doi:10.7659/j.issn.1005-6947.2023.04.002
http://dx.doi.org/10.7659/j.issn.1005-6947.2023.04.002
China Journal of General Surgery, 2023, 32(4):488-496.

·专题研究·

神经网络预测模型辅助诊断结直肠癌微卫星状态的研究

郝俊[1]，王帅[2]，朱军[3]，徐春盛[4]

（中国人民解放军空军军医大学第一附属医院 1. 实验外科 4. 胃肠外科，陕西 西安 710000；2. 中国人民解放军空军西安飞行学院一旅明港场站医院 门诊部，河南 信阳 463200；3. 中国人民解放军南部战区空军医院 普通外科，广东 广州 510000）

摘　要

背景和目的：微卫星不稳定（MSI）已经成为结直肠癌（CRC）临床诊断、辅助治疗和预后指导的重要生物学标志物。MSI往往伴随DNA错配修复蛋白（dMMR）的缺失。目前错配修复蛋白缺失的诊断主要依靠4种修复蛋白（MLH1、MSH2、MSH6和PMS2）病理免疫组化的结果，而且MSI已经成为CRC免疫治疗重要的生物学标志物。然而MSI精准预测模型和新型特征基因的研究很少。随着人工智能（AI）在医学的发展，精准预测和数据挖掘成为研究的热点。本研究的目的是建立MSI预测的神经网络模型和挖掘新型MSI特征基因。

方法：将3个CRC的GEO数据集（GSE39582、GSE29638和GSE75315）作为模型训练集，将1个TCGA CRC数据集作为独立的外部验证集。基于数据集测序数据和芯片数据，使用差异分析，随机森林算法和弹性反向传播算法建立CRC MSI的神经网络预测模型。用K-临近算法（KNN）和支持向量机（SVM）算法建立MSI传统机器学习网络模型。用混淆矩阵，受试者工作特征曲线（ROC）与曲线下面积（AUC）评价模型的预测能力。

结果：在训练集中，共纳入787例，其中微卫星高不稳定（MSI-H）111例（14.10%），微卫星低不稳定（MSI-L）/微卫星稳定（MSS）676例（85.90%）。在验证集中，TCGA数据集最终纳入389例，其中MSI-H 67例（17.22%），MSI-L/MSS 322例（82.78%）。通过差异分析计算出与MSI的相关基因100个，其中上调61个，下调39个。通过差异分析和随机森林算法，筛选出前30个贡献最大的MSI的特征基因。基于MSI相关基因的表达矩阵，建立了基于23个基因表达矩阵的神经网络预测模型。该模型在训练集（敏感度0.993，特异度0.973，诊断符合率0.990，AUC为0.991）和验证集（敏感度0.950，特异度0.828，诊断符合率0.933，AUC为0.922）模型均体现出精准的预测能力。此外，对比神经网络模型和机器学习的其他模型，结果表明神经网络模型在预测MSI方面更加准确。

结论：神经网络预测模型结合组织深度测序可以较好地辅助临床医生诊断CRC的MSI状态，为肿瘤免疫治疗方案的选择提供了参考和决策依据。同时，所鉴定的MSI的特征基因为深入研究相关的功能及机制提供了线索和方向。

关键词　结直肠肿瘤；微卫星不稳定性；人工智能；神经网络，计算机

中图分类号：R735.3

基金项目：国家自然科学基金资助项目（82100680）。

收稿日期：2022-03-01；**修订日期：**2022-05-07。

作者简介：郝俊，中国人民解放军空军军医大学第一附属医院助理研究员，主要从事结直肠癌发生机制方面的研究。

通信作者：徐春盛，Email: xucs1205@163.com

Neural network prediction model for assisting diagnosis of microsatellite status in colorectal cancer

HAO Jun[1], WANG Shuai[2], ZHU Jun[3], XU Chunsheng[4]

(1. Department of Experimental Surgery 4. Department of Gastrointestinal Surgery, the First Affiliated Hospital, Air Force Medical University, Xi'an 710000, China; 2. Department of Outpatient Services, Ming Gang Station Hospital, Xi'an Institute of Flight of the Air Force, Xingyang, Henan 463200, China; 3. Department of General Surgery, PLA Southern Theater Command General Hospital, Guangzhou 510000, China)

Abstract

Backgrounds and Aims: Microsatellite instability (MSI) has become an important biological marker for clinical diagnosis, adjuvant therapy, and prognostic guidance in colorectal cancer (CRC). Microsatellite instability often accompanies the loss of DNA mismatch repair proteins (dMMR). Currently, the diagnosis of mismatch repair protein deficiency mainly relies on the results of pathological immunohistochemistry for four repair proteins (MLH1, MSH2, MSH6, and PMS2), and MSI has become an important biological marker for immunotherapy in CRC. However, there are few studies on precise MSI prediction models and new signature genes. With the development of artificial intelligence in medicine, precise prediction and data mining have become research hotspots. The aim of this study was to establish a neural network model for MSI prediction and to discern new MSI signature genes.

Methods: Three CRC GEO datasets (GSE39582, GSE29638, and GSE75315) were used as model training sets, and one TCGA CRC dataset was used as an independent external validation set. Based on the sequencing data and microarray data of the datasets, a neural network prediction model for CRC MSI was established using differential analysis, random forest algorithm, and elastic backpropagation algorithm. Traditional machine learning models for MSI were established using K-nearest neighbor algorithm (KNN) and support vector machine (SVM) algorithm. The prediction ability of the models was evaluated using confusion matrices, receiver operating characteristic (ROC) curves, and the area under the curve (AUC).

Results: In the training set, a total of 787 cases were included, including 111 cases (14.10%) of microsatellite instability-high (MSI-H) and 676 cases (85.90%) of microsatellite instability-low/microsatellite stability (MSI-L/MSS). In the validation set, 389 cases in the TCGA dataset were finally included, including 67 cases (17.22%) of MSI-H and 322 cases (82.78%) of MSI-L/MSS. One hundred MSI-related genes were identified by differential analysis, including 61 up-regulated genes and 39 down-regulated genes. By combining differential analysis and random forest algorithm, the top 30 most significant MSI-related genes were screened out. Based on the expression matrix of the MSI-related genes, a neural network prediction model was established using 23 gene expression matrices. The model showed accurate prediction ability in both the training set (sensitivity: 0.993, specificity: 0.973, diagnostic coincidence rate: 0.990, AUC: 0.991) and the validation set (sensitivity: 0.950, specificity: 0.828, diagnostic coincidence rate: 0.933, AUC 0.922). Moreover, compared with other machine learning models, the neural network model demonstrated more accurate prediction ability in predicting MSI.

Conclusion: The neural network prediction model combined with tissue deep sequencing can assist clinicians in diagnosing the MSI status of CRC, and provide references and decision-making basis for the selection of tumor immunotherapy schemes. At the same time, the identified MSI signature genes provide clues and directions for in-depth research on related functions and mechanisms.

Key words Colorectal Neoplasms; Microsatellite Instability; Artificial Intelligence; Neural Networks, Computer

CLC number: R735.3

在我国，结直肠癌（colorectal cancer，CRC）的发病率居常见恶性肿瘤的第三位，病死率居第五位[1]。在散发性CRC发病机制中，微卫星不稳定（microsatellite instability，MSI）占15%，已成为林奇综合征诊断的标志。MSI往往伴有DNA错配修复蛋白缺失（DNA mismatch repair deficiency，dMMR）[2-3]，目前已作为CRC诊断、治疗和预后的标志物[4-7]。随着免疫治疗在实体肿瘤治疗中的兴起，免疫检查点抑制剂为晚期CRC患者的治疗带来了希望。然而，在临床运用的过程中，CRC患者对免疫检查点抑制剂的应答率不是很高。研究发现CRC微卫星高不稳定（high-microsatellite instability，MSI-H）与高水平肿瘤突变负荷[8-9]和细胞毒性T细胞的富集有关[10]，MSI-H是免疫检查点抑制剂疗效敏感的生物学标志物。因此，MSI-H已经成为CRC免疫治疗的有效人群[11]。然而MSI-H的人群只占CRC患者的15%左右[12]，而且目前MSI的检出方法各种各样，存在较大的敏感度和特异度差异[13]。生物信息学和人工智能（artificial intelligence，AI）网络的发展，为CRC患者高效诊断、精准分型和预后评价提供了可能。

本研究基于多个CRC的基因测序数据，使用随机森林和AI的方法，建立了新型的神经网络模型。该模型在训练集和验证集中都表现出极佳的预测准确性。同时，本研究也发掘了新的MSI相关基因，为MSI与免疫检查点治疗的机制研究提供了潜在的分子。

1 材料与方法

1.1 研究对象与数据收集

研究对象为临床诊断为CRC患者。纳入标准：(1)病例具有完整的二代测序数据；(2)病理检测微卫星状态或者免疫组化法检测错配修复蛋白缺失情况。排除标准：(1)合并其他肿瘤；(2)癌旁正常组织或转移组织的测序结果。

数据收集：在GEO官网（https://www.ncbi.nlm.nih.gov/geo）下载CRC完整测序数据GSE39582、GSE29638和GSE75315，在TCGA官网（https://portal.gdc.cancer.gov）下载CRC测序数据TCGA-COAD。GEO的3个数据集作为训练集，TCGA数据集作为外部验证集。使用Linear Models for Microarray Data（LIMMA）包中normalizeBetweenArrays函数对以上的数据进行标准化处理，使用SVA包中的Combat函数去除3个训练集的批次效应。

1.2 差异基因筛选

以微卫星稳定（microsatellite stability，MSS）/微卫星低不稳定（low-microsatellite instability，MSI-L）为对照组，以MSI-H为实验组。使用LIMMA包进行筛选差异基因，其校正方法为FDR法。筛选条件为：log差异倍数（FC）的绝对值>1，并且FDR值<0.05。

1.3 随机森林算法与神经网络学习

使用randomForest函数包，设置最大随机数为500，绘制误差与随机树的曲线。此外，按照基因贡献的重要性进行排序，本研究选择重要性评分前30个的基因作为后续人工神经网络预测（artificial neural network，ANN）的筛选基因。ANN算法前数据准备：对随机森林算法筛选的基因进行数据的预处理，构建基因表达矩阵。以基因表达的中位数为标准线，将>中位数的基因表达设置为1，≤中位数设置为0。使用neuralnet和NeuralNetTools函数包构建神经网络模型，隐藏层设置为5，并且绘制神经网络结构图。用建立好的神经网络模型，对训练集和验证集数据进行MSI状态的预测。将CRC MSI-L/MSS与MSH-H分别设置为0和1，以0.5为临界值，0.6~1.0为MSI-H组，0~0.5为MSI-L/MSS组。为了对比神经网络模型和传统机器学习算法的预测效能，本研究建立经典机器学习的支持向量机（support vector machine，SVM）和K-临近算法（K-nearest neighbor，KNN）来预测MSI。诊断一致性来验证模型之间的预测能力。

1.4 预测模型评价

MSI神经网络模型的评价方式为混淆矩阵，受试者工作特征曲线（receiver operating characteristic curve，ROC）。使用pROC函数包，绘制ROC，并且计算曲线下面积（area under curve，AUC）。AUC的范围为0~1，越接近1表示模型预测得越准确。使用随机抽样的方式（bootstrap法）计算AUC的95%可信区间（confidence interval，*CI*）。

1.5 统计学处理

本研究使用的统计软件为R语言（版本为4.02）。基线资料中，符合正态分布且方差齐性的计量资料使用平均值±标准差（$\bar{x} \pm s$），检验方法为*t*检验。不符合正态分布或方差不齐的计量资料使用中位数（四分位数间距）[*M*（*IQR*）]，检验

方法为非参数检验。随机森林与神经网络的模型构建都在 R 语言中完成，相关的 R 语言包如上所述。本研究采用的其余 R 语言包括：ggplot2、dplyr、kknn、reshape2、pROC。统计分析中采用双侧检验，$P<0.05$ 为差异有统计学意义。

2　结　果

2.1　基线资料

在训练集中：GSE75315 数据集最终纳入 206 例，其中 MSI-H 24 例（11.65%），MSI-L/MSS 182 例（88.35%）。GSE29638 数据集最终纳入 45 例，MSI-H 10 例（22.22%），MSI-L/MSS 35 例（77.78%）。GSE39582 数据集最终纳入 536 例，其中 MSI-H 77 例（14.37%），MSI-L/MSS 459 例（85.63%）。在验证集中：TCGA 数据集最终纳入 389 例，其中 MSI-H 67 例（17.22%），MSI-L/MSS 322 例（82.78%）。CRC 数据集的其他详细临床信息见表 1。

表 1　CRC 数据集的基本临床特征[*n*（%）]

Table 1　Basic clinical characteristics in CRC datasets [*n* (%)]

项目	训练集			验证集	项目	训练集			验证集
	GSE75315（*n*=206）	GSE29638（*n*=45）	GSE39582（*n*=536）	TCGA（*n*=389）		GSE75315（*n*=206）	GSE29638（*n*=45）	GSE39582（*n*=536）	TCGA（*n*=389）
MSI 状态					N 分期				
MSI-L/MSS	182(88.35)	35(77.78)	459(85.63)	322(82.78)	N0	—	—	270(50.37)	228(58.61)
MSI-H	24(11.65)	10(22.22)	77(14.37)	67(17.22)	N1	—	—	135(25.19)	94(24.16)
生存情况					N2	—	—	105(19.59)	67(17.23)
存活	—	—	354(66.04)	315(80.98)	未知	—	—	26(4.85)	—
死亡	—	—	179(33.40)	74(19.02)	M 分期				
未知	—	—	3(0.56)	0(0.00)	M0	—	—	455(84.89)	301(77.38)
性别					M1	—	—	58(10.82)	56(14.40)
男	118(57.28)	—	299(55.78)	182(46.79)	未知	—	—	23(4.29)	32(8.22)
女	88(42.72)	—	237(44.22)	207(53.21)	BRAF				
临床分期					野生型	177(85.92)	—	420(78.35)	—
I	1(0.49)	26(57.78)	—	69(17.73)	突变型	29(14.08)	—	45(8.40)	—
II	1(0.49)	4(8.88)	—	151(38.82)	未知	0(0.00)	—	71(13.25)	—
III	195(94.65)	3(6.67)	—	113(29.05)	KRAS				
IV	9(4.37)	12(26.67)	—	56(14.40)	野生型	132(64.08)	—	295(55.04)	—
T 分期					突变型	74(35.92)	—	203(37.81)	—
T1	—	—	16(2.99)	8(2.06)	未知	0(0.00)	—	38(7.09)	—
T2	—	—	49(9.13)	69(17.74)					
T3	—	—	348(64.93)	268(68.89)					
T4	—	—	103(19.22)	44(11.31)					
未知	—	—	20(3.73)	—					

2.2　MSI 差异基因筛选

LIMMA 包筛选差异基因，其中上调的分子数目为 39，在图 1 表示，使用红色点表示。下调的分子数目为 61，使用绿色的点表示。无差异表达的分子使用黑色的点表示。上升或下调的分子作为后续模型筛选的目标基因。

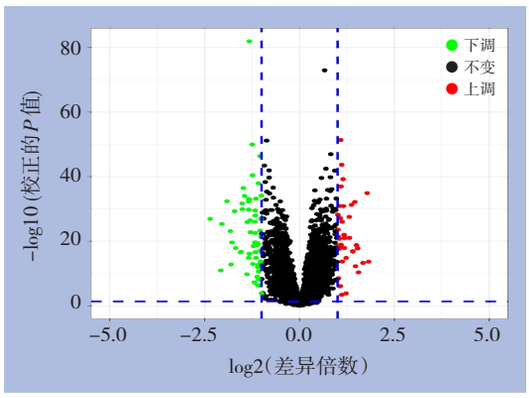

图1 MSI基因差异分析的火山图（红色的点表示高表达基因，绿色的表示低表达基因，黑色的表示无差异表达的基因）

Figure 1 Volcano plot for MSI differentially expressed genes (red dots representing upregulated genes, green dots represent downregulated genes, and black dots representing genes with no differential expression)

2.3 随机森林筛选MSI特征基因

在得到100个差异基因后，使用随机森林的方式筛选MSI的特征基因。随机森林曲线表明：随机树在30左右时，误差线基本保持稳定，基本处于最小值（图2）。因此，按照基因重要性展示排在前30位的基因，Gini系数越大，表示该基因在参与诊断MSI时越重要。

2.4 神经网络学习模型

对随机森林算法筛选出的前30位重要基因，进行神经网络模型的构建。30个基因的表达水平作为输入神经元，是否存在MSI-H为输出神经元，不断调整ANN参数使得模型的评价质量达到最优化。最终确定了“23–5–2”的结构模型（图3），其中输入层23个基因，隐藏层5个，输出层2个。在运行27 828步后得到的最终模型误差为6.98。输入层各变量与隐藏层之间的权重值见表2，隐藏层与输出层之间的权重值见表3。在输入层对隐藏层贡献力的评测分析中发现，AXIN2、ENO2、EIF5A和HPSE最为重要，表明这些基因与MSI或MMR高度相关。

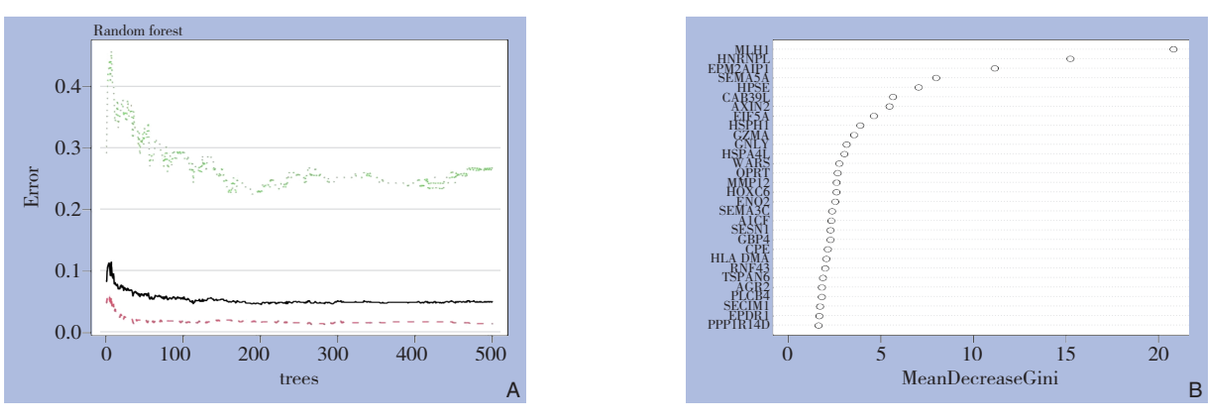

图2 随机森林筛选的MSI特征基因 A：决策树目与误差的关系；B：前30位贡献最大的基因的重要程度

Figure 2 MSI-related genes screened by random forest A: The relationship between the decision tree node and the error rate; B: The importance of the top 30 genes with the largest contributions

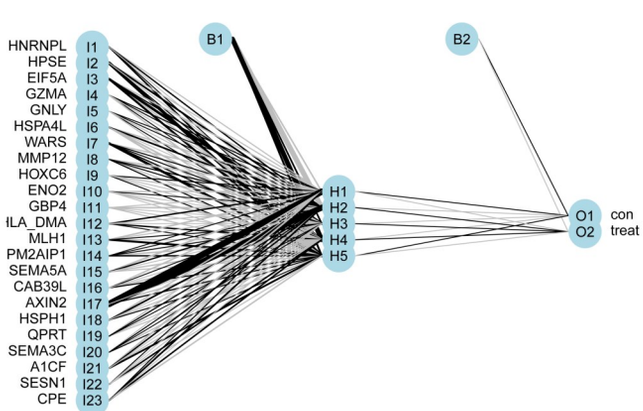

图3 人工神经网络

Figure 3 Diagram of artificial neural network

表 2　输入层与隐藏层的权重系数

Table 2　The weight coefficients of the input layer and the hidden layer

项目	H1	H2	H3	H4	H5
截距	−13.48	−99.23	16.77	89.09	23.59
HNRNPL	−5.40	11.02	−1.36	1.06	17.54
HPSE	−6.46	19.76	−2.04	4.26	52.63
EIF5A	−11.71	8.10	38.87	−11.59	26.65
GZMA	4.66	1.66	−3.45	0.34	−12.09
GNLY	1.25	1.61	−2.24	0.87	−6.36
HSPA4L	−10.64	16.74	−0.71	−18.26	−17.48
WARS	−10.35	4.26	3.67	9.56	37.86
MMP12	−16.66	5.76	3.58	−5.28	5.22
HOXC6	4.49	−8.25	−2.05	−2.57	1.74
ENO2	−10.05	0.99	−0.48	−21.76	−0.88
GBP4	−2.40	−15.43	−0.86	−3.14	−5.22
HLA-DMA	12.76	−2.03	−0.65	11.95	0.60
MLH1	−16.32	−2.13	2.34	−4.66	25.00
EPM2AIP1	0.41	1.17	1.69	−14.82	11.84
SEMA5A	−21.51	11.22	−0.12	−10.39	−3.44
CAB39L	−13.42	0.44	−0.77	−10.87	7.89
AXIN2	124.59	114.41	−1.33	−10.06	6.66
HSPH1	−12.57	16.72	−2.43	−3.97	4.08
QPRT	−10.79	7.56	1.50	−0.92	−3.41
SEMA3C	−2.80	4.50	1.60	−8.53	1.70
A1CF	−18.84	12.80	−2.48	−6.36	20.19
SESN1	1.98	−0.71	−2.10	14.19	−4.80
CPE	4.79	17.13	−0.75	14.04	−12.32

表 3　隐藏层与输出层的权重系数

Table 3　The weight coefficients of the hidden layer and the output layer

项目	MSI-L/MSS	MSI-H
截距	1.59	−1.08
H1	0.97	−0.97
H2	−0.98	0.98
H3	−0.53	1.01
H4	−1.06	1.06
H5	1.00	−0.99

2.5　模型的评价与运用

本研究使用神经网络模型预测训练集和验证集中的 CRC MSI 状态。在训练集中，神经网络模型预测的结果为：敏感度 99.26%，特异度 97.30%，诊断符合率 98.98%，阳性预测值 99.55%，阴性预测值 95.37%，AUC 0.991，95% *CI*=0.982~0.998；在验证集中（表 4）（图 4），敏感度 94.92%，特异度 82.76%，诊断符合率 93.03%，阳性预测值 96.76%，阴性预测值 75.00%，AUC 0.922，95% *CI*=0.866~0.969。以上结果表明，无论在训练集中，还是在验证集中，该神经网络模型的 AUC 接近 1，体现出较好的诊断效能。此外，与 KNN 和 SVM 等算法相比，神经网络模型在预测一致性上也表现出一定的优势（表 5）。

表 4　模型评价指标

Table 4　Model evaluation variables

评价指标	训练集	验证集
敏感度(%)	99.26	94.92
特异度(%)	97.30	82.76
阳性预测值(%)	99.55	96.76
阴性预测值(%)	95.37	75.00
诊断符合率(%)	98.98	93.30
假阳性率(%)	2.70	17.24
假阴性率(%)	0.74	5.08
AUC	0.991	0.922

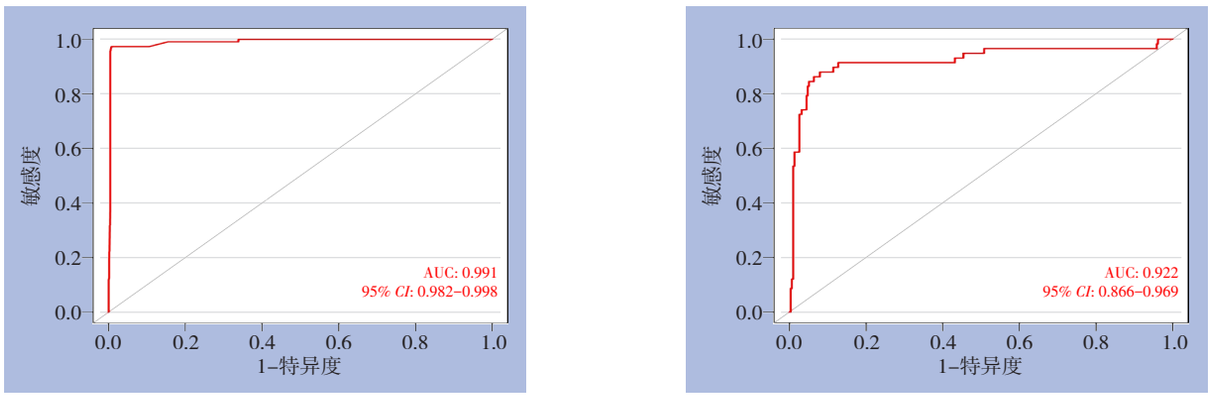

图4 训练集和验证集的ROC曲线

Figure 4 ROC curves of the training and validation sets

表5 不同算法预测CRC MSI状态的准确性

Table 5 Accuracy of different algorithms for predicting MSI status in CRC

项目	训练集	验证集
SVM	0.859	0.844
KNN	0.927	0.914
神经网络	0.989	0.933

3 讨 论

MSI已经成为CRC十分重要的临床特征之一。在CRC发病人群中，MSI途径占散发性CRC的15%，而且与林奇综合征泛癌的发生息息相关[14-15]。在肿瘤治疗中，MSI-H的临床II期CRC患者不适用以5-氟尿嘧啶为主的化疗方案，而MSI-H的CRC患者对伊立替康等化疗药物较为敏感[16]。在局部进展期低位直肠癌中，肠镜初诊活检组织中dMMR蛋白表型预示较好的新辅助放化疗疗效[17]。在肿瘤预后方面，MSI-H的肿瘤患者较MSI-L/MSS的肿瘤患者具有更好的预后[6]，此现象在临床II期的CRC患者中更为明显[18]。在免疫治疗中，多项RCT研究[19-20]一致表明，MSI成为CRC免疫治疗尤其是免疫检查点治疗的新型肿瘤标志。目前错配修复蛋白缺失的诊断主要依靠4种修复蛋白（MLH1、MSH2、MSH6和PMS2[21]）病理免疫组化的结果。然而，目前运用神经网络模型预测和筛选MSI相关基因的研究较少，并且缺乏对预测模型机制的深入研究。本研究基于多中心的测序数据，构建了评价效能和可信度较高的神经网络模型。ROC和多个评价指标均表明，无论是训练集还是验证集，该神经网络模型均体现出精准的预测能力。

在MSI诊断模型的构建研究中，学者们利用病理信息、病理切片信息和测序数据，对MSI状态进行预测。Hyde等[22]基于病理数据，构建了PREDICT的MSI-H预测模型。Cao等[23]利用TCGA和亚洲CRC人群的病理切片数据，使用反卷积神经网络的方式预测MSI状态。同时，他们发现该AI模型对肿瘤突变负荷的测定、免疫相关通路激活状态的判定和免疫检查点治疗的评价，有较高的契合度和预测效能[23]。此外，Lin等[24]在生物测序的基础上，研究MSI-H与免疫治疗之间的关联性，发现MSI-H人群与免疫细胞浸润、炎性免疫微环境的形成，以及免疫检查点治疗的敏感性密切相关。本研究基于多中心的测序数据，使用随机森林算法筛选出了30个重要基因，构建了相关的神经网络模型，并且在TCGA CRC数据集中进行了验证。结果表明，该神经网络模型预测准确度高、临床实用性强。此外，基于该神经网络模型，本研究筛选出4个贡献最大的核心基因，AXIN2，ENO2，EIF5A和HPSE。

AXIN2是对隐藏层1和2贡献最大的基因。AXIN2是CRC经典Wnt/β-catenin/TCF通路的负反馈蛋白基因[25]。AXIN2基因突变是导致CRC发生的基因缺失突变，且在肿瘤进展过程中发挥着重要的作用[26]。IDO1是抑制T细胞免疫的重要分子，也是免疫治疗的靶向分子之一[27]。有文献[28]报道称，抑制IDO1可以减少Wnt通路中β-catenin的入核与活化，从而抑制AXIN2的转录表达和肿瘤生长。MSI-AXIN2-IDO1调控轴可能是MSI-H患者具有良好免疫治疗敏感性的潜在机制。EIF5A是一个翻译起始因子，受羟腐胺赖氨酸作用调节。羟腐胺赖氨酸化的EIF5A可通过直接调节特定暂停状态下的

MYC 生物合成，从而促进 CRC 细胞的生长；而抑制 EIF5A 的羟腐胺赖氨酸化作用，可以抑制 CRC 细胞的生长[29]。但是，关于 EIF5A 与 MSI 之间的关系，目前尚未被深入研究和报道。ENO2 是糖酵解通路的烯醇化酶，在 BRAF V600E 突变的 CRC 中发挥重要作用。在 ENO2 受到抑制后，BRAF V600E 突变的 CRC 细胞生长也受到了明显抑制[30]。HPSE（乙酰肝素酶）是目前已知的唯一一种负责硫酸乙酰肝素裂解的哺乳动物内糖苷酶，是一种影响癌细胞多种恶性行为的多面蛋白[30]。研究发现，HPSE 和 CRC 的肝转移相关，与转移分子 MMP1 呈正相关关系。因此，HSPE 可能作为 CRC 治疗的一个方向[30]。以上文献表明，本研究发掘出的 MSI 相关基因 AXIN2 与免疫检查点分子 IDO1 相关，EIF5A、ENO2 与肿瘤生长相关，同时 ENO2、HPSE 还与 CRC 转移密切相关。

本研究虽然基于多中心测序数据集，但是也存在以下不足：(1) 二代测序花费较大，临床运用不普遍；(2) AXIN2、EIF5A、ENO2 和 HPSE 与 CRC MSI 的因果关系尚未进行深入探索。

总之，本研究为寻找更准确的 MSI 预测模型，使用神经网络的方式建立模型，并在相关训练集和验证集中得到了充分验证。最后基于神经网络学习的方法，筛选出了 4 个 MSI 的特征基因，为 MSI 免疫治疗提供了更多的研究方向和线索。

利益冲突：所有作者均声明不存在利益冲突。

作者贡献声明：徐春盛设计课题、数据分析；郝俊撰写论文；王帅辅助数据分析、参与论文撰写；朱军文章投稿、文稿校正。

参考文献

[1] Chen WQ, Zheng RS, Baade PD, et al. Cancer statistics in China, 2015[J]. CA Cancer J Clin, 2016, 66(2): 115−132. doi: 10.3322/caac.21338.

[2] 王乐, 李江, 朱陈, 等. 结直肠癌适宜筛查开始年龄的探讨[J]. 中华流行病学杂志, 2021, 42(6): 1113−1117. doi: 10.3760/cma.j.cn112338−20200807−01041.
Wang L, Li J, Zhu C, et al. Controversy on the age of initiation in colorectal cancer screening[J]. Chinese Journal of Epidemiology, 2021, 42(6): 1113−1117. doi: 10.3760/cma.j.cn112338−20200807−01041.

[3] Boland CR, Goel A. Microsatellite instability in colorectal cancer[J]. Gastroenterology, 2010, 138(6):2073−2087. doi: 10.1053/j.gastro.2009.12.064.

[4] Le DT, Durham JN, Smith KN, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade[J]. Science, 2017, 357(6349):409−413. doi: 10.1126/science.aan6733.

[5] Sargent DJ, Marsoni S, Monges G, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer[J]. J Clin Oncol, 2010, 28(20): 3219−3226. doi: 10.1200/JCO.2009.27.1825.

[6] Ma HY, Brosens LAA, Offerhaus GJA, et al. Pathology and genetics of hereditary colorectal cancer[J]. Pathology, 2018, 50(1): 49−59. doi: 10.1016/j.pathol.2017.09.004.

[7] Guastadisegni C, Colafranceschi M, Ottini L, et al. Microsatellite instability as a marker of prognosis and response to therapy: a meta-analysis of colorectal cancer survival data[J]. Eur J Cancer, 2010, 46(15):2788−2798. doi: 10.1016/j.ejca.2010.05.009.

[8] Germano G, Lamba S, Rospo G, et al. Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth[J]. Nature, 2017, 552(7683):116−120. doi: 10.1038/nature24673.

[9] Niu BF, Ye K, Zhang QY, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data[J]. Bioinformatics, 2014, 30(7): 1015−1016. doi: 10.1093/bioinformatics/btt755.

[10] Narayanan S, Kawaguchi T, Peng X, et al. Tumor infiltrating lymphocytes and macrophages improve survival in microsatellite unstable colorectal cancer[J]. Sci Rep, 2019, 9(1): 13455. doi: 10.1038/s41598−019−49878−4.

[11] Ganesh K, Stadler ZK, Cercek A, et al. Immunotherapy in colorectal cancer: rationale, challenges and potential[J]. Nat Rev Gastroenterol Hepatol, 2019, 16(6):361−375. doi: 10.1038/s41575−019−0126−x.

[12] Vilar E, Gruber SB. Microsatellite instability in colorectal cancer-the stable evidence[J]. Nat Rev Clin Oncol, 2010, 7(3): 153−162. doi: 10.1038/nrclinonc.2009.237.

[13] Lindor NM, Burgart LJ, Leontovich O, et al. Immunohistochemistry versus microsatellite instability testing in phenotyping colorectal tumors[J]. J Clin Oncol, 2002, 20(4): 1043−1048. doi: 10.1200/jco.2002.20.4.1043.

[14] Alicia L, Preethi S, Yelena K, et al. Microsatellite instability is associated with the presence of lynch syndrome pan-cancer[J]. J Clin Oncol Off J Am Soc Clin Oncol, 2019, 37(4): 286−295. doi: 10.1200/JCO.18.00283.

[15] 王玲玲, 刘正, 王锡山. Lynch综合征相关胃癌研究进展[J]. 中国普通外科杂志, 2020, 29(10):1243−1250. doi: 10.7659/j.issn.1005−6947.2020.10.011.

Wang LL, Liu Z, Wang XS. Progress in Lynch syndrome associated gastric cancer[J]. China Journal of General Surgery, 2020, 29(10): 1243−1250. doi: 10.7659/j.issn.1005−6947.2020.10.011.

[16] Bertagnolli MM, Niedzwiecki D, Compton CC, et al. Microsatellite instability predicts improved response to adjuvant therapy with irinotecan, fluorouracil, and leucovorin in stage III colon cancer: cancer and Leukemia Group B Protocol 89803[J]. J Clin Oncol, 2009, 27(11):1814−1821. doi: 10.1200/JCO.2008.18.2071.

[17] 程康文, 李佳, 王贵和, 等. 错配修复蛋白在直肠癌中的表达及其对新辅助放化疗敏感性的预测价值[J]. 中国普通外科杂志, 2020, 29(10): 1178−1186. doi: 10.7659/j. issn. 1005−6947.2020.10.004.
Cheng KW, Li J, Wang GH, et al. Expression of mismatch repair proteins in rectal cancer and its predictive value for sensitivity of neoadjuvant chemoradiotherapy[J]. China Journal of General Surgery, 2020, 29(10): 1178−1186. doi: 10.7659/j. issn. 1005−6947.2020.10.004.

[18] Merok MA, Ahlquist T, Røyrvik EC, et al. Microsatellite instability has a positive prognostic impact on stage II colorectal cancer after complete resection: results from a large, consecutive Norwegian series[J]. Ann Oncol, 2013, 24(5):1274−1282. doi: 10.1093/annonc/mds614.

[19] Le DT, Uram JN, Wang H, et al. PD-1 blockade in tumors with mismatch-repair deficiency[J]. N Engl J Med, 2015, 372(26):2509−2520. doi: 10.1056/NEJMoa1500596.

[20] Le DT, Kim TW, van Cutsem E, et al. Phase II open-label study of pembrolizumab in treatment-refractory, microsatellite instability-high/mismatch repair-deficient metastatic colorectal cancer: KEYNOTE−164[J]. J Clin Oncol, 2020, 38(1):11−19. doi: 10.1200/JCO.19.02107.

[21] 唐伟森, 廖明媚, 屈展, 等. 结直肠癌肿瘤组织PMS2蛋白表达状态与其临床病理特征的关系[J]. 中国普通外科杂志, 2019, 28(10):1297−1301. doi:10.7659/j.issn.1005−6947.2019.10.019.
Tang WS, Liao MM, Qu Z, et al. Expression status of PMS2 protein in colorectal cancer tumor tissue and the relationship with its clinicopathological characteristics[J]. China Journal of General Surgery, 2019, 28(10): 1297−1301. doi: 10.7659/j. issn. 1005−6947.2019.10.019.

[22] Hyde A, Fontaine D, Stuckless S, et al. A histology-based model for predicting microsatellite instability in colorectal cancers[J]. Am J Surg Pathol, 2010, 34(12):1820−1829. doi: 10.1097/PAS.0b013e3181f6a912.

[23] Cao R, Yang F, Ma SC, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer[J]. Theranostics, 2020, 10(24): 11080−11091. doi: 10.7150/thno.49864.

[24] Lin AQ, Zhang J, Luo P. Crosstalk between the MSI status and tumor microenvironment in colorectal cancer[J]. Front Immunol, 2020, 11:2039. doi: 10.3389/fimmu.2020.02039.

[25] Jho EH, Zhang T, Domon C, et al. Wnt/beta-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway[J]. Mol Cell Biol, 2002, 22(4):1172−1183. doi: 10.1128/MCB.22.4.1172−1183.2002.

[26] Mazzoni SM, Fearon ER. AXIN1 and AXIN2 variants in gastrointestinal cancers[J]. Cancer Lett, 2014, 355(1): 1−8. doi: 10.1016/j.canlet.2014.09.018.

[27] Zhai LJ, Ladomersky E, Lenzen A, et al. IDO1 in cancer: a gemini of immune checkpoints[J]. Cell Mol Immunol, 2018, 15(5):447−457. doi: 10.1038/cmi.2017.143.

[28] Thaker AI, Rao MS, Bishnupuri KS, et al. IDO1 metabolites activate β-catenin signaling to promote cancer cell proliferation and colon tumorigenesis in mice[J]. Gastroenterology, 2013, 145(2): 416−425. doi: 10.1053/j.gastro.2013.05.002.

[29] Coni S, Serrao SM, Yurtsever ZN, et al. Blockade of EIF5A hypusination limits colorectal cancer growth by inhibiting MYC elongation[J]. Cell Death Dis, 2020, 11(12): 1045. doi: 10.1038/s41419−020−03174−6.

[30] Yukimoto R, Nishida N, Hata T, et al. Specific activation of glycolytic enzyme enolase 2 in BRAF V600E-mutated colorectal cancer[J]. Cancer Sci, 2021, 112(7): 2884−2894. doi: 10.1111/cas.14929.

（本文编辑　宋涛）

本文引用格式：郝俊, 王帅, 朱军, 等. 神经网络预测模型辅助诊断结直肠癌微卫星状态的研究[J]. 中国普通外科杂志, 2023, 32(4):488−496. doi：10.7659/j.issn.1005−6947.2023.04.002

***Cite this article as*:** Hao J, Wang S, Zhu J, et al. Neural network prediction model for assisting diagnosis of microsatellite status in colorectal cancer[J]. Chin J Gen Surg, 2023, 32(4): 488−496. doi：10.7659/j.issn.1005−6947.2023.04.002