5 种机器学习算法在肾结石筛查诊断中的应用及比较*

庄锡伟[1]，黎志全[1]，刘建雷[1]，岳　赟[2]
（1. 佛山复星禅诚医院，广东 佛山 528000；2. 佛山科学技术学院医学院，广东 佛山 528000）

［摘　要］ **目的**　通过收集尿液骨桥蛋白（OPN）、pH、白细胞和结晶的实验室检测结果，构建机器学习（ML）模型，评估模型对肾结石的筛查诊断价值。**方法**　选择 2021 年 12 月至 2022 年 8 月佛山复星禅诚医院就诊的肾结石患者 88 例为研究对象（肾结石组），采用“金标准”确诊，合并 CT、超声检测和临床症状等手段辅助鉴定，同时纳入 88 例非结石患者作为对照组，按照 5 折交叉验证随机分为训练集和验证集。收集研究对象尿 OPN、pH、白细胞、结晶及临床诊断信息，构建逻辑回归（LR）、决策树（DT）、随机森林（RF）、支持向量机（SVM）和 AdaBoost 共 5 种诊断模型，验证集中评价其诊断效能。**结果**　采用二元 logistic 回归从研究对象尿 OPN、pH、白细胞、结晶 4 个指标中筛选出白细胞、结晶 2 个诊断效率较高的指标，构建 ML 模型。验证集中各模型的诊断性能如下：（1）LR，精确度 0.882 4，召回率 0.340 9，F1 值 0.491 8，受试者工作特征曲线（ROC 曲线）下面积（AUC）为 0.722 4；（2）DT，精确度 0.662 2，召回率 0.556 8，F1 值 0.604 9，AUC 为 0.679 4；（3）RF，精确度 0.666 7，召回率 0.613 6，F1 值 0.639 1，AUC 0.700 2；（4）SVM，精确度 0.822 2，召回率 0.420 5，F1 值 0.556 4，AUC 为 0.727 1；（5）AdaBoost，精确度 0.729 7，召回率 0.613 6，F1 值 0.666 7，AUC 为 0.715 5，其中 AdaBoost 的肾结石诊断效能最佳。**结论**　收集尿 OPN、pH、白细胞和结晶检测结果，构建基于 ML 并具有可解释性的鉴别诊断模型，对肾结石的筛查诊断有一定的临床应用意义。

［关键词］ 肾结石；　实验室结果；　机器学习；　筛查诊断
DOI：10.3969/j.issn.1009-5519.2023.12.013　　**中图法分类号**：R319
文章编号：1009-5519（2023）12-2045-04　　**文献标识码**：A

Application and comparison of five machine learning algorithms in screening and diagnosis of renal calculi*

ZHUANG Xiwei[1], *LI Zhiquan*[1], *LIU Jianlei*[1], *YUE Yun*[2]
（1. *Foshan Fuxing Chancheng Hospital, Foshan, Guangdong* 528000, *China*;
2. *School of Medicine, Foshan University, Foshan, Guangdong* 528000, *China*）

［**Abstract**］ **Objective**　To construct a machine learning（ML）model by collecting the laboratory results of urine osteopontin（OPN）, pH, leukocyte and crystallization, and evaluate the diagnostic value of the model for kidney calculi. **Methods**　From December 2021 to August 2022, a total of 88 patients with kidney calculi who were treated in Foshan Fuxing Chancheng Hospital were selected as the study objects. The patients were diagnosed by "gold standard" and identified with CT, ultrasound detection and clinical symptoms, and 88 non-calculi patients were included as the control group. They were randomly divided into the training set and the verification set according to the five-fold cross-validation. The results of urine OPN, pH, white blood cells, crystallization test and clinical diagnosis information of the subjects were collected, and construct five diagnostic models, including logical regression（LR）, decision tree（DT）, random forest（RF）, support vector machine（SVM）and AdaBoost, to evaluate their diagnostic effectiveness in the verification set. **Results**　The binary logistic regression analysis was used to construct a ML model by screening two indicators with high diagnostic efficiency, namely, leukocyte exclusion and crystallization, from the four indicators of urine OPN, pH, leukocyte and crystallization. The diagnostic performance of each model in the validation set is: （1）LR, accuracy rate 0.882 4, recall rate 0.340 9, F1 value 0.498 7, area under the ROC curve（AUC）0.722 4; （2）DT, accuracy rate 0.662 2, recall rate 0.556 8, F1 value 0.604 9, AUC 0.679 4; （3）RF, accuracy rate 0.666 7, recall rate 0.613 6, F1 value 0.639 1, AUC 0.700 2; （4）SVM, accuracy rate 0.822 2, recall rate 0.420 5, F1 value 0.556 4, AUC

* **基金项目**：广东省佛山市自筹经费类科技计划项目（1920001001021）。
作者简介：庄锡伟（1983—），副主任技师，主要从事检验项目诊断价值方面的研究。

0.727 1; (5) AdaBoost, accuracy rate 0.729 7, recall rate 0.613 6, F1 value 0.666 7, and AUC 0.715 5, among which AdaBoost had the best diagnostic efficiency for kidney calculi. **Conclusion** Urine OPN, pH, leukocyte and crystal detection results are collected to construct an interpretable differential diagnosis model based on ML, which has certain clinical significance for the screening and diagnosis of kidney calculus.

[**Key words**] Kidney calculus; Laboratory results; Machine learning; Screening and diagnosis

肾结石是广东地区最常见的泌尿系疾病，其临床表现主要为肾区疼痛、血尿。综合结石性质、大小、位置、形态等因素，临床治疗手段主要有手术取石、体外碎石、药物排石等[1]。目前，肾结石的诊断“金标准”为影像学检查，不适用于早期诊断。因此，在肾结石的形成早期，及时发现并进行干预具有很好的临床价值。通常，疾病发生早期，实验室生化及代谢指标变化早于影像学变化，而目前单独的实验室诊断指标如尿常规用于筛查肾结石发病风险或评价治疗效果的灵敏度、特异度不高，因此，寻找构建肾结石临床筛查诊断的模型及方法成为临床研究的一个热点。近年来，人工智能技术在临床领域的应用发展迅猛，通过临床信息的挖掘，构建疾病筛查诊断模型，辅助临床进行疾病筛查或诊断[2]。本研究拟通过收集肾结石及非肾结石患者尿液骨桥蛋白（OPN）及尿常规中pH、白细胞、结晶的检验信息，构建基于机器学习（ML）的肾结石筛查诊断模型，并比较各模型对肾结石的筛查诊断价值，提供一种新的肾结石筛查诊断方法，以期实现肾结石筛查诊断，早期干预治疗肾结石，降低外科手术取石及体外碎石的治疗率。

1　资料与方法

1.1　资料

1.1.1　一般资料　选择 2021 年 12 月至 2022 年 8 月本院收治的肾结石及非肾结石患者，采用 CT、超声等“金标准”确诊肾结石，最终纳入研究的肾结石组患者 88 例，并选择同期非肾结石患者 88 例为对照组。采用患者检测的剩余样本进行尿液 OPN、pH、白细胞、结晶检测，方案得到医院伦理委员会批准同意。最终将纳入研究的 176 例患者信息按照 5 折交叉验证随机分为训练集和验证集。

1.1.2　肾结石诊断及鉴别诊断　（1）患者出现腰痛或血尿等临床症状；（2）患者泌尿系 B 超表现肾脏集合系统中强回声光团伴声影，伴或不伴肾盂、肾脏扩张或泌尿系 X 线平片肾脏显示致密影；（3）依据临床表现及影像学依据，与其他泌尿系结石进行鉴别诊断（表 1），最终临床诊断为肾结石的患者。

表 1　常见泌尿系结石的鉴别诊断

疾病	临床表现	影像学
肾结石	腰腹部疼痛甚至绞痛，血尿、恶心、呕吐、腹胀等	B 超表现肾脏集合系统中强回声光团伴声影，伴或不伴肾盂、肾脏扩张或泌尿系 X 线平片肾脏显示致密影
输尿管结石	腰腹部疼痛、血尿	B 超下输尿管结石表现是输尿管内可以出现强回声，后方伴明显的声影，同时伴结石以上的输尿管扩张。CT 表现输尿管走行区发现高密度的结节影
膀胱结石	下腹部疼痛、排尿困难、血尿	超声检查膀胱腔内强回声并有明显的声影，CT 平扫图像上表现为块状高密度灶，具有移动性

1.1.3　分组依据　（1）肾结石组：泌尿系 B 超或泌尿系 X 线平片诊断及医生诊断为肾结石的患者。（2）对照组：泌尿系 B 超或泌尿系平片没有结石诊断，同时，医生临床诊断没有肾结石的患者或体检人群。

1.2　方法　收集 2 组患者的尿液 OPN、pH、白细胞、结晶检测信息。

1.2.1　样本采集　（1）尿 OPN 检测样本：采集 10 mL 无稳定剂的尿液样品，5 000 r/min 离心 15 min，吸取上清液 1 mL，移入 1.5 mL 离心管，－80 ℃保存。（2）pH、白细胞和结晶尿液样本：采集患者尿液 10 mL，分别采用尿干化学分析仪、尿沉渣分析仪、显微镜等仪器对新鲜尿液进行 pH、白细胞、结晶的测定。

1.2.2　样本检测方法　（1）尿 OPN 检测：采用 CUSABIO 商品化试剂盒（ELISA 方法）检测，严格按照试剂盒的要求进行操作，读取 450 nm 波长的吸光度，通过标准曲线计算尿 OPN 的含量。（2）尿 pH、白细胞和结晶检测：尿干化学分析分析采用深圳美桥干化学分析仪检测；尿沉渣检测采用贝克曼 IQ-200 尿沉渣分析仪检测；尿结晶镜检，将尿液 1 500 r/min 离心 5 min，去上清，采用显微镜检查。尿 pH、白细胞和结晶操作过程均在标本采集后 2 h 内完成。

1.3　统计学处理　采用 SPSS25.0 统计软件进行数据处理，将肾结石组设定为 1，对照组设定为 0。对尿

OPN、pH、白细胞采用秩和检验，对结晶采用 χ^2 检验，进行单因素分析。采用二元 logistic 回归分进行指标筛选。各种假设检验中 $P<0.05$ 为差异有统计学意义。将筛选出的有效指标作为输入变量在训练集中构建逻辑回归(LR)、决策树(DT)、随机森林(RF)、支持向量机(SVM)和 AdaBoost 共 5 种诊断模型。验证集中采用精确率、召回率、F1 值、受试者工作特征曲线(ROC 曲线)及曲线下面积(AUC)评价模型的诊断性能，通过交叉验证计算出各模型 AUC 值对应的 95%可信区间(95%*CI*)。使用 F1 值作为本研究模型对比的最终指标。

2　结　　果

2.1　单因素分析结果　收集尿 OPN、pH、白细胞、结晶 4 个实验室检查指标。单因素分析结果显示，尿白细胞、结晶 2 个指标在肾结石组和对照组中比较，差异有统计学意义($P<0.05$)。见表 2。

表 2　尿 OPN、pH、白细胞和结晶单因素分析

指标	肾结石组($n=88$)	对照组($n=88$)	U/χ^2	P
尿 OPN(ng/mL)	616.145(1 894.747～14 029.677)	10 888.260(2 271.665～18 622.130)	4 446.0	0.090
尿 pH	6.000(6.000～6.500)	6.000(6.000～6.500)	4 087.5	0.507
尿白细胞(个/μL)	66.000(13.000～466.250)	7.000(0～22.250)	1 871.0	<0.001
尿结晶[n(%)]	79(89.8)	1(1.1)	5.195	0.023

2.2　二元 logistic 回归分析　将指标进行 logistic 回归分析，结果显示，白细胞、结晶 2 个指标在肾结石组和对照组比较，差异有统计学意义($P<0.05$)。见表 3。

表 3　尿 OPN、pH、白细胞和结晶的二元逻辑回归分析

指标	回归系数	P	OR	95%*CI*
尿 pH	−0.056	0.843	0.946	0.035～33.239
尿白细胞(个/μL)	0.002	0.005	1.002	1.001～1.003
尿结晶(阳性=1,阴性=0)	2.219	0.039	9.197	1.116～75.775
尿 OPN(ng/mL)	0	0.261	1.000	1.000～1.000

注：*OR* 表示比值比。

2.3　肾结石诊断模型建立及 ROC 曲线分析　将二元 logistic 回归分析中筛选出的尿白细胞、结晶 2 个指标作为输入变量，构建基于 ML 的肾结石筛查诊断模型，并在验证集中对模型诊断性能进行评价，结果显示，AdaBoost F1 值最高，为 0.666 7，见表 4。比较 5 种基于 ML 肾结石筛查诊断模型的 ROC 曲线，分析 AUC，结果显示，SVM 的 AUC 最大，为 0.727 1，见表 4。各指标权重值见表 5。

表 4　5 种基于 ML 的肾结石筛查诊断模型的诊断效能

ML 模型	精确度	召回率	F1 值	AUC
LR	0.882 4	0.340 9	0.491 8	0.722 4
DT	0.662 2	0.556 8	0.604 9	0.679 4
RF	0.666 7	0.613 6	0.639 1	0.700 2
SVM	0.822 2	0.420 5	0.556 4	0.727 1
AdaBoost	0.729 7	0.613 6	0.666 7	0.715 5

表 5　各指标权重值

ML 模型	尿白细胞权重	尿结晶权重
LR	1	0.491 4
DT	1	0.061 0
RF	1	0.062 1
SVM	1	0.161 8
AdaBoost	1	0.763 0

3　讨　　论

肾结石是泌尿系统的常见病、多发病，患病率高达 5%～10%。肾结石易复发，10 年内复发率高达 50%，并发症多，饮食习惯、肥胖程度、饮酒状况、饮水量、睡眠时间、早餐规律、焦虑压力为其发病的危险因素[3]。肾结石与遗传、代谢、环境等多种因素相关，同时存在地方、性别等差异[4-5]。目前，尚无肾结石诊断特别是早期诊断的实验室特异性生物标志物及筛查诊断模型，因此，寻找肾结石新的实验室筛查诊断指标、构建新的诊断筛查模型，在临床上具有很好的临床意义及应用前景。

OPN 是一种存在于细胞外基质中具有分泌性的磷酸化糖蛋白，广泛存在于肾脏组织及尿液等组织、体液中，其相对分子质量约为 44×10^3[6]。有实验研究表明，尿 OPN 通过抑制草酸钙结晶成核、聚集，减轻肾小管上皮细胞的氧化损伤，抑制结石的形成[7-8]，有希望成为结石筛查的实验室标志物。同时研究发现，肾结石与泌尿系感染密切相关[9]，通过尿白细胞可反映感染的严重程度，而细菌的感染可能影响尿液的 pH。尿结晶显微镜检查可通过结晶种类、形状、镜下数量作为肾结石诊断的有力佐证。

近年来，人工智能在医学等各个领域及日常生活中广泛应用，逐步改变人们的工作生产模式及生活方

式。ML 是人工智能的一大分支，其核心是源于大数据自动挖掘学习模式，通过建立学习模式对相关的问题输出决策[10]。临床上，疾病筛查诊断需要多学科、多指标进行综合判断，特别是给疾病的早期筛查诊断，因此，决定了临床医生必须具备丰富的知识储备和较高的综合判断能力，需要长时间的经验积累和知识更新。20 世纪 50 年代末，LEDLEY 等[11]第 1 次将数学模型引入临床医学，提出可以使用数学和推理模型协助进行疾病诊断。人工智能技术具有出色的处理大数据、分析挖掘复杂信息的能力，通过 ML 建立疾病筛查诊断的模型可以在特定疾病的诊断中显示较强的能力，为临床疾病诊断，特别是给疾病多学科诊断等应用带来新的契机，如近期肆虐全球的新型冠状病毒感染、肝硬化、肝癌等[12-13]，但由于临床用户难以理解模型的原理及决策过程，为其在临床应用也带来一定的挑战。ML 通过构建高维及多模态生物医学数据，采用复杂、自动和目标算法进行自动计算，挖掘生物医学大数据中隐含的诊断模型，实现疾病的智能筛查诊断。ML 主要算法包含人工神经网络（ANN）、DT、SVM、AdaBoost、集成学习（EL）等，通过采用不同算法对构建模型的诊断效率进行比较，优化算法及策略，建立疾病筛查诊断的最优模型。AdaBoost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器（弱分类器），根据弱分类器的错误率分配不同的权重参数，最后累加加权的预测结果作为输出。把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器），具有较高的检测速率，且不易出现过适应现象，广泛应用于乳腺癌疾病的临床诊断研究[14]。

本研究通过收集实验室尿 OPN、pH、白细胞、结晶 4 个检测指标信息，构建 5 种基于 ML 的肾结石筛查诊断模型。结果综合分析显示，5 种模型中 AdaBoost 的诊断效能最佳，其 AUC 为 0.715 5，F1 值为 0.666 7，诊断效能较高，可为临床肾结石的筛查诊断提供一种新的思路及方法。

本研究存在一定的局限性：(1)研究收集的样本量较少，限制了 ML 模型的构建，可能影响研究结果的准确性；(2)样本收集的实验室信息较少，如缺乏收集其他肾功能相关实验室检测指标、患者疾病分期、是否为复发患者等信息，诊断模型的代表性可能有所欠缺。下一步，作者将进一步加大纳入样本数量及实验室其他相关检验信息的收集，构建更具代表性的基于 ML 的肾结石筛查诊断模型。

参考文献

[1] 韩冬，张万生. 肾结石微创技术治疗的进展综述[J]. 中国医药指南，2016，14(10)：40.

[2] 张千，方丽华，王庆玮，等. 基于机器学习的疾病诊断模型研究[J]. 计算机与数学工程，2020，48(7)：1705-1709.

[3] 郭禹封，郭亚明，楚甜甜，等. 多种生活因素对肾结石发病的流行病学调查[J]. 现代医学与健康研究，2018，2(1)：135-137.

[4] 王冠怡，李胜，李刚，等. 高钙尿性肾结石相关遗传性疾病研究进展[J]. 中华泌尿外科杂志，2022，43(5)：393-396.

[5] 冯星亮. 肾结石与代谢综合征相关性的研究进展[J]. 国际泌尿系统杂，2022，40(3)：564-568.

[6] 覃诗婷，银锡靖，杨柯. 骨桥蛋白与草酸钙结石关系的研究进展[J]. 药学研究，2017，36(12)：729-732.

[7] 夏煜琦，程帆，袁润，等. 骨桥蛋白与肾结石形成机制的研究进展[J]. 中国医药导报，2018，15(1)：45-48.

[8] THURGOOD L A，SORENSEN E S，RYALL R L. Theeffect of intracrysta-lline and surface-bound osteopontin onthe degradation and dissolution of calcium oxalatedihydrate crystals in MDCKII cells[J]. Urol Res，2012，40(1)：1-15.

[9] SCALES C D JR，SMITH A C，HANLEY J M，et al. Prevalence of kidneystones in the United States[J]. Eur J Urol，2012，62(5)：160-165.

[10] 刘通. 智能优化极限学习机方法研究及在疾病诊断中的应用[D]. 长春：吉林大学计算机科学与技术学院，2020.

[11] LEDLEY R S，LUSTED L B. Reasoning foundations of medical diagnosis[J]. Science，1959，130(3366)：9-21.

[12] 王冰，栾锋，刘满仓，等. 基于线性判别式和支撑向量机的肾结石分类方法[J]. 兰州大学学报(自然科学版)，2006，42(2)：77-80.

[13] 刘巧，曾燕，王浩林，等. 机器学习方法对新型冠状病毒肺炎的诊断价值[J]. 中国医学影像学杂志，2021，29(4)：293-299.

[14] 叶琳，石胜源，罗铁清. AdaBoost 算法在乳腺癌疾病预测中的研究[J]. 计算机时代，2021，9(7)：61-64.

（收稿日期：2022-11-22　修回日期：2023-02-15）