Statistics and Application 统计学与应用, 2025, 14(8), 85-96
Published Online August 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.148218

Hans汉斯

基于机器学习的前列腺癌致病基因筛选与疾病预测

王秉基[1*]，高　翔[2]

[1]中国海洋大学海德学院，山东　青岛
[2]中国海洋大学数学科学学院，山东　青岛

收稿日期：2025年7月8日；录用日期：2025年7月28日；发布日期：2025年8月11日

摘　要

为识别与前列腺癌相关的遗传特征，文章提出了一种集成式机器学习方法，用于筛选前列腺癌的关键基因，并深入分析这些靶基因的生物学意义，从而建立高效的疾病诊断预测模型。研究通过UCSC Xena数据库收集了151例前列腺癌组织和152例正常组织的转录组数据，并采用PCA等方法进行批次效应校正。通过差异表达分析筛选出了2586个上下调基因，并结合GO和KEGG富集分析，揭示了与前列腺癌相关的关键致病通路。进一步集成随机森林、LASSO回归和梯度提升机(GBM)三种机器学习算法进行基因二次筛选，最终确定了12个对前列腺癌具有重要影响的关键基因。基于这些基因，构建了8种前列腺癌诊断预测模型，采用混淆矩阵和ROC曲线对模型性能进行评估。结果显示，极限梯度提升(XGBoost)模型的准确度和AUC值分别达到了93%和97%，验证了该模型在前列腺癌诊断中的应用潜力。

关键词

前列腺癌，差异表达分析，机器学习，诊断预测模型

Screening and Disease Prediction of Prostate Cancer-Causing Genes Based on Machine Learning

Bingji Wang[1*], Xiang Gao[2]

[1]Haide College, Ocean University of China, Qingdao Shandong
[2]School of Mathematical Sciences, Ocean University of China, Qingdao Shandong

Received: Jul. 8th, 2025; accepted: Jul. 28th, 2025; published: Aug. 11th, 2025

[*]通讯作者。

文章引用: 王秉基, 高翔. 基于机器学习的前列腺癌致病基因筛选与疾病预测[J]. 统计学与应用, 2025, 14(8): 85-96.
DOI: 10.12677/sa.2025.148218

Abstract

To identify genetic features associated with prostate cancer, the article proposes an integrated machine learning approach for screening key genes for prostate cancer and analyzing the biological significance of these target genes in-depth to build an efficient predictive model for disease diagnosis. The study collected transcriptome data from 151 prostate cancer tissues and 152 normal tissues through the UCSC Xena database, and corrected for batch effects using PCA and other methods. A total of 2586 up- and down-regulated genes were screened by differential expression analysis and combined with GO and KEGG enrichment analysis to reveal the key pathogenic pathways associated with prostate cancer. A secondary screening was further integrated with three machine learning algorithms, namely Random Forest, LASSO regression and Gradient Boosting Machine (GBM), and 12 key genes with significant impact on prostate cancer were finally identified. Based on these genes, eight prostate cancer diagnosis prediction models were constructed, and the model performance was evaluated using confusion matrix and ROC curve. The results showed that the accuracy and AUC value of the Extreme Gradient Boosting (XGBoost) model reached 93% and 97%, respectively, which verified the potential application of the model in prostate cancer diagnosis.

Keywords

Prostate Cancer, Differential Expression Analysis, Machine Learning, Diagnostic Predictive Modeling

1. 引言

前列腺癌(Prostate Cancer)是一种严重危害男性健康的恶性肿瘤[1]。从流行病学数据来看，前列腺癌的发病率和病死率在全球范围内均处于较高水平[2]。前列腺癌的发生受多种因素的影响，包括遗传因素、年龄、种族、雄激素水平及生活方式等[3]。尽管近年来，前列腺特异性抗原筛查的广泛应用提高了早期诊断率，并促进了手术、放疗及内分泌治疗等治疗手段的发展，但由于肿瘤的异质性和耐药性的存在，部分患者仍可能进展为去势抵抗性前列腺癌，导致预后较差[4]。因此，探索前列腺癌的分子机制并寻找有效的生物标志物和治疗靶点对于提高患者生存率具有重要意义[5]。

近年来，越来越多的研究发现，前列腺癌的发生和进展与特定的差异表达基因(DEGs)密切相关，这些基因可能在肿瘤的发生、侵袭和转移过程中发挥重要作用[6]。高通量基因测序技术的快速发展，使研究者能够更全面地解析前列腺癌的基因表达谱，并为个性化治疗提供新的可能性。随着机器学习和生物信息学的融合应用，研究者们利用特征选择算法和分类模型对关键基因进行筛选，从而构建更精准的前列腺癌预测和分类模型[7]。这些研究不仅有助于前列腺癌的早期诊断，还为潜在的治疗策略提供了新的方向。

2. 数据来源与处理

本文所用数据均来源于公开的 UCSC Xena 数据库(https://xena.ucsc.edu/)。首先从 UCSC Xena 数据库中的 TCGA 板块选取了 203 例样本，其中分为 152 例前列腺癌样本和 51 例正常样本。这里为了弥补正

常样本的不足，又从 GTEX 板块中选取了 122 例样本，并将两数据集进行合并。将合并后的数据集进行数据预处理，这个过程中去除了基因表达量较低的基因以及基因表达异常的样本。再根据数据库特征，对样本进行分类，最终得到 151 个前列腺癌样本和 152 个正常样本。

3. 方法

3.1. 差异基因筛选

考虑到 TCGA 和 GTEX 数据库之间具有批次效应，本文首先对合并后的数据进行了 PCA 和数据归一化来去除批次效应，如图 1，展示了部分样本去除批次效应后的变化。接着利用 R 语言中的 Deseq2 软件包对去除批次效应后的数据进行了差异基因筛选。这里，本文设置域值($P < 0.05$, $|\log_2 FC| > 1$)来筛序差异表达基因，然后使用火山图对结果进行可视化。并通过 R 语言程序对 DEGs 进行了功能富集分析(GO 与 KEGG)，得到了与差异基因相关的生物学过程(BP)、细胞组分(CC)、分子功能(MF)以及相关的信号通路。

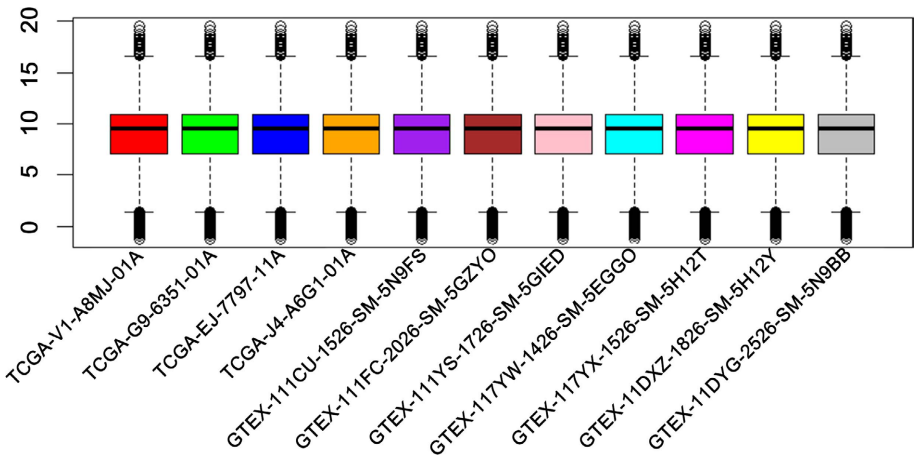


Figure 1. Distribution of samples after removal of batch effects
图 1. 去除批次效应后样本分布图

3.2. 机器学习筛选关键基因

在创建基因筛选模型时，首先通过差异表达分析(如 DESeq2)进行初步筛选。然而，单独依赖 DESeq2 包进行差异表达分析可能无法充分捕捉到所有关键基因，因此研究进一步应用了三种机器学习中的特征选择算法：随机森林(RF)、LASSO 算法和 GBM (Gradient Boosting Machine)算法。这些算法通过不断训练和调整模型参数，有效地优化了筛选基因的范围，为后面建立预测模型提供了有效帮助。

3.2.1. 随机森林

随机森林(如图 2)是一种基于集成学习的算法，通过构建多个决策树并结合投票(分类)或平均(回归)结果，提高模型的准确性和鲁棒性。它具有良好的抗过拟合能力和容错性，能有效处理高维数据与变量间的非线性关系。在基因筛选和疾病分类中，随机森林不仅能识别出具有重要诊断价值的关键特征，还能提供特征的重要性排序，有助于发现潜在的生物标志物和构建稳定的预测模型。

3.2.2. LASSO 回归

LASSO (Least Absolute Shrinkage and Selection Operator)是一种改进的线性回归方法，通过引入 L1 正则化项，实现变量筛选和模型压缩。其目标函数为：

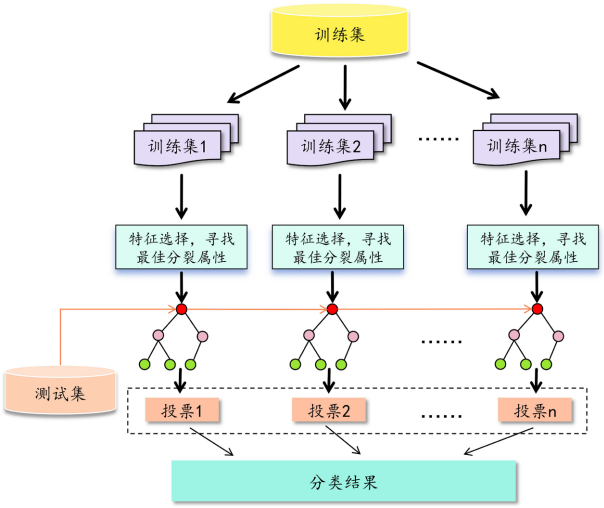


Figure 2. Random forest flowchart
图 2. 随机森林流程图

$$\hat{\beta} = \text{argmin}\left(\sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right) \tag{1}$$

其中，$X \in R^{n\times p}, \beta \in R^p, y \in R^n, \sum_{j=1}^{p} \left| \beta_j \right|$ 是 L1 正则化项，λ 是正则化强度控制参数。当 λ 较大时，LASSO 会将部分系数 $\left| \beta_j \right|$ 收缩至 0，从而实现特征的自动选择。这种机制使 LASSO 不仅可以降低模型复杂度、防止过拟合，还能提升模型的可解释性，尤其适用于高维数据。在基因筛选或疾病分类等生物信息学应用中，LASSO 能够从海量特征中识别出与疾病高度相关的核心基因，为生物标志物的发现和精准医学研究提供了有力支持。

3.2.3. GBM 算法

GBM (梯度提升机)算法是一种集成学习算法，通过逐步构建多个弱学习器(通常是决策树)，每一步利用前一轮的残差进行学习，逐步提升模型的预测精度。其核心思想是最小化损失函数，通过以下公式更新模型：

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2}$$

上式中，$F_m(x)$ 是第 m 轮的预测模型，γ_m 是学习率，$h_m(x)$ 是当前训练的基学习器。GBM 通过优化损失函数(如均方误差)来训练模型，在每轮迭代中减小预测误差。此外，GBM 能够计算特征的重要性，帮助识别关键特征或生物标志物，特别适用于基因筛选和疾病分类等任务。

3.3. 预测模型构建与评估

使用 R 语言程序构建了基于广义线性模型(GLM)、随机森林(RF)、支持向量机(SVM)、朴素贝叶斯(NBM)、K 近邻(KNN)、LASSO 回归(LASSO)、极致梯度提升(XGBoost)、多层感知机(MLP) 8 种方法的前列腺癌诊断预测模型，同时采用了十折交叉验证，以减少模型训练过程中因数据划分不同而导致的偏差，提升模型的泛化能力和稳健性。

在训练不同的分类器模型后，需要对其性能进行评估，研究主要采用了 ROC 曲线和 AUC 值来评估各分类模型的预测能力。通过绘制 ROC 曲线，可以观察模型在不同判别阈值下的灵敏度(Sensitivity)与特

异性(Specificity)变化，并通过 AUC 值定量评估模型的整体分类能力。AUC 值越大，说明模型在区分阳性和阴性样本方面的性能越优。

4. 结果

4.1. 基因筛选

本研究对 303 例样本进行了分析，包含 151 个 tumor 样本和 152 个 normal 样本。在进行数据预处理之后，我们使用“DESeq2”包对数据集进行分析，同时设置域值($P < 0.05$, $|\log_2 FC| > 1$)，最后共获得 16,833 个 差异表达基因，其中 1499 个差异表达基因下调，1087 个差异表达基因上调，DEGs 的结果显示在如下火山图和热图中(如图 3、图 4)。

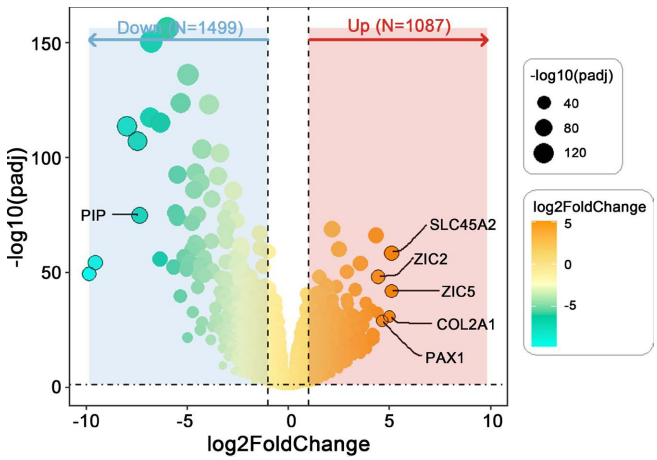
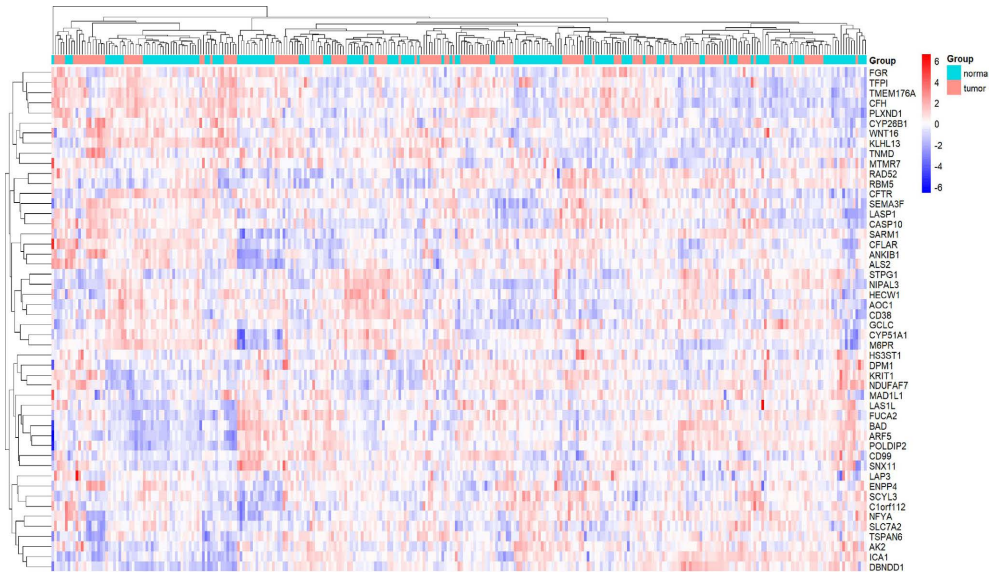


Figure 3. Differential gene volcano map
图 3. 差异基因火山图



注：方格中红色代表正相关，蓝色代表负相关。颜色越深，相关性越强。

Figure 4. Heat map of differential genes (top 50)
图 4. 差异基因(前 50 个)热图

4.2. 富集分析

我们通过 R 语言程序对 DEGs 进行了功能富集分析，筛选出了校正后概率 P < 0.05 的富集途径。从图中可以看出，基因本体论(GO)包含 869 条目：生物过程(BP)为 643 条，细胞组分(CC)为 89 条，分子功能(MF)为 137 条。我们按照特定顺序排列条目的 P 值，选择每个过程中的前条记录，并在图 5 中展示这 8 条记录。如图 5(A)~(C)，生物过程主要和肌肉收缩、细胞间黏附、嗜同性细胞粘附等有关，细胞组分主要和细胞外基质、离子通道复合物、肌浆等密切相关，分子功能主要和被动跨膜转运体活性、通道活性、阳离子通道活性等密切相关。

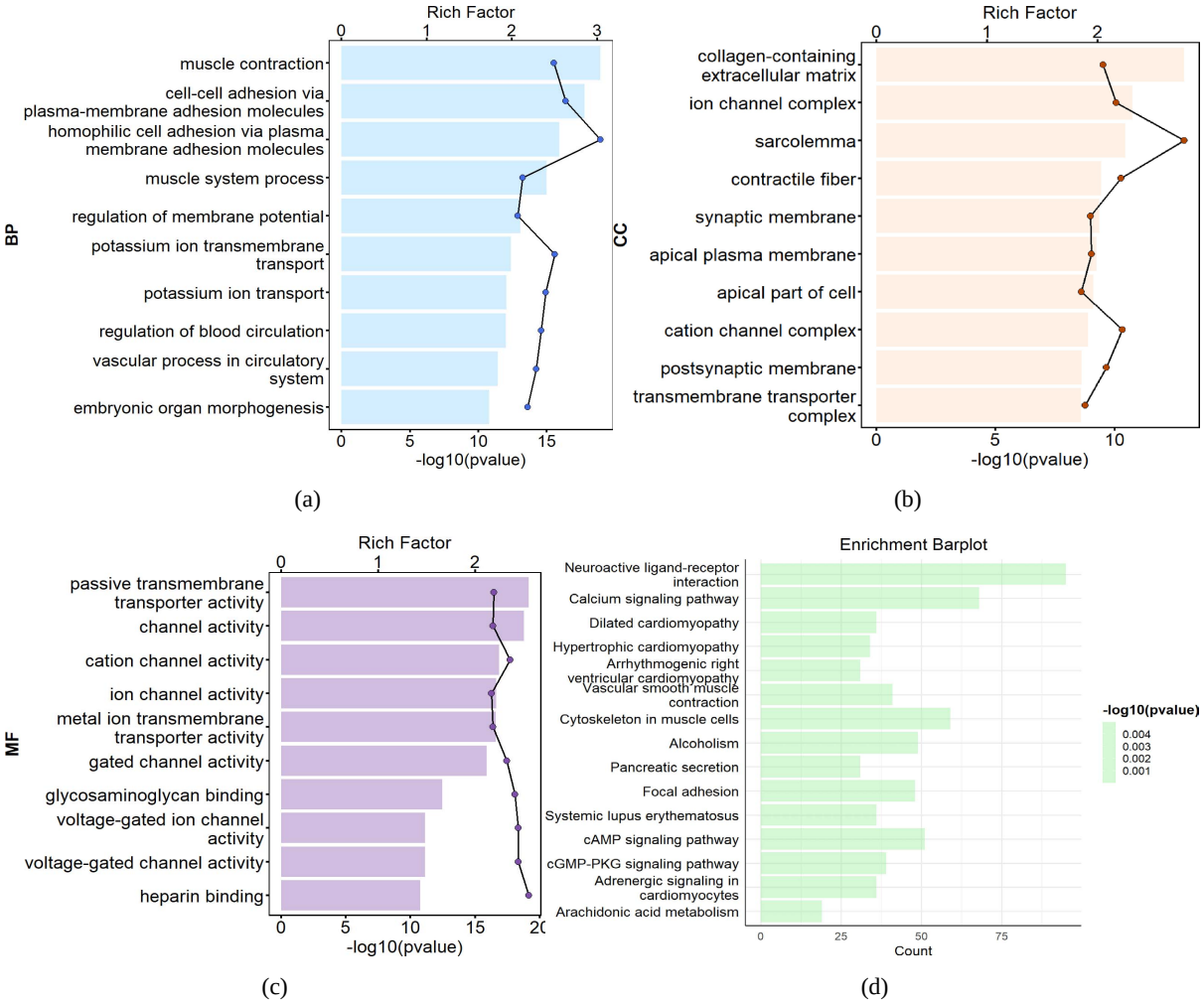


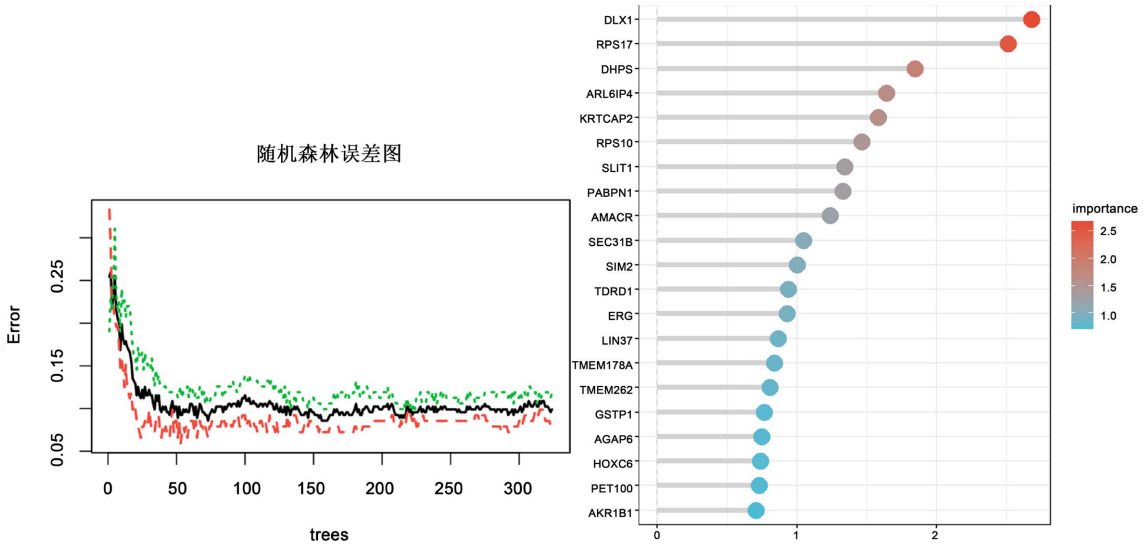
Figure 5. Enrichment analysis. (a) BP in GO enrichment analysis; (b) CC in GO enrichment analysis; (c) MF in GO enrichment analysis; (d) KEGG enrichment analysis

图 5. 富集分析。(a) GO 富集分析中的 BP 方面；(b) GO 富集分析中的 CC 方面；(c) GO 富集分析中的 MF 方面；(d) KEGG 富集分析

KEGG 富集结果则包含 335 个条目，关键差异基因主要在神经活性配体与受体的相互作用、钙信号通路、肌肉细胞的细胞骨架和 cAMP 信号通路等通路富集明显。如图 5(D)，我们按照校正后概率 P 的升序顺序展示了前 15 个通路。

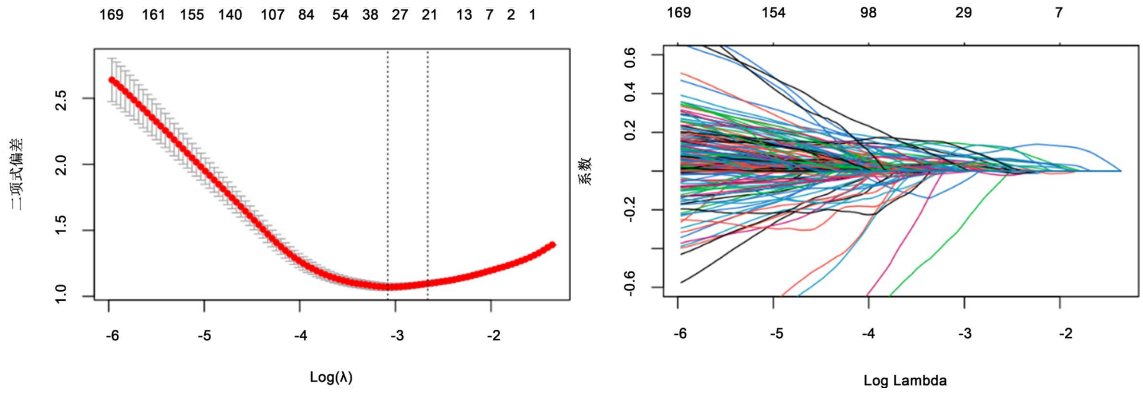
4.3. 关键基因筛选

在进行差异分析后，我们又分别使用了 RF、LASSO 回归和 GBM 算法重新对上面得到的 2586 个差异表达基因进行了筛选。随机森林通过生成多个决策树并计算每个基因的特征重要性，最终选出最具预测性的特征基因。在模型训练过程中，RF 的参数设置为 ntree = 1000，通过调整树的数量以优化模型的性能。如图 6(a)和图 6(b)，最终筛选出 21 个较为显著的基因；LASSO 算法则通过调整正则化参数来获得最佳的基因子集，从而提高模型的预测能力并减少过拟合。如图 7(a)和图 7(b)，通过选择 lambda.min = 0.04587769，筛选出 32 个关键基因；如图 8，GBM 算法通过十折交叉验证选择出评分最高的 13 个基因。如图 9，RF 与 LASSO 交集为 5 个标志基因，RF 与 GBM 交集为 8 个标志基因，LASSO 与 GBM 的交集为 3 个标志基因。这里为防止筛选过度，研究将上述三个交集合并，最终得到 12 个关键基因，分别为 DLX1、SLIT1、SIM2、TDRD1、HOXC6、RPS17、DHPS、ARL6IP4、RPS10、SEC31B、GSTP1、GOLM1。



(a) (b)

Figure 6. Random forest screening for biomarkers
图 6. 随机森林筛选生物标志物



(a) (b)

Figure 7. LASSO screening biomarkers
图 7. LASSO 筛选生物标志物

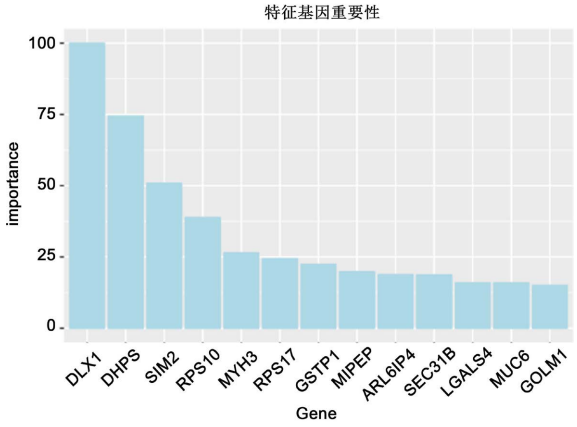


Figure 8. GBM screening biomarkers
图 8. GBM 筛选生物标志物

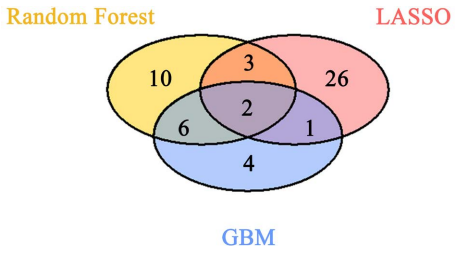


Figure 9. Wayne's map
图 9. 韦恩图

4.4. 前列腺癌预测模型构建

利用上文筛选出的 12 个关键标志基因，研究构建了前列腺癌早期预测模型。构建步骤如下：将来源于 UCSC Xena 数据库的数据划分为测试集和训练集，其中，选取了 $\frac{1}{10}$ 的数据作为测试集，用于验证模型的鲁棒性和泛化能力，剩下的数据则作为模型训练集，流程如图 10。

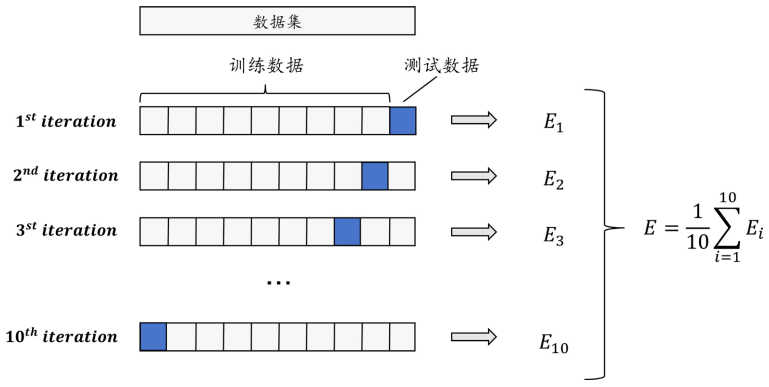


Figure 10. Flowchart of ten-fold cross-validation
图 10. 十折交叉验证流程图

采用十折交叉验证法对训练集构建了基于广义线性模型(GLM)、随机森林(RF)、支持向量机(SVM)、朴素贝叶斯(NBM)、K 近邻(KNN)、LASSO 回归(LASSO)、极限梯度提升(XGBoost)、多层感知机(MLP)

8 种方法的前列腺癌诊断预测模型。接下来，我们将使用独立的测试集验证八种前列腺癌诊断预测模型的性能。同时，我们将通过计算混淆矩阵(如图 11)和 AUC 值来评估这些模型的准确性和可靠性。根据图 12 可知，测试集在 8 种模型下表现效果良好，AUC 均在 0.7 以上。由表 1、图 12 知，测试集在 XGBoost 模型下表现最优，AUC 值高达 97%，准确度高达 93%。结果表明，基于 XGBoost 构建的前列腺癌诊断预测模型性能最好，鲁棒性最强。

Table 1. Evaluation results of 8 models
表 1. 8 种模型评价结果

模型	准确度	精确度	召回率	F1 分数
GLM	0.762	0.758	0.768	0.763
RF	0.925	0.932	0.914	0.923
SVM	0.911	0.925	0.894	0.909
KNN	0.881	0.857	0.914	0.885
Bayesian	0.918	0.909	0.927	0.918
LASSO	0.733	0.756	0.724	0.739
XGBoost	0.931	0.915	0.927	0.921
MLP	0.776	0.833	0.690	0.755

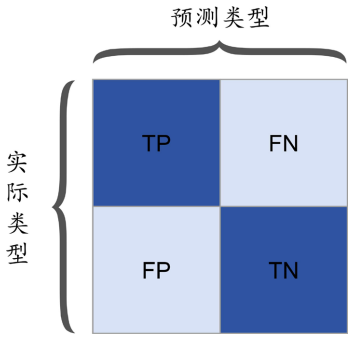


Figure 11. Confusion matrix
图 11. 混淆矩阵

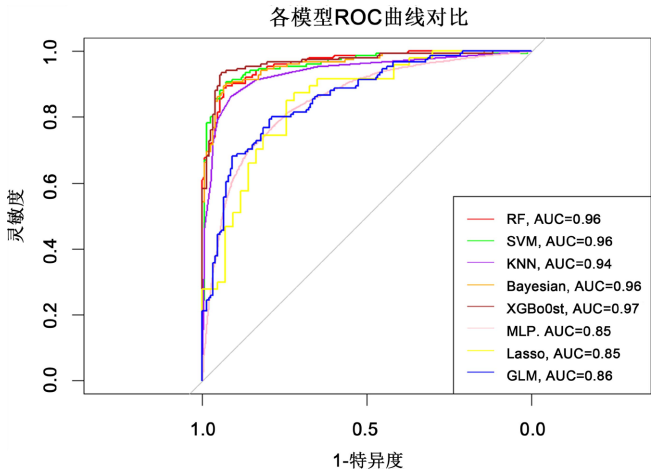


Figure 12. ROC curve of test set under 8 models
图 12. 8 种模型下测试集 ROC 曲线

5. 讨论

前列腺癌作为男性中最常见的恶性肿瘤之一，近年来其发病率和死亡率的不断上升引发了全球范围内的广泛关注。本研究旨在通过生物信息学和机器学习技术融合，探究前列腺癌与基因表达的关系并建立良好的预测疾病模型，从而为前列腺癌的诊断与治疗提供理论依据。

本研究首先使用生物信息学方法初步筛选出 2586 个差异基因，包含 1499 上调基因和 1087 个下调差异基因，然后对这些差异基因进行 GO 富集分析和 KEGG 富集分析。通过 GO 富集分析发现前列腺癌生物过程主要和肌肉收缩、细胞间黏附、嗜同性细胞粘附等有关等密切相关；通过 KEGG 富集分析发现其与神经活性配体与受体的相互作用、钙信号通路、肌肉细胞的细胞骨架和 cAMP 信号通路等密切相关。

研究基于三种机器学习方法(RF，LASSO 和 BGM)对上述所得的差异基因进行特征选择，最终得到了 12 个差异基因。同时，对上述 12 种关键基因构建了 8 种前列腺癌诊断预测模型。通过测试集验证并绘制 ROC 曲线、计算混淆矩阵，结果表明 8 种机器学习模型在诊断上表现出色，准确性和可靠性均较高，其中，XGBoost 模型准确率和 AUC 值分别为 93%、97%。

研究中所筛选出的这些基因，在已有文献中已有一定的相关性。Goel 等[8]表明 DLX1 在前列腺癌组织中显著高表达，并且在晚期和转移性前列腺癌患者中呈上调趋势；其高水平与肿瘤侵袭性增强和预后不良密切相关，可作为前列腺癌诊断的潜在非侵入性生物标志物。Wu 等[9]的研究表明，SLIT1 基因在前列腺癌中的表达升高。SLIT1 的高表达可能与前列腺癌的侵袭性及较差预后相关，在激素抵抗性的前列腺癌病例中尤为明显。这一发现提示 SLIT1 信号通路在前列腺癌进展中具有重要作用。通过基因敲降实验发现，Lu 等[10]发现转录因子 SIM2 在前列腺癌细胞中发挥促肿瘤作用。SIM2 在前列腺癌组织中呈上调表达；抑制 SIM2 可引发广泛的基因表达改变，影响细胞增殖和代谢通路，暗示 SIM2 的高表达有助于前列腺肿瘤细胞的增殖和分化异常，从而促进前列腺癌的发生发展。Chen 等[11]研究指出，TDRD1 在前列腺癌患者中异常高表达。TDRD1 属于胚系特异基因，在多达 68%的前列腺肿瘤中错误表达；其表达水平与雄激素受体驱动的 TMPRSS2-ERG 基因融合高度相关，是 ERG 转录因子的直接下游靶基因[11]。由于 TDRD1 过表达与前列腺癌早期复发风险相关，因此该基因有望用作前列腺癌的诊断及预后指标。通过基因组分析，Luo 等[12]发现 HOXC6 在前列腺癌中异常高表达，其高水平与肿瘤侵袭性增加及治疗后复发密切相关。HOXC6 被认为是侵袭性前列腺癌的临床相关生物标志物之一，可能通过与雄激素受体通路相关的转录调控机制促进前列腺癌的进展。Nasr 等[13]的综述指出，多种核糖体蛋白在前列腺癌组织中呈现高表达趋势，是驱动肿瘤发生的重要因素。RPS17 作为核糖体小亚基蛋白，其过表达可能通过增强肿瘤细胞的蛋白质合成能力，在前列腺癌的发生发展过程中起关键作用，提示其有潜力成为新的肿瘤标志物。Connell 等[14]发现 RPS10 基因在前列腺癌组织和泌尿生殖液中均高水平表达。Proteomics 分析显示 RPS10 在前列腺癌中蛋白水平显著上调。尽管其作为尿液标志物的研究尚不多见，RPS10 的过表达可能促进肿瘤细胞的蛋白合成和增殖，在前列腺癌中发挥促进作用。Zhao 等[15]发现 DHPS (脱氧高胺合成酶)通过催化真核翻译因子 eIF5A 的假亮氨酸化，在肿瘤细胞生长中扮演角色。虽缺乏针对前列腺癌的直接研究，其他肿瘤研究表明抑制 DHPS 可减少 eIF5A 活化，从而抑制癌细胞增殖和迁移[15]。这暗示 DHPS 可能影响前列腺癌细胞的生长调控，对肿瘤发生发展产生作用。

ARL6IP4 被 Javier-DesLoges 等[16]初步发现参与调控内质网应激(ER stress)，而内质网应激是肿瘤细胞在化疗或缺氧条件下存活的重要机制。在前列腺癌中，内质网应激的激活可能通过未折叠蛋白反应(UPR)促进癌细胞存活和耐药性，这对前列腺癌的研究有着一定价值。Stankewich 等[17]发现 SEC31B 作为 COPII 囊泡蛋白的组成部分，参与内质网和高尔基体之间的蛋白质转运。吴等[18]研究表明，内质网－高尔基体转运过程对肿瘤细胞的生长、迁移和侵袭至关重要。SEC31B 可能通过调节细胞内蛋白的分泌和

运输，在肿瘤细胞的适应性和生长中起到关键作用，对前列腺癌有一定研究价值。GSTP1 是前列腺癌研究中最常见的分子标志物之一。几乎 90%以上的前列腺癌标本中检测到 GSTP1 基因启动子区域的高水平异常甲基化[19]，导致基因沉默和蛋白表达缺失。这种启动子高甲基化在良性前列腺组织中罕见，却从前列腺上皮内瘤变(PIN)阶段即出现并贯穿癌变全过程[19]。因此，GSTP1 的甲基化状态被广泛用于前列腺癌的早期诊断，是公认的前列腺癌发生早期事件和诊断生物标志物。Varambally 等[20]基于表达谱分析发现，GOLM1 在前列腺癌组织中显著高表达，且主要由前列腺上皮肿瘤细胞产生。随后研究通过尿液检测证实，前列腺癌患者尿液中 GOLM1 mRNA 水平明显升高，并可作为前列腺癌的诊断指标，其诊断效能优于血清 PSA。此外，近期研究显示 GOLM1 过表达还能通过激活 TGF-β1/Smad2 通路促进上皮－间质转化(EMT)，在前列腺癌进展和转移中发挥促癌作用，体现了其作为诊断和预后生物标志物以及治疗靶点的潜力[21]。

6. 结论

本研究基于机器学习的方法，成功筛选出了关于前列腺癌的 12 个关键基因，这些基因在前列腺癌样本与正常样本中呈差异表达，可能作为未来诊断前列腺癌的潜在标志物。同时，构建了 8 种基于机器学习的前列腺癌预测诊断模型，其中 XGBoost 模型在测试集中表现最好，AUC 高达 97%，准确率高达 93%，这也进一步表明这些基因与前列腺癌的产生和发展有着密切的联系，可作为前列腺癌早期诊断及研究的潜在靶点。

综上所述，本文结合生物信息学与机器学习技术，完成了对前列腺癌关键生物标志物的筛选，同时构建了对前列腺癌的早期诊断预测的模型，AUC 值和准确率分别为 97%和 93%。本研究虽构建了基于 12 个关键基因的前列腺癌预测模型，但仍存在数据集规模有限、生物学机制缺乏实验验证、模型可解释性不足及临床因素未纳入等不足。未来研究将扩大数据集，提高模型泛化能力；通过细胞与分子实验验证关键基因功能；引入深度学习及特征选择算法优化模型性能；采用 SHAP、LIME 等方法增强模型可解释性；整合多组学数据提升诊断精准度；并推进模型临床转化，与 PSA 等生物标志物结合，开发精准检测工具，以促进前列腺癌的早期诊断与个性化治疗。

参考文献

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., *et al.* (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA*: *A Cancer Journal for Clinicians*, **71**, 209-249. https://doi.org/10.3322/caac.21660

[2] Siegel, R.L., Miller, K.D., Fuchs, H.E. and Jemal, A. (2022) Cancer Statistics, 2022. *CA*: *A Cancer Journal for Clinicians*, **72**, 7-33. https://doi.org/10.3322/caac.21708

[3] Pernar, C.H., Ebot, E.M., Wilson, K.M. and Mucci, L.A. (2018) The Epidemiology of Prostate Cancer. *Cold Spring Harbor Perspectives in Medicine*, **8**, a030361. https://doi.org/10.1101/cshperspect.a030361

[4] Attard, G., Parker, C., Eeles, R.A., Schröder, F., Tomlins, S.A., Tannock, I., *et al.* (2016) Prostate Cancer. *The Lancet*, **387**, 70-82. https://doi.org/10.1016/s0140-6736(14)61947-4

[5] Robinson, D., Van Allen, E.M., Wu, Y., Schultz, N., Lonigro, R.J., Mosquera, J., *et al.* (2015) Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell*, **161**, 1215-1228. https://doi.org/10.1016/j.cell.2015.05.001

[6] Barbieri, C.E. and Tomlins, S.A. (2014) The Prostate Cancer Genome: Perspectives and Potential. *Urologic Oncology*: *Seminars and Original Investigations*, **32**, 53.e15-53.e22. https://doi.org/10.1016/j.urolonc.2013.08.025

[7] Kothari, V., Wei, J.S., Shukla, S.K., *et al.* (2020) Machine Learning-Based Clinical Genomics Analysis of Prostate Cancer Outcomes. *Cancers*, **12**, Article 1164.

[8] Goel, S., Bhatia, V., Kundu, S., Biswas, T., Carskadon, S., Gupta, N., *et al.* (2021) Transcriptional Network Involving ERG and AR Orchestrates Distal-Less Homeobox-1 Mediated Prostate Cancer Progression. *Nature Communications*, **12**, Article No. 5325. https://doi.org/10.1038/s41467-021-25623-2

[9] Gara, R.K., Kumari, S., Ganju, A., Yallapu, M.M., Jaggi, M. and Chauhan, S.C. (2015) Slit/Robo Pathway: A Promising

Therapeutic Target for Cancer. *Drug Discovery Today*, **20**, 156-164. https://doi.org/10.1016/j.drudis.2014.09.008

[10] Lu, B., Asara, J.M., Sanda, M.G. and Arredouani, M.S. (2011) The Role of the Transcription Factor SIM2 in Prostate Cancer. *PLOS ONE*, **6**, e28837. https://doi.org/10.1371/journal.pone.0028837

[11] Feng, Q., Kim, H., Barua, A., Huang, L., Bolaji, M., Zachariah, S., Jung, S.Y., He, B., Zhou, T. and Mitra, A. (2023) The Cancer Testis Antigen TDRD1 Regulates Prostate Cancer Proliferation by Associating with snRNP Biogenesis Machinery. *Research Square*. This is a preprint. https://doi.org/10.21203/rs.3.rs-2035901/v1

[12] Luo, Z. and Farnham, P.J. (2020) Genome-Wide Analysis of HOXC4 and HOXC6 Regulated Genes and Binding Sites in Prostate Cancer Cells. *PLOS ONE*, **15**, e0228590. https://doi.org/10.1371/journal.pone.0228590

[13] El Khoury, W. and Nasr, Z. (2021) Deregulation of Ribosomal Proteins in Human Cancers. *Bioscience Reports*, **41**, BSR20211577. https://doi.org/10.1042/bsr20211577

[14] Yazbek Hanna, M., Winterbone, M., O'Connell, S.P., Olivan, M., Hurst, R., Mills, R., *et al.* (2023) Gene-Transcript Expression in Urine Supernatant and Urine Cell-Sediment Are Different but Equally Useful for Detecting Prostate Cancer. *Cancers*, **15**, Article 789. https://doi.org/10.3390/cancers15030789

[15] Zhao, G., Zhao, X., Liu, Z., Wang, B., Dong, P., Watari, H., *et al.* (2025) Knockout or Inhibition of DHPS Suppresses Ovarian Tumor Growth and Metastasis by Attenuating the TGFβ Pathway. *Scientific Reports*, **15**, Article No. 917. https://doi.org/10.1038/s41598-025-85466-5

[16] Javier-DesLoges, J., McKay, R.R., Swafford, A.D., Sepich-Poore, G.D., Knight, R. and Parsons, J.K. (2021) The Microbiome and Prostate Cancer. *Prostate Cancer and Prostatic Diseases*, **25**, 159-164. https://doi.org/10.1038/s41391-021-00413-5

[17] Stankewich, M.C., Stabach, P.R. and Morrow, J.S. (2006) Human Sec31B: A Family of New Mammalian Orthologues of Yeast Sec31p That Associate with the COPII Coat. *Journal of Cell Science*, **119**, 958-969. https://doi.org/10.1242/jcs.02751

[18] 吴诗洋, 常爽, 陈晴, 等. 肿瘤微环境调节型细胞器靶向递药系统的研究进展[J]. 药学学报, 2022, 57(6): 1771-1780.

[19] Martignano, F., Gurioli, G., Salvi, S., Calistri, D., Costantini, M., Gunelli, R., *et al.* (2016) *GSTP1* Methylation and Protein Expression in Prostate Cancer: Diagnostic Implications. *Disease Markers*, **2016**, Article ID: 4358292. https://doi.org/10.1155/2016/4358292

[20] Varambally, S., Laxman, B., Mehra, R., Cao, Q., Dhanasekaran, S.M., Tomlins, S.A., *et al.* (2008) Golgi Protein GOLM1 Is a Tissue and Urine Biomarker of Prostate Cancer. *Neoplasia*, **10**, 1285-IN35. https://doi.org/10.1593/neo.08922

[21] Qin, X., Liu, L., Li, Y., Luo, H., Chen, H. and Weng, X. (2023) GOLM1 Promotes Epithelial-Mesenchymal Transition by Activating TGFβ1/Smad2 Signaling in Prostate Cancer. *Technology in Cancer Research & Treatment*, **22**, 1-8. https://doi.org/10.1177/15330338231153618