·临床研究·

机器学习构建多基因模型预测前列腺癌

陈志远，杨　瑞，刘修恒
（武汉大学人民医院泌尿外科，湖北武汉　430060）

Construction of a multigene predictive model of prostate cancer based on machine learning

CHEN Zhi-yuan，YANG Rui，LIU Xiu-heng
（Department of Urology，Renmin Hospital of Wuhan University，Wuhan 430060，China）

ABSTRACT：Objective　To construct a multigene model to predict prostate cancer based on machine learning. Methods　The RNA sequencing data of prostate cancer and normal prostate tissues were downloaded. The data were filtered and the differentially expressed genes were analyzed. The key genes were selected and the model was established and verified. The performance of decision tree，random forest，k-nearest neighbor（KNN），logistic regression and support vector machine under the default parameters were tested. The model with higher test efficiency were chosen for parameter adjustment and optimization. Results　Among the five models，random forest had the highest test efficiency，followed by decision tree. The accuracy of the optimized random forest model was 94%，and the area under the ROC was 0.94. Conclusion　The machine learning model based on gene expression data can predict prostate cancer.

KEY WORDS：prostate cancer；transcriptome；machine learning；random forest；predictive model

摘要：**目的**　基于基因表达数据，通过机器学习的方法构建模型鉴别前列腺癌。**方法**　下载前列腺癌和前列腺正常组织的 RNA 测序数据，进行数据过滤并分析差异表达基因，选择关键基因、建立模型并验证模型效能。验证决策树、随机森林、KNN 近邻、逻辑回归和支持向量机这 5 个模型在默认参数下的性能并选取具有较高检验效能的模型进行优化。**结果**　在 5 个模型中随机森林的检验效能最高，决策树次之。优化之后的随机森林模型鉴别前列腺癌的准确度为 94%，受试者工作（ROC）曲线下面积为 0.94。**结论**　通过基因表达数据构建机器学习模型能够较好地预测前列腺癌。

关键词：前列腺癌；转录组；机器学习；随机森林；预测模型
中图分类号：R737.25　　**文献标志码**：A　　**DOI**：10.3969/j.issn.1009-8291.2020.07.004

前列腺癌是中老年男性多发的一种恶性肿瘤。美国男性患者中，2018 年前列腺癌新发患者占到所有新发癌症患者的 19%，成为危害男性健康第一位的肿瘤[1]。随着近年来我国平均寿命的延长、生活习惯的改变和诊断水平的提高，前列腺癌在我国的发病率呈现逐年上升的趋势[2]。直肠指诊、前列腺特异性抗原（prostate specific antigen，PSA）、前列腺彩超广泛应用于前列腺癌的筛查，但均有一定的局限性。前列腺磁共振检查，特别是多参数前列腺磁共振检查能够较好地检测前列腺癌，但也有检查时间长、价格昂贵等问题[3]。为了更好地诊断前列腺癌，部分研究者尝试通过机器学习构建模型来预测前列腺癌[4]。目前在医学领域应用较多的机器学习算法有支持向量机、随机森林、朴素贝叶斯和神经网络等。随着测序技术的发展，TCGA 和 GTEx 等项目产生了大量的生物学数据，包含肿瘤和正常组织的基因表达、蛋白表达和临床特征等，这些数据是构建模型的基础[5-6]。本研究通过机器学习对前列腺癌和正常的前列腺基因表达数据进行分析，旨在构建新的前列腺癌预测模型。

1　材料与方法

1.1　数据获取　TCGA 包含前列腺癌组织和癌旁组织的测序数据，GTEx 包含了正常前列腺组织的测序数据。WANG 等[7]通过去除两者的批间差，采用 RNA 序列的期望最大化（RNA-Seq by expectation maximization，RSEM）进行定量并将结果以每千个碱基的转录每百万映射读取的片段数（Fragments per Kilobase Million，FPKM）形式进行表示，这使人们可以直接使用两个数据库中的数据。我们下载转录本表达数据，并按照肿瘤组织和正常组织进行分组。肿瘤组包含样本 427 例，正常组包含样本 155 例。两组都包含了 19 047 个 mRNA 的表达数据。由于部分 mRNA 的表达量在大部分样本中的表达量很低，我们采用在一半样本中的表达量大于 1 作为规则进行过滤。

收稿日期：2020-02-17　　**修回日期**：2020-04-02
基金项目：国家自然科学基金面上项目（No. 81972408）；武汉市应用基础前沿项目（No. 2018060401011321）
通信作者：刘修恒，教授，主任医师. E-mail：drliuxiuheng@sina.com
作者简介：陈志远，副教授，副主任医师. 研究方向：前列腺癌. E-mail：420132602@qq.com

1.2 差异基因的获取和富集分析 首先计算前列腺癌和正常前列腺组织中具有差异表达的基因。我们采用 limma 包和 R 语言进行计算，将 $\log_2$FC 的绝对值取 1.5，FDR 为 0.05 作为阈值。计算得到了 1 033 个差异基因，其中 233 个在前列腺癌组织中上调，800 个在癌组织中下调。基因本体论（Gene Ontology，GO）数据库分别从基因功能、基因参与的生物途径和细胞中的定位对基因产物进行了描述，对差异基因进行 GO 富集分析可以了解差异基因在哪些生物学功能、参与的生物学途径和细胞定位出现了富集。对差异基因进行可视化和富集分析。

1.3 特征基因的选取 我们利用 sklearn 中单变量特征选择的方法从差异基因中选择特征基因。用 feature_selection 类的 selectKBest 命令选择前 20 个和分组信息具有高相关性的基因，并计算这 20 个基因之间的相关性。由于具有较高相关性的基因在建模时出现的多重共线性问题容易导致建模失败，我们排除了这 20 个基因中具有较高皮尔森相关系数的基因，最终纳入建模的有 7 个基因。随机森林可以判断每个特征对分类所作的贡献程度，我们使用这一方法来计算这 7 个基因对分类的重要性。

1.4 基本模型的构建 我们首先观察默认参数下模型的效能，纳入分析的模型有决策树、随机森林、KNN（K-nearest neighbor）近邻、逻辑回归和支持向量机。对于每一个模型，我们使用准确度（Accuracy）、精确度（Precision）、召回率（Recall）和 F1-score 来评估模型的效能。其中，准确度表示模型预测正确的正样本和负样本占所有样本的比例，精确度表示模型预测正确的正样本占所有预测正样本的比例，召回率表示模型预测正确的正样本占所有实际正样本的比例，F1-score 是精确度和召回率的调和平均值，可以较好地评估模型的整体效果。结果表明，决策树和随机森林显示出了较好的分类效果。我们选取决策树和随机森林模型进行进一步的优化调参。

1.5 决策树和随机森林模型的优化 剪枝策略对决策树模型的效能有重要的影响，控制决策树的深度是一种常用的剪枝策略。我们通过计算不同决策树深度下的模型准确度来寻找最佳深度。对随机森林而言，其包含的子评估器的个数对模型效能较为重要，我们同样计算了不同的评估器个数对模型效能的影响。学习曲线是一种判断模型训练状态的方法，它评估了训练集和验证集的拟合程度。我们计算了决策树和随机森林模型的学习曲线。

1.6 数据处理 本研究所有的数据处理均在个人电脑上进行，配置如下：CPU 为英特尔 i5，内存 8 G，差异基因的计算以 R 计算机语言完成，其余分析在 python 3.6 语言环境下进行，环境同时配有 scikit-learn 数据分析包和 numpy、pandas、matplotlib 等数据处理和可视化模块。$P<0.05$ 为差异有统计学意义。

2 结 果

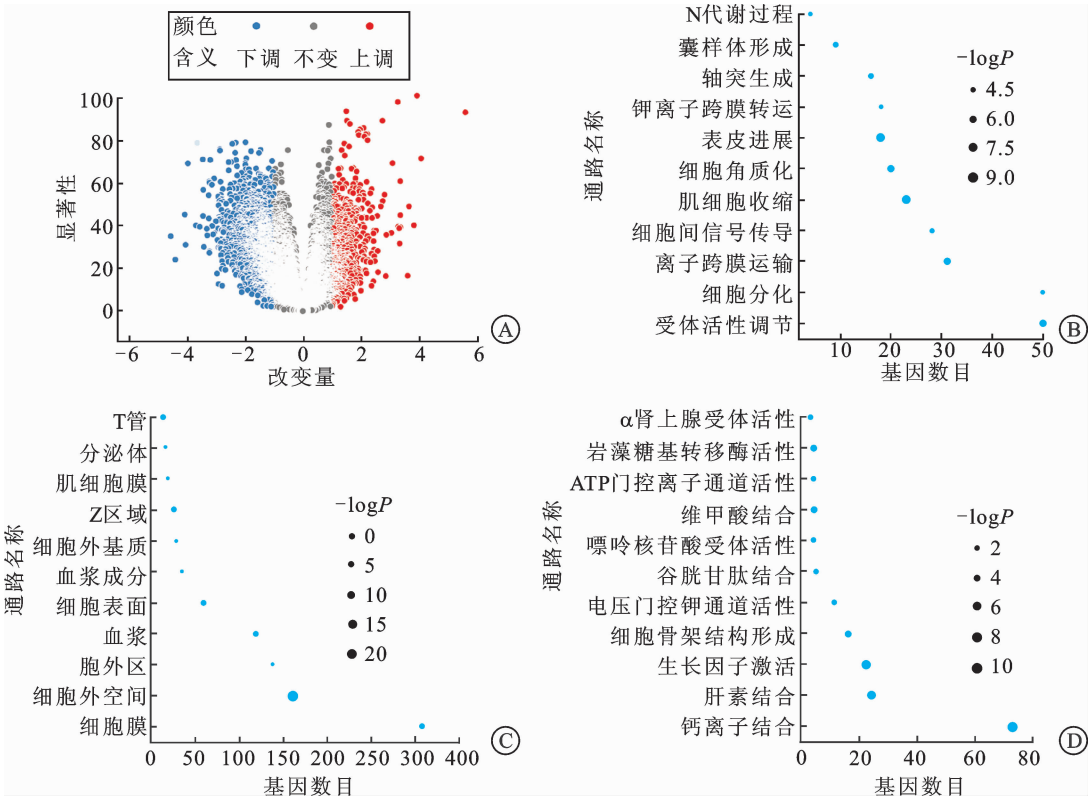
2.1 差异基因的可视化和富集分析 我们用火山图来表示差异基因。红色表示在肿瘤组织中上调的基因、有 233 个，蓝色表示在肿瘤组织中下调的基因、有 800 个。随后我们将这些差异表达基因进行基因富集分析。结果显示，在生物学过程这一分类中，差异基因在表皮细胞进展和受体激活调节等通路中出现了富集；在细胞组分这一分类中，差异基因在血浆成分、T 管和细胞表层通路中出现了富集；在分子功能这一分类中，差异基因在钙离子结合、肝素结合和生长因子激活通路中出现了富集（图 1）。

2.2 特征基因的选取结果 我们首先根据每个基因区分前列腺癌组和正常组的能力对基因进行了排序，并选择了前 20 个基因。如图 2A 所示，颜色的不同和深浅代表着不同的相关性。可以看到，这样挑选出来的 20 个基因中的部分基因具有较强的相互作用。我们过滤掉相关性在 0.5 之上的基因，最后得到了 7 个基因，分别是 KLK2、SPON2、PLA2G2A、MYLK、WFDC2、SEMG1 和 LTF，它们之间的相关性如图 2B 所示。这 7 个基因对模型分类的重要性排序如图 2C 所示，可以看到 KLK2 对分类的重要性较强。

2.3 基础模型的效能 首先将数据按照 7∶3分为训练集和验证集，随后计算不同的模型对这一分类任务的效果。特别是由于不同的核函数对支持向量机的影响较大，我们将每一个核函数单列。如表 1 所示，可见模型效能基于前两位的是随机森林和决策树，剩下的模型检验效能较为接近，其 F1-score 都在 0.8 到 0.9 之间。

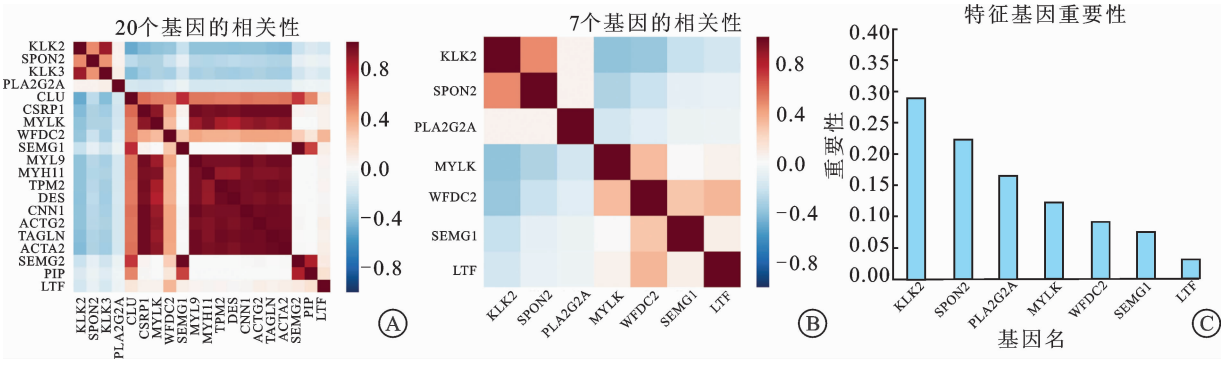
2.4 随机森林和决策树的优化和验证 鉴于随机森林和决策树模型的检验效能，对其进行进一步优化有可能达到最佳的效果。我们首先探究了决策树的深度对决策树的影响和随机森林中子分类器个数对随机森林的影响。从图 3A 中可以看出，当树的深度在 16 时决策树模型的准确度最高。此时模型识别前列腺癌的准确度为 0.941 4，比默认模型上升了 0.1 个百分点左右。当子评估器的个数取值为 21 的时候，随机森林模型的分类效果最高，此时的准确度为 0.948，相比默认参数上升了 0.07 个百分点（图 3B）。我们重新计算了优化之后的决策树和随机森林的受试者工作曲线（receiver operating characteristic，

ROC)，如图 3C、E 所示。决策树的 ROC 曲线下面积为 0.925 3，随机森林的 ROC 曲线下面积为 0.945 1，两者的学习曲线如图 3D、F 所示，训练集和验证集的评分较为接近，可见模型的训练达到了较好的效果。



A：差异基因火山图。横坐标表示两组基因表达量的的定量关系，纵坐标表示差异的显著性；B：差异基因在生物学过程上的富集结果；C：差异基因在细胞组分上的富集结果；D：差异基因在分子功能上的富集结果。

图 1　差异基因和富集分析图



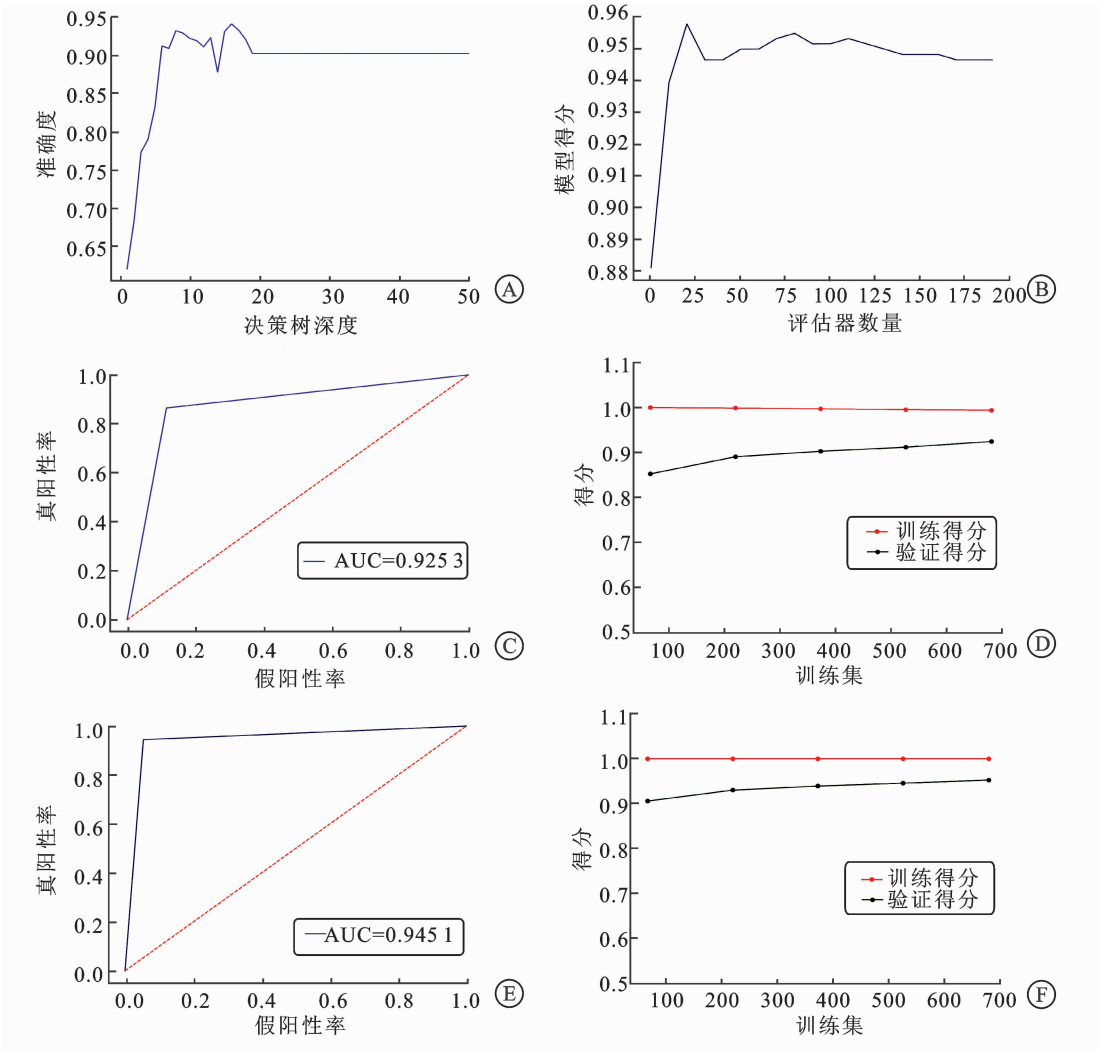
A：特征工程选取出来的 20 个基因；B：去掉高相关性之后的 7 个基因；C：7 个基因对建模的贡献程度。

图 2　特征基因的选取和重要性排序

表 1　不同模型的检验效能

模型	准确度	精确度	召回率	F1 值	模型	准确度	精确度	召回率	F1 值
决策树	0.937 5	0.931 4	0.931 1	0.931 2	支持向量机（径向基核函数）	0.859 3	0.814 6	0.936 3	0.871 2
随机森林	0.941 4	0.934 1	0.941 2	0.938 8					
KNN	0.898 4	0.890 1	0.903 0	0.896 5	支持向量机（sigmoid 核函数）	0.875 0	0.845 2	0.931 4	0.876 2
逻辑回归	0.882 8	0.840 4	0.935 2	0.885 2					
支持向量机（线性核函数）	0.878 9	0.847 8	0.927 5	0.885 8	支持向量机（二项式核函数）	0.855 4	0.900 2	0.785 3	0.838 8

KNN：(K-nearest neighbor)。



A：决策树深度对模型的影响；B：子评估器的数目对模型的影响；C：决策树的受试者工作曲线；D：决策树的学习曲线；E：随机森林的受试者工作曲线；F：随机森林的学习曲线。

图 3 随机森林和决策树的优化和验证

3 讨 论

前列腺癌的早期治疗有赖于早期诊断。诸多研究尝试通过各种生物标记物和模型来进行前列腺癌的诊断。李方龙等[8]对解放军总医院预测前列腺癌的 301 模型进行了单中心验证，证明了当 PSA 水平在 10～20 ng/mL 时该模型能够较为准确的诊断前列腺癌。张志昱等[9]对 PSA 及其衍生指标进行了研究，提示前列腺特异性抗原密度（prostate specific antigen density，PSAD）在 PSA 处于灰区时诊断效能较好，取 0.475 作为阈值时候的灵敏度为 72%，特异度为 61%。而前列腺特异性抗原密度比值（The rate of Prostate Specific Antigen Density，PSAMR）在灰区之外的诊断效能较好，当 0.135 作为阈值时灵敏度为 66%，特异度为 55%。机器学习在前列腺癌的诊断和治疗中显示了很好的前景。SAHRAN 等[10]采用基于径向基的支持向量机模型构建特征选择算法来进行前列腺癌 Gleason 评分的预测，达到了较好的效果。HUSSAIN 等[11]利用朴素贝叶斯、支持向量机和决策树来对前列腺磁共振图像所提取的特征进行建模，其得到的模型鉴别前列腺癌的准确度达到了 99.71%。在本研究中，我们通过机器学习的方法，基于 TCGA 和 GTEx 数据库中的 RNA 表达数据建立了数个前列腺癌诊断模型，其中最佳模型是基于决策树的随机森林模型。随机森林模型的子评估器的个数为 21 时，模型的性能达到最高，此时模型识别前列腺癌的准确度为 94%，ROC 曲线下面积为 0.94。和常用的 PSA、多参数磁共振等检测指标相比，本研究构建的模型在准确率和检验效能上有一定优势。

机器学习是人工智能的核心，主要研究如何使计算机可以通过数据来建立模型，并且利用建立好的模型来对新的输入进行预测。Sklearn 这一软件包是实现这一理论的重要工具，sklearn 依赖 python 计算机语言，包含了分类、回归、聚类、降维、模型选择和数据

预处理六大模块，可以轻松完成机器学习全流程任务[12]。在本研究中，我们通过sklearn中的特征选择方法纳入了7个基因进行建模。它们分别是KLK2、SPON2、PLA2G2A、MYLK、WFDC2、SEMG1和LTF。KLK2编码的蛋白质是一种高活性的类胰蛋白酶丝氨酸蛋白酶，这种蛋白主要在前列腺组织中表达，负责将前列腺前特异性抗原切割成酶活性形式。该基因在前列腺癌细胞中高度表达，可能是前列腺癌风险的预后指标[13]。SPON2也是一种蛋白质编码基因，与蛋白质代谢和ERK信号转导相关[14]。PLA2G2A的蛋白质是磷脂酶A2家族(PLA2)的成员，参与生物膜磷脂代谢的调节[15]。MYLK编码肌球蛋白轻链激酶，它是一种钙/钙调素依赖性酶，可以磷酸化肌球蛋白来促进肌球蛋白与肌动蛋白丝的相互作用[16]。WFDC2编码一种属于WFDC结构域家族的蛋白质。WFDC结构域(WAP-Signature motif)包含8个半胱氨酸，在蛋白质的核心形成4个二硫键。有研究显示，前列腺癌和正常前列腺组织中该基因启动子的甲基化有着显著差异[17-18]。SEMG1编码的蛋白质是精液中的主要蛋白质[19]，LTF编码的蛋白质是人体分泌物中主要的铁结合蛋白，具有抗菌活性，是非特异性免疫系统的重要组成部分[20]。这些基因在前列腺癌中的作用尚待进一步研究。

值得指出的是，机器学习技术和大数据息息相关，样本越多，训练的模型才能越准确。随着测序技术的进一步普及，用于训练和评估的数据集将进一步增多，并且随着机器学习算法的优化及提高，更多优秀的模型将会被开发出来，更多的患者也将会因此受益。

参考文献：

[1] SIEGEL RL, MILLER KD, JEMAL A. Cancer statistics, 2018[J]. CA Cancer J Clin, 2018, 68(1): 7-30.

[2] 顾秀英，郑荣寿，张思维，等. 2000—2014年中国肿瘤登记地区前列腺癌发病趋势及年龄变化分析[J]. 中华预防医学杂志，2018，52(6)：586-592.

[3] HAIDER MA, YAO X, LOBLAW A, et al. multiparametric magnetic resonance imaging in the diagnosis of prostate cancer: a systematic review[J]. Clin Oncol, 2016, 28(9): 550-567.

[4] GOLDENBERG SL, NIR G, SALCUDEAN SE. A new era: artificial intelligence and machine learning in prostate cancer[J]. Nature Reviews Urol, 2019, 16(7): 391-403.

[5] WANG Z, JENSEN MA, ZENKLUSEN JC. A Practical guide to the cancer genome atlas (TCGA)[J]. Methods Mol Biol, 2016, 1418: 111-141.

[6] CARITHERS LJ, MOORE HM. The Genotype-Tissue Expression (GTEx) project[J]. Biopreserv Biobank, 2015, 13(5): 307-308.

[7] WANG Q, ARMENIA J, ZHANG C, et al. Unifying cancer and normal RNA sequencing data from different sources[J]. Sci Data, 2018, 5(1): 180061.

[8] 李方龙，刘健，邱建宏，等. 前列腺特异抗原及其相关参数在前列腺癌诊断中的意义[J]. 现代泌尿外科杂志，2017，22(7)：534-540.

[9] 张志昱，张江磊，臧晋，等. 前列腺特异性抗原新参数在早期前列腺癌筛查中的作用[J]. 现代泌尿外科杂志，2019，24(10)：828-832.

[10] SAHRAN S, ALBASHISH D, ABDULLAH A, et al. Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading[J]. Artif Intell Med, 2018, 87: 78-90.

[11] HUSSAIN L, AHMED A, SAEED S, et al. Prostate cancer detection using Machine learning techniques by employing combination of features extracting strategies[J]. Cancer Biomark, 2018, 21(2): 393-413.

[12] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python[J]. J Mach Learn Res, 2011, 12: 2825-2830.

[13] KOHLI M, ROTHBERG PG, FENG C, et al. Exploratory study of a KLK2 polymorphism as a prognostic marker in prostate cancer[J]. Cancer Biomark, 2010, 7(2): 101-108.

[14] QIAN X, LI C, PANG B, et al. Spondin-2 (SPON2), a more prostate-cancer-specific diagnostic biomarker[J]. PLoS One, 2012, 7(5): e37225.

[15] HALPERN AL, KOHTZ PD, ROVE JY, et al. Inhibition of secretory phospholipase A2 IIa attenuates prostaglandin E2-induced invasiveness in lung adenocarcinoma[J]. Mol Cell Biochem, 2019, 456(1-2): 145-156.

[16] WALLACE SE, REGALADO ES, GONG L, et al. MYLK pathogenic variants aortic disease presentation, pregnancy risk, and characterization of pathogenic missense variants[J]. Genet Med, 2019, 21(1): 144-151.

[17] TOWNES CL, ALI A, GROSS N, et al. Prostate specific antigen enhances the innate defence of prostatic epithelium against Escherichia coli infection[J]. Prostate, 2013, 73(14): 1529-1537.

[18] KIM JH, DHANASEKARAN SM, PRENSNER JR, et al. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer[J]. Genome Res, 2011, 21(7): 1028-1041.

[19] YU Q, ZHOU Q, WEI Q, et al. SEMG1 may be the candidate gene for idiopathic asthenozoospermia[J]. Andrologia, 2014, 46(2): 158-166.

[20] WANG M, QIN M. Lack of association between LTF gene polymorphisms and different caries status in primary dentition[J]. Oral Dis, 2018, 24(8): 1545-1553.

（编辑　杨婉婉）