昆明医科大学学报　2024, 45(3):99~105　　DOI: 10.12259/j.issn.2095-610X.S20240315
Journal of Kunming Medical University　　CN 53－1221/R

# 机器学习算法构建慢性肾脏病伴高血压或糖尿病的预测模型

曾慧娟[1)]，田　波[2)]，袁红伶[1)]，何　杰[1)]，李冠羲[1)]，茹国佳[1)]，许　敏[1)]，詹　东[3)]
（1)昆明医科大学第一附属医院肾脏内二科，云南 昆明　650032; 2)云南省第一人民医院科研科，云南 昆明　650034; 3)昆明医科大学基础医学院，云南 昆明　650500）

［**摘要**］**目的**　构建社区高血压、糖尿病患者中慢性肾脏病(CKD)早期预测模型。**方法**　群随机抽样昆明市 4 个城区的社区服务中心。对各中心建档居民分为疾病组($n$ = 1267)和对照组($n$ = 566)，疾病组居民患有高血压和或糖尿病，对照组未患。分别调查 2 组 CKD 患病情况并进行问卷调查、实验室检查和人浆细胞瘤变异易位基因(PVT1)基因中 5 个单核苷酸多态位点等检测。Logistics 回归筛选有统计学意义的危险因素纳入机器学习模型构建。算法模型包括支持向量机(SVM)、随机森林模型(RF)，朴素贝叶斯(NB)模型和人工神经网络(ANN)，并对比评价 4 个模型的效能和准确性进行比较分析。**结果**　筛选出 13 项具有统计学意义的指标($P$ < 0.05)，包括年龄、疾病类型、民族、血尿素氮、血肌酐、eGFR、PAM13 量表分数、睡眠质量调查、熬夜情况、PVT1 基因单核苷酸多态位点 rs11993333 及 rs2720659。基于危险指标建立机器学习算法模型。ANN 模型的准确度达 94.6%、特异性为 66.67%、Kappa 值为 0.7294、ROC 和 PRC 曲线下面积(0.9418 和 0.9261)均高于其它 3 种模型；RF 模型敏感性最高位 100%。**结论**　机器学习算法构建的 CKD 早期诊断模型在社区高血压或糖尿病患者中有较好的预测效能。尤其 ANN 模型各项性能优于其它。

［**关键词**］慢性肾脏病；机器学习；预测模型；高血压；糖尿病
［**中图分类号**］ R319；R692；R34 ［**文献标志码**］ A ［**文章编号**］ 2095－610X(2024)03－0099－07

# Predictive Modeling of Chronic Kidney Disease with Hypertension or Diabetes Based on Machine Learning Algorithms

ZENG Huijuan[1)], TIAN Bo[2)], YUAN Hongling[1)], HE Jie[1)], LI Guanxi[1)], RU Guojia[1)], XU Min[1)], ZHAN Dong[3)]
（1) *Dept. of Nephrology*, 1*st Affiliated Hospital of Kunming Medical University*; 2) *Dept. Scientific Research*, *Yunnan* 1*st People's Hospital*; 3) *School of Basic Medical Sciences*, *Kunming Medical University*, *Kunming Yunnan* 650500, *China*）

［**Abstract**］ Objective　To build the early predictive model for chronic kidney disease (CKD) in hypertension and diabetes patients in the community. Methods　The CKD patients were recruited from 4 health care centers in 4 urban areas in Kunming. The control group was residents without hypertension and diabetes ($n$ = 1267). The disease group was residents with hypertension and/or diabetes ($n$ = 566). The questionnaire survey, physical examination, laboratory testing, and 5 SNPs gene types in the PVT1 gene. The risk factors, which were filtered with logistics regression, were used to build predictive models. Four machine learning algorithms were built: support vector machine (SVM), random forest (RF), Naïve Bayes (NB), and artificial neural network (ANN) models. Results　Thirteen indicators included in the final diagnostic model: age, disease type, ethnicity, blood

［收稿日期］ 2023－12－13
［基金项目］ 云南省教育厅科学研究基金资助项目(2022J0268)；昆明医科大学第一附属医院博士科研基金资助项目(2021BS018)
［作者简介］ 曾慧娟(1985～)，女，云南大理人，医学博士，主治医师，主要从事肾脏内科学研究工作。
［通信作者］ 詹东，E-mail：zhandong@kmmu.edu.cn

urea nitrogen，creatinine，eGFR from MDRD，ACR，eGFR from EPI2009，PAM13 score，sleep quality survey，staying-up late，PVT1 SNP rs11993333 and rs2720659. The accuracy，specificity，Kappa value，AUC of ROC，and PRC of ANN are greater than those of the other 3 models. The sensitivity of RF is the highest among 4 types of machine learning. **Conclusions** The ANN predictive model has a good ability of efficiency and classification to predict CKD with hypertension and/or diabetes patients in the community.

［**Key words**］Chronic kidney disease；Machine learning；Predictive modeling；Hypertension；Diabetes

慢性肾脏病（chronic kidney disease，CKD）早期发现和诊断较为困难，随疾病不断进展，最后成为终末期肾脏病而使肾功能衰竭[1]。若能早期识别、发现 CKD 并准确诊断，便可及早干预和治疗[2]。全球约 5%～7% 的人口患有中等程度 CKD，其主要病因为糖尿病和高血压，尤其发展中国家、贫困地区、少数民族人群中慢性肾病发病率更高[3-4]。此外，CKD 患者治疗费用昂贵，给家庭带来沉重经济压力、给医疗保障带来沉重社会负担，已成为一个世界性公共健康问题[5]。然而，CKD 的诊断基于回顾性数据，起病隐匿、症状不明显，难以早期发现，延迟治疗和干预会增加肾衰竭可能[6]。本研究结合社区问卷调查、基本资料、实验室检查、PVT1 基因多态性等多领域交叉，采用 4 种机器学习算法构建 CKD 预测模型，辅助医生、患者及家人早识别，为评估提供参照、为诊断提供参考。

## 1 对象与方法

### 1.1 一般资料

随机抽样 4 个昆明市区域中的 1 个社区卫生服务中心建档居民。于 2019 年 11 月至 2023 年 11 间招募 CKD 患者 256 例和健康志愿者 1577 例。CKD 纳入标准：（1）CKD 诊断符合《慢性肾脏病早期筛查、诊断及防治指南》[7]；（2）患有血尿、蛋白尿、水肿、高血压或肾功能异常等临床表现伴有肾小球滤过率或肾组织学异常、肾脏影像学异常，病程持续 3 个月以上；（3）年满 18 周岁及以上居民；（4）诊断明确的 2 型糖尿病患者，且至少 6 个月以上或已建立慢性病健康管理档案；（5）诊断明确的高血压病患者，且至少 6 个月以上或已建立慢性病健康管理档案。排除标准：（1）患有其他系统疾病或脏器功能异常者；（2）患有恶性肿瘤者；（3）精神疾病患者；（4）妊娠及哺乳期女性患者。本研究获得昆明医科大学第一附属医院医学伦理委员会批准［（2022）伦审 L 第 264 号］，研究人员严格遵照《赫尔辛基宣言》实施。

### 1.2 研究方法

（1）问卷调查：内容包括一般人口学资料，13 条目患者积极度量表测量（PAM13）[8]，居民个人生活习惯调查（饮食、睡眠、烟酒摄入等）。调查员培训后上岗，调查员与患者一对一完成问卷。（2）体格检查：调查员行身高、体重、腰围、臀围等测量。（3）实验室检查：收集志愿者尿液行尿常规、肾功能、随机尿微量白蛋白测定等检查；抽取外周血提取 DNA 进行人浆细胞瘤变异易位基因（PVT1）基因进行单核苷酸多态位点（rs1499368、rs1121947/rs2608030、rs11993333、rs2720659 和 rs2720660）检测。（4）构建预测模型：采用 Logistic 回归对变量进行筛选。变量被随机分为训练集和测试集，分别占全体数据 2/3 和 1/3，用于建立预测模型和评价预测模型。归一化处理，二分类变量取值 0 或 1，计量资料变量值取值范围为 0～1 之间。分类变量因素不存在赋值 0，存在赋值 1。使用 R 软件工具包（e1071，caret，nnet 和 Neural NetTools）构建基于支持向量机（SVM）、随机森林（RF）、朴素贝叶斯（N B）和人工神经网络（ANN）算法的 CKD 预测模型。依据构建的 CKD 预测模型计算结果对检测集数据进行对比分析，评价指标包括灵敏度、特异度、准确率、原错率、Kappa 系数、阳性预测值、阴性预测值等。Kappa 系数用于评价模型预测值和真实值间的一致性。若 Kappa 系数 ≤0.2 则认为一致性极低；若 0.2< Kappa 系数 ≤0.4 则认为一致性一般；若 0.4< Kappa 系数 ≤0.6 则认为一致性中等；若 0.6< Kappa 系数≤0.8 则认为一致性较高；若>0.8 则认为几乎完全一致。采用受试者工作特征曲线（receiver operating characteristic curve，ROC）和精确召回曲线（precision recall curves，PRC）比较各模型性能优劣，并计算各模型曲线下面积（area under curve，AUC）。若 AUC 小于 0.5 说明模型不具有预测价值；若 AUC 值位于 0.7～0.5 区间时说明模型真实程度较差；若 AUC 值位于 0.9～0.7 区间时说明模型真实程度较好；总之，AUC 值越接近 1.0，模型预测真实程度越高，见图 1。

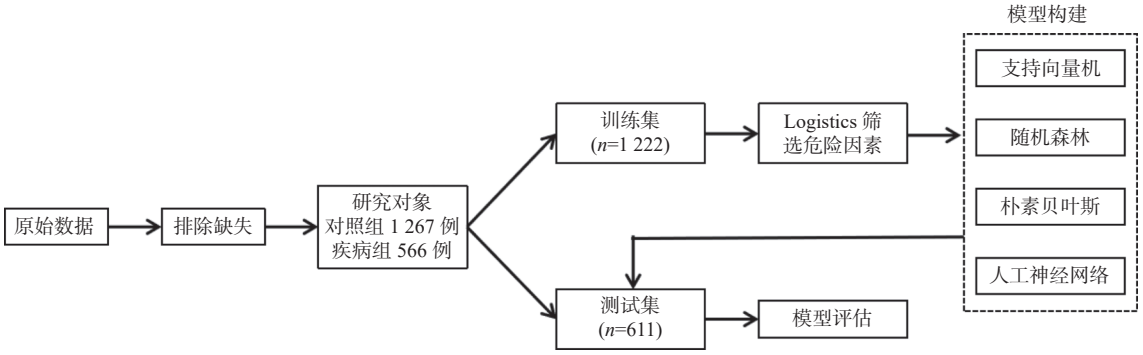


**图 1　预测模型构建流程图**

**Fig. 1　Flowchart of predictive modeling**

### 1.3　统计学处理

采用 R 软件(版本 4.1.3)处理数据。符合正态分布且方差齐的计量资料，行 Student's T 检验；不符合正态分布或方差不齐的计量资料则采用 Wilcox 检验。计数资料比较用卡方检验。检验水准设置为 $\alpha = 0.05$，且双尾设置。$P < 0.05$ 认为差异具有统计学意义。

## 2　结果

### 2.1　基线数据特征

对照组共有 1267 人纳入研究，平均年龄(65.90 ± 9.01)岁；疾病组患单纯高血压 344 例，患单纯糖尿病 96 例，同时患高血压和糖尿病者 126 例，平均年龄(65.67 ± 9.77)岁，2 组间差异无统计学意义($P = 0.314$)。对照组和疾病组男女比例分别为 45.30%、54.70% 和 41.67%、58.33%，差异无统计学意义($P = 0.699$)。疾病组患 CKD 率为 30.41% 显著高于对照组 7.58%，差异具有统计学意义($P < 0.0001$)，见表 1。对比训练集($n = 1222$)和测试集($n = 611$)，各项指标差异均无统计学意义，$P > 0.05$。

### 2.2　Logistic 回归筛选 CKD 风险指标

采用 Logisitc 回归分析发现 13 项指标对判定非 CKD 和 CKD 具有统计学意义，分别是年龄($P = 0.699$)、疾病类型(高血压、糖尿病、高血压合并糖尿病)($P < 0.0001$)、民族($P = 0.040$)、血尿素氮($P = 0.032$)、血肌酐($P = 0.015$)、MDRD 公式计算 eGFR≤60 mL/(min · 1.73 $m^2$)($P = 0.007$)、ACR≥30 mg/g($P < 0.0001$)、EPI2009 肌酐方程公式计算 eGFR≤60 mL/(min · 1.73 $m^2$)($P = 0.017$)、PAM13 量表分数($P = 0.001$)、睡眠质量调查表($P = 0.016$)、熬夜情况($P = 0.012$)、PVT1 基因 SNP 位点 rs11993333($P = 0.026$)和 rs2720659($P = 0.012$)，见图 2。

### 2.3　模型的建立和评估

13 项指标纳入机械学习算法，用于构建模型。PVT1 基因 SNP 位点 rs11993333 非优势基因型 TC 和 TT，位点 rs2720659 非优势基因型 AG 和 GG。

SVM 算法建立模型的准确率为 86.25%(95%*CI*：83.26% ~ 88.88%)小于原错率 87.23%，差异无统计学意义($P = 0.7863$)。该模型的 Kappa 值为 0.081，该模型预测值与真实值间一致性极低，模型预测精度极差。同时，SVM 模型灵敏度为 97.75%，而特异度仅为 7.69%。阳性和阴性预测值分别为 95.29% 和 33.33%。SVM 模型中 ROC 和 PRC 的 AUC 分别为 0.8957 和 0.7139 均大于 0.70，SVM 模型的真实度和精确度较高，见图 3A 和图 3B。

RF 算法建立模型准确率为 88.54%(95%*CI*：85.75% ~ 90.96%)小于原错率 87.23%，差异无统计

**表 1　研究对象分组数据分析 [*n*(%)]**

**Tab. 1　Base line data analysis between control group and disease group[*n*(%)]**

| 组别 | | *n* | non-CKD | CKD | $\chi^2$ | *P* |
|---|---|---|---|---|---|---|
| 对照组 | | 1267 | 1171(92.43) | 96(7.58) | | |
| 疾病组 | 高血压 | 344 | 258(75.00) | 86(25.00) | | |
| | 糖尿病 | 96 | 68 (70.83) | 28 (29.17) | 149.64 | <0.0001* |
| | 高血压合并糖尿病 | 126 | 80(63.49) | 46(36.51) | | |

*$P < 0.05$。

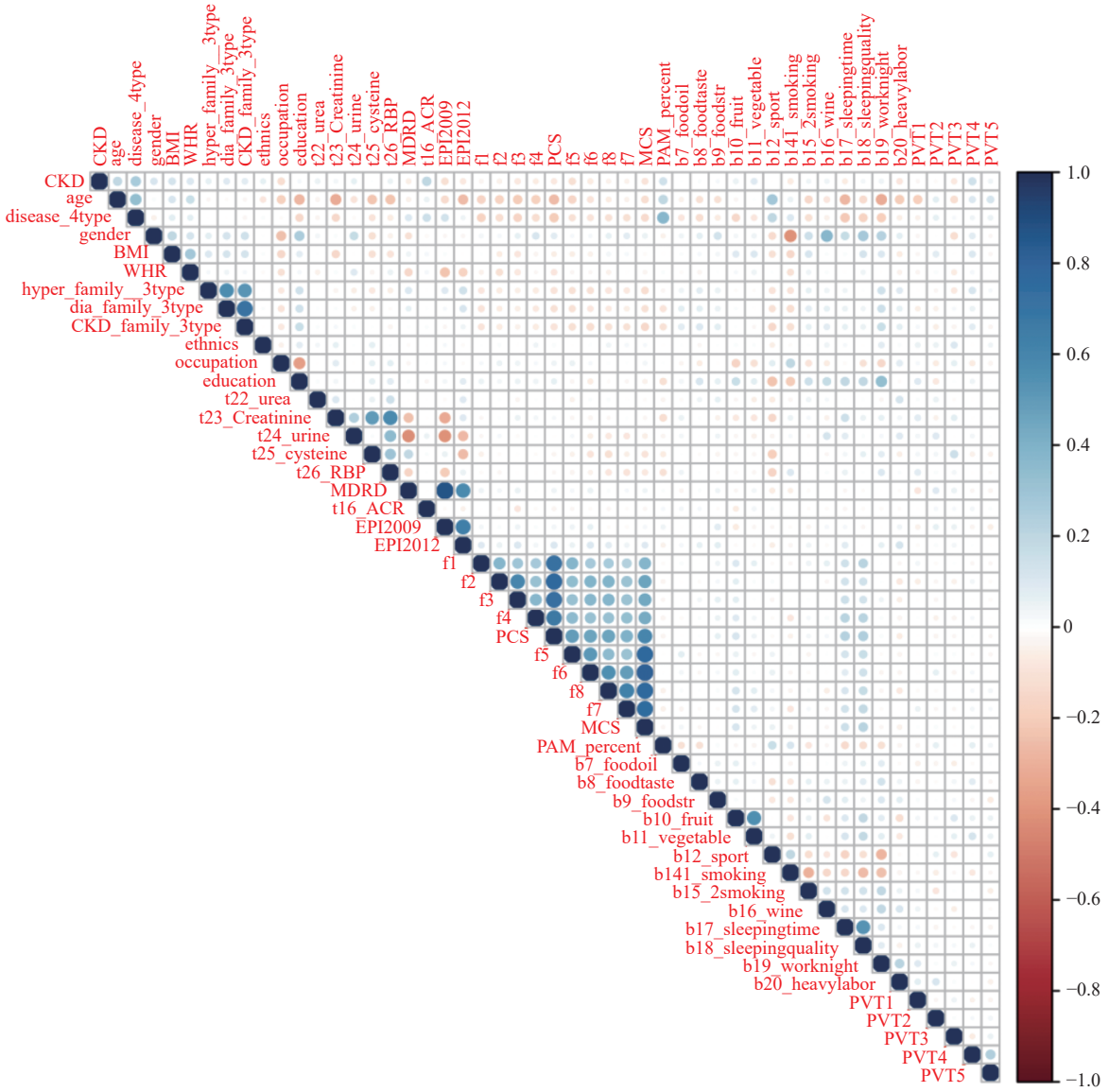


**图 2　Logisitc 回归分析热图**

**Fig. 2　Heatmap of Logistic regression**

学意义（$P=0.1823$）。该模型的 Kappa 值为 0.1662 < 0.2，预测值与真实值间一致性极低，模型预测精度极差。同时，RF 模型灵敏度为 100%，而特异度仅为 10.26%。阳性和阴性预测值分别为 88.29% 和 100%。RF 模型中，ROC 的 AUC 为 0.9210 大于 0.90，说明模型准确度较高，见图 4A，但 PRC 的 AUC 为 0.6502 小于 0.7，说明模型精确性较差，见图 4B。

NB 算法建立模型的准确率为 92.14%（95%*CI*：89.72% ~ 94.15%）大于原错率 87.23%，差异具有统计学意义（$P<0.0001$）。该模型的 Kappa 值为 0.6039，大于 0.41 而小于 0.60，预测值与真实值间一致性中等，模型预测精度尚可。同时，NB 模型灵敏度为 97.37%，而特异度仅为 56.41%。阳性和阴性预测值分别为 93.85% 和 79.86%。

NB 模型中，ROC 的 AUC 为 0.9369 大于 0.90，见图 5A，说明模型准确度较高，而且 PRC 的 AUC 为 0.7793 大于 0.7，说明模型精确性较也好，见图 5B。

ANN 模型输入层包含 15 个神经节点，隐藏层包含 11 个神经节点，输出层为目标疾病，包含 1 个神经节点，见图 6A。各变量对建立 ANN 模型的相对重要性不同，采用 Garson 算法评价各变量对 ANN 模型的相对重要性见图 6B。相对重要性贡献最大的是 ACR，其次为肌酐 creatinine，第三为 EPI2009。相对重要性贡献最小的为疾病类型。ANN 算法建立模型的准确率为 94.60%（95%*CI*：92.50% ~ 96.25%）大于原错率 87.23%，差异具有统计学意义（$P<0.0001$）。该模型的 Kappa 值为 0.7294，大于 0.60，预测值与真实值

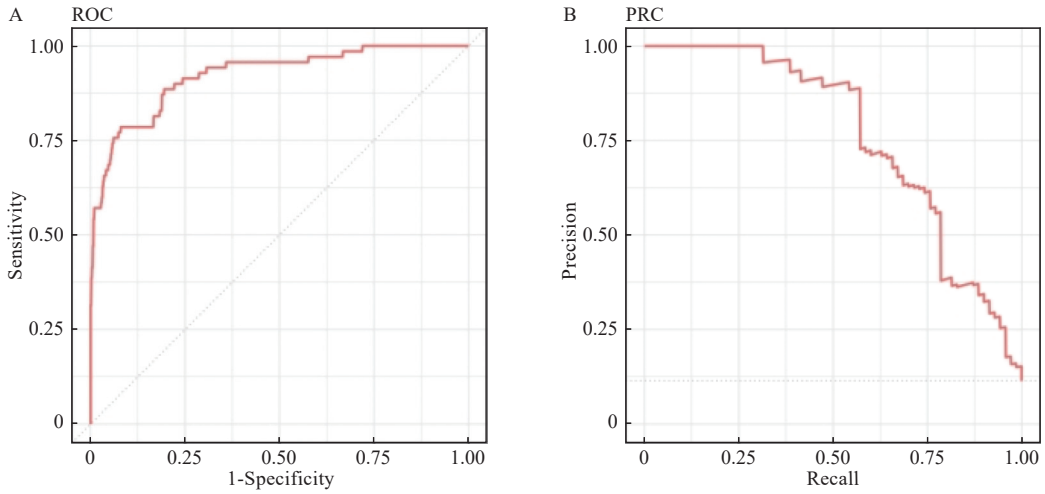


**图 3　支持向量机模型 ROC 和 PRC 的 AUC**

**Fig. 3　AUC of ROC and PRC in Support Vector Machine (SVM)**

A：支持向量机模型 ROC；B：支持向量机模型的 PRC。

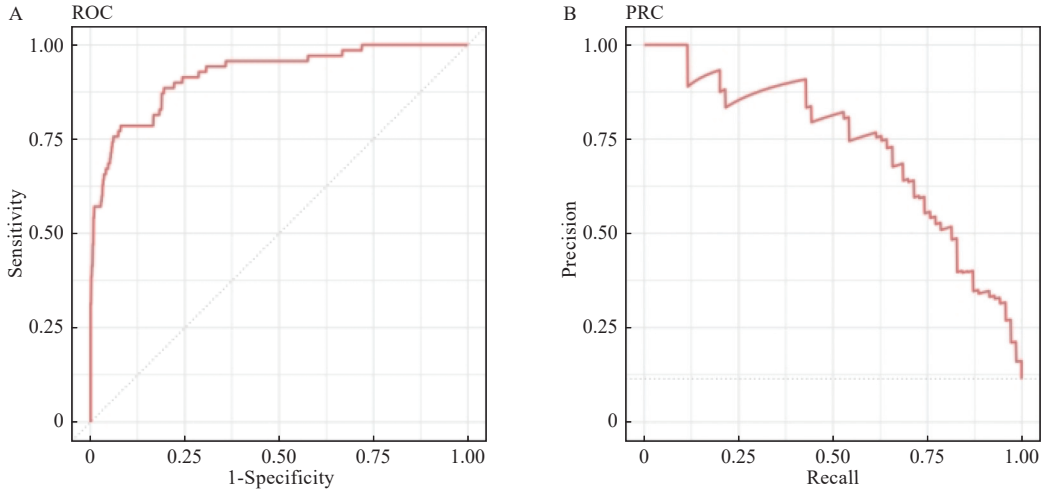


**图 4　自由森林模型 ROC 和 PRC 的 AUC**

**Fig. 4　AUC of ROC and PRC in Random Forest (RF)**

A：自由森林模型 ROC；B：自由森林的 PRC。

间一致性较高，模型预测精度较高。同时，ANN 模型灵敏度为 98.69%，而特异度仅为 66.67%。阳性和阴性预测值分别为 95.29% 和 88.14%。ANN 模型中，ROC 的 AUC 为 0.9418 大于 0.90，说明模型准确度较高，见图 6C，而且 PRC 的 AUC 为 0.9261 大于 0.9，说明模型精确性较也高，见图 6D。ANN 模型的准确率、特异性、Kappa 系数和 AUC 均高于 SVM 模型、RF 模型和 NB 模型，然而 ANN 模型灵敏度(98.69%)却低于 RF 灵敏度(100%)。

## 3　讨论

SVM、RF、NB 和 ANN 是目前较常见的机器学习算法用于 CKD 诊断。我国学者也采用机器学习建立 IgA 肾病的诊断模型，其准确率及可信度高[9]。国外学者采用 CKD 患者临床资料和症状建立 SVM 模型能够区分 CKD 患者和非 CKD 患者，准确率达到 99%，明显好于本研究的 SVM 模型准确率(86.25%)[10]。另 1 项研究通过对指标进行等级排序算法从 25 项指标中选取了 15 项建立 SVM 模型，可提高模型的准确率和 Kappa 值[11]。通过筛选算法选取 CKD 特征性指标建立 SVM 模型，使其准确率提高到 98.5%。由此 CKD 患者指标的选择、参数的设置可影响 SVM 模型准确率[12]。

本研究中选取了这 4 种算法对 CKD 患者临床资料、流行病学特征和分子基因 SNP 等建立模型。结果发现建立的 ANN 模型准确率高于其他 3 种模

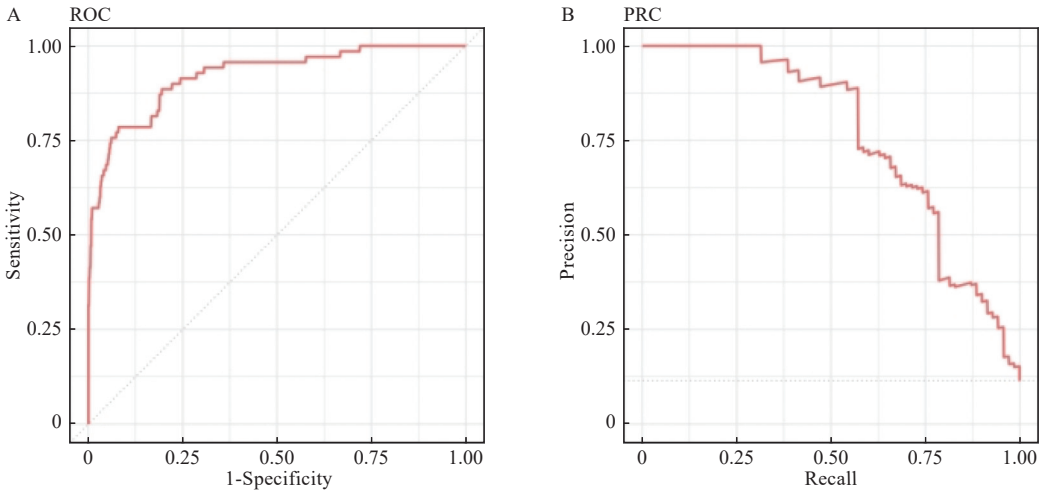


**图 5 朴素贝叶斯模型 ROC 和 PRC 的 AUC**

**Fig. 5 AUC of ROC and PRC in Naïve Bayes (NB)**

A：朴素贝叶斯模型 ROC；B：朴素贝叶斯模型的 PRC。

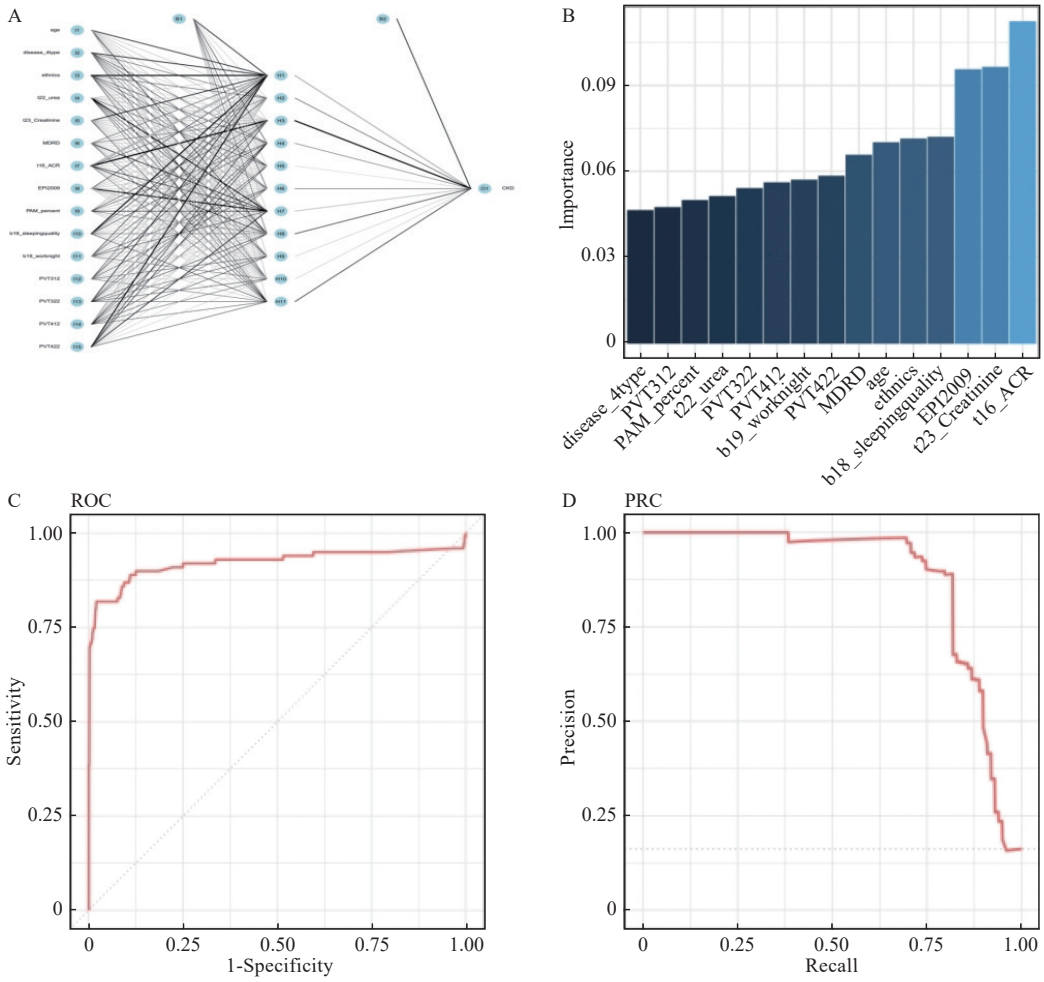


**图 6 人工神经网络评价**

**Fig. 6 Evaluation for artificial neuron net (ANN)**

A：人工神经网络模型各层图，I 代表输入层，H 代表隐藏层，O 代表输出层，B 代表偏倚矫正神经节点；B：ANN 模型各指标重要性评价；C：人工神经网络模型的 ROC；D：人工神经网络模型的 PRC。

型，达到 94.60%；同时 ANN 的 Kappa 值大于其他 3 种模型，ANN 预测值和真实值间一致性较高，精确度较好。基于前期调查和实验数据，通过 Logistic 回归分析，发现 13 个指标对模型建立起

主要作用，分别为年龄、疾病类型(高血压、糖尿病、同时患高血压糖尿病)、民族、血尿素氮 urine、血肌酐 creatinine、MDRD 公式计算 eGFR≤60 mL/(min·1.73 $m^2$)、ACR≥30 mg/g、EPI2009 肌酐方程公式计算 eGFR≤60 mL/(min·1.73 $m^2$)、PAM 量表分数、睡眠质量调查、熬夜情况、PVT1 基因 rs11993333 及 rs2720659($P<0.05$)。纳入指标建立 SVM、RF、NB 和 ANN 模型预测社区卫生服务中心中 CKD 患者。结果显示，在社区糖尿病高血压人群中，需要通过一些重要的因素早期筛查 CKD，这些指标主要为：患者血尿素氮及血肌酐检测，患者 eGFR 测定(主要通过 MDRD 公式计算及 EPI2009 肌酐方程公式)，筛查 ACR，PAM 量表分数、睡眠质量、熬夜情况；如能开展基因检测，可以检测 PVT1 基因 rs11993333 及 rs2720659[13-14]。另一方面，ANN 模型各项性能优于其他 3 种模型，ANN 模型的准确率和精确率较高、分类效果较好；但特异性欠佳，有待完善特征选择算法，剔除无关和冗余特征。建立社区 CKD 诊断 ANN 模型，目的是在社区卫生服务中心为社区医护提供方便实用的诊断预测模型，让社区医护、社区慢性病患者提高对 CKD 的认识及早期预警，逐步实现筛查、疾病追踪、诊断疾病及预测疾病预后的社区 CKD 管理模型[15]。下一步研究团队期望继续开发基于社区卫生服务中心 CKD 早期诊断小程序、APP 等，更方便模型的使用，逐步实现 CKD 诊断模型、风险预测模型、预测预后模型及评估 CKD 进展的一系列模型，最终实现早期发现 CKD、延缓 CKD 进展，让更多的人群不走进尿毒症、减轻医疗负担。

## [参考文献]


[1] 陈婷, 邓云蕾, 龚蓉. 终末期肾病合并感染的生物标志物检测意义及研究进展[J]. 临床肾脏病杂志, 2022, 22(3): 243-247.

[2] Santos M, Yin H, Steffick D, et al. Predictors of kidney function recovery among incident ESRD patients[J]. BMC Nephrol, 2021, 22(1): 142-153.

[3] Bikbov B, Purcell C, Levey A, et al. Global, regional, and national burden of chronic kidney disease, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017[J]. Lancet (London, England), 2020, 395(10225): 709-733.

[4] 王仕鸿, 令垚, 杨子华, 等. 基于时间序列模型的中国2020—2029年慢性肾病发病和患病情况预测研究[J]. 中国慢性病预防与控制, 2023, 31(11): 801-806.

[5] Zhang L, Wang F, Wang L, et al. Prevalence of chronic kidney disease in China: A cross-sectional survey[J]. Lancet (London, England), 2012, 379(9818): 815-822.

[6] 郑旭彤, 张曼, 秦朱珠, 等. 慢性肾病患者肾脏替代治疗辅助决策工具开发与验证研究的范围综述[J]. 中华护理教育, 2023, 20(4): 500-507.

[7] 高翔, 梅长林. 慢性肾脏病筛查诊断及防治指南[J]. 中国实用内科杂志, 2017, 37(1): 28-34.

[8] Zeng H, Jiang R, Zhou M, et al. Measuring patient activation in Chinese patients with hypertension and/or diabetes: Reliability and validity of the PAM13[J]. J Int Med Res, 2019, 47(12): 5967-5976.

[9] Chen T, Li X, Li Y, et al. Prediction and risk stratification of kidney outcomes in IgA nephropathy[J]. Am J Kidney Dis, 2019, 74(3): 300-309.

[10] Chen Z, Zhang X, Zhang Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models[J]. Int Urol Nephrol, 2016, 48(12): 2069-2675.

[11] Polat H, Danaei Mehr H, Cetin A. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods[J]. J Med Syst, 2017, 41(4): 55-66.

[12] Almansour A, Syed F, Khayat R, et al. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study[J]. Comput Biol Med, 2019, 109(6): 101-111.

[13] Wolfgram F, Garcia K, Evans G, et al. Association of albuminuria and estimated glomerular filtration rate with functional performance measures in older adults with chronic kidney disease[J]. Am J Nephrol, 2017, 45(2): 172-179.

[14] Zhang L, Zuo L, Xu G, et al. Community-based screening for chronic kidney disease among populations older than 40 years in Beijing[J]. Nephrology, dialysis, transplantation:official publication of the European Dialysis and Transplant Association - European Renal Association, 2007, 22(4): 1093-1099.

[15] Wouters J, O'donoghue J, Ritchie J, et al. Early chronic kidney disease: Diagnosis, management and models of care[J]. Nature Reviews Nephrology, 2015, 11(8): 491-502.