



利用人工智能预测癌症的易感性、复发性和生存期*

高美虹 尚学群**

(西北工业大学计算机学院, 西安 710072)

摘要 癌症具有较高的发病率和致死率, 对人类健康具有重大威胁。癌症预后分析可以有效避免过度治疗及医疗资源的浪费, 为医务人员及家属进行医疗决策提供科学依据, 已成为癌症研究的必要条件。随着近年来人工智能技术的迅速发展, 对癌症患者的预后情况进行自动化分析成为可能。此外, 随着医疗信息化的发展, 智慧医疗的理念受到广泛关注。癌症患者作为智慧医疗的重要组成部分, 对其进行有效的智能预后分析十分必要。本文综述现有基于机器学习的癌症预后方法。首先, 对机器学习与癌症预后进行概述, 介绍癌症预后及相关的机器学习方法, 分析机器学习在癌症预后中的应用; 然后, 对基于机器学习的癌症预后方法进行归纳, 包括癌症易感性预测、癌症复发性预测、癌症生存期预测, 梳理了它们的研究现状、涉及到的癌症类型与数据集、用到的机器学习方法及预后性能、特点、优势与不足; 最后, 对癌症预后方法进行总结与展望。

关键词 癌症预后分析, 人工智能, 智慧医疗, 易感性预测, 复发性预测, 生存期预测

中图分类号 R730, R730.7

DOI: 10.16476/j.pibb.2021.0334

癌症具有较高的发病率和死亡率, 是人类健康的巨大威胁^[1-3]。2018年, 全世界大约有1 810万癌症新生病例和960万癌症死亡病例^[2-3]。癌症预后, 是指根据病人当前状况对其治疗结果进行预测。癌症预后分析可以有效避免过度治疗及医疗资源的浪费, 为医务人员及家属进行医疗决策提供科学依据。癌症预后与基因突变相关, 常见的癌症预后相关的突变基因有TP53、KRAS、BRAF和PIK3CA等^[4-7]。对癌症预后相关的突变基因进行分析, 有助于癌症预后研究, 并可以为癌症患者的治疗提供一个参考。癌症预后受多种临床因素影响, 如患者的性别、年龄及其肿瘤分期等^[8-10]。常见的癌症预后方法及相关问题包括癌症易感性预测、癌症复发性预测和癌症生存期预测。

随着近年来电子计算机技术在医疗诊断中的应用及测序技术的发展, 对癌症的易感性、复发性及生存期进行自动化预测成为可能^[11-14]。一系列开源数据库(如TCGA和GEO等)提供了大量医疗数据, 创造了构建计算模型以有效进行癌症预后的机

会。医疗数据具有数据量大、模式复杂、个体表达特异等特点。机器学习方法可以从一系列复杂的医疗数据集中挖掘重要的模式, 极大推动了医疗信息化发展, 使得智慧医疗受到广泛关注。癌症患者作为智慧医疗的重要组成部分, 对其进行有效的智能预测十分必要。常见的基于机器学习的癌症预后方法及相关问题包括: a. 基于机器学习的癌症易感性的预测, 即用机器学习方法预测癌症发生的可能性; b. 基于机器学习的癌症复发性的预测, 即用机器学习方法预测癌症复发的可能性; c. 基于机器学习的癌症生存期的预测, 即用机器学习方法预测癌症的生存结果。常用的癌症预后相关的机器学习方法包括COX比例风险回归、随机生存森林(RSF)、LASSO回归、人工神经网络(ANN)、贝叶斯网络(BN)、支持向量机(SVM)和决策树(DT)等^[15-19]。机器学习方法已被广泛应用于癌症

* 国家自然科学基金(61772426, U1811262)资助项目。

** 通讯联系人。

Tel: 13484907397, E-mail: shang@nwpu.edu.cn

收稿日期: 2021-10-30, 接受日期: 2022-02-08

预后研究中,其可以基于癌症样本数据得出有效和准确的预后结论^[17, 20-26]。

本文重点对已有的基于机器学习的癌症预后方法进行综述。第1节对机器学习方法与癌症预后进行了概述,包括癌症预后分析、相关机器学习方法及机器学习在癌症预后中的应用。第2~4节对基于机器学习的癌症预后方法进行了详细分析,统计了近20年来基于机器学习方法的癌症预后相关研究,并对常用的癌症预后分析方法进行了比较。第5节对全文进行了总结,并对未来值得关注的研究方向进行初步探讨。

1 机器学习与癌症预后

机器学习方法与癌症预后的关系如图1所示,其可以基于回归分析方法、神经网络分析方法、SVM分析方法及一些常用的机器学习方法,对癌症患者的mRNA表达数据、lncRNA表达数据、miRNA表达数据及DNA甲基化数据进行分析,构建预后模型,获得癌症患者的易感性、复发性及生存期相关的预后情况。下面分别对癌症预后、机器学习分析过程及预后相关的机器学习方法进行详细介绍。

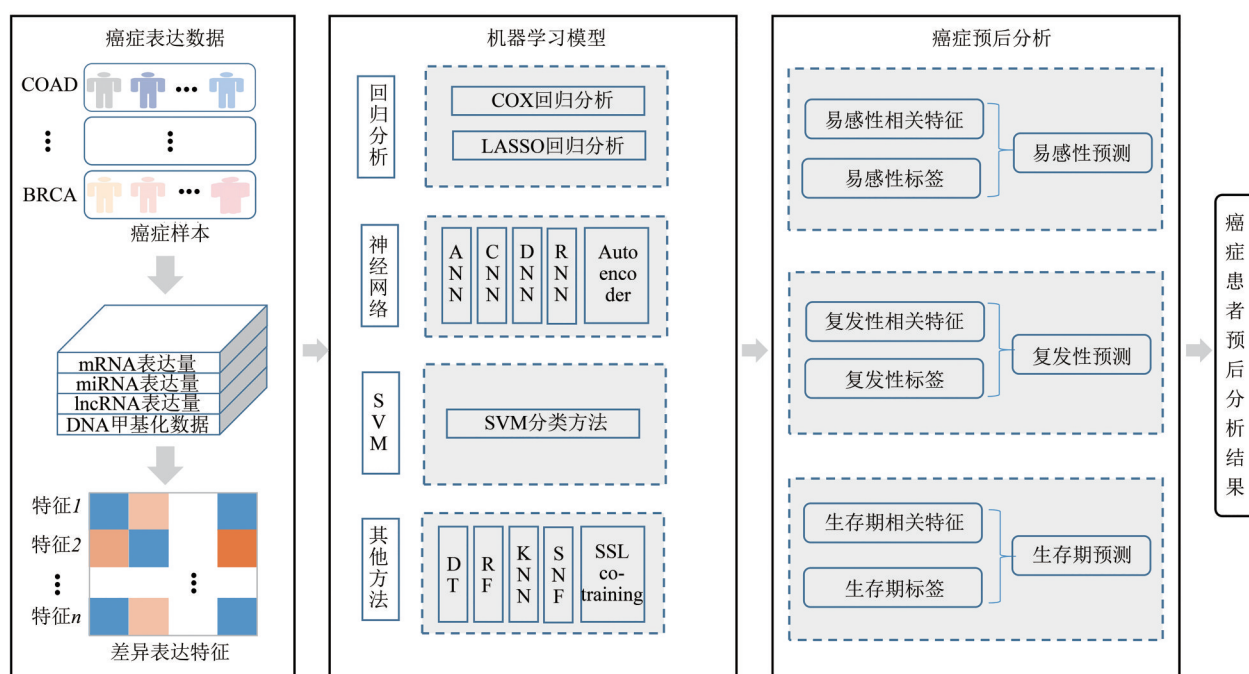


Fig. 1 Schematic diagram of cancer prognosis analysis based on machine learning

图1 基于机器学习的癌症预后分析示意图

1.1 预后分析概述

癌症预后分析可以对患者的预后风险进行评估,通过患者的基因表达情况,将患者的预后情况分为高风险和低风险两类。图2和图3所示为癌症预后分析的一个实例:首先,获得癌症样本和癌旁样本的分布情况(图2),其中FC表示两种样本间基因表达值的倍数变化,FC越大则样本间的表达差异越大,图2a~c分别表示logFC的绝对值大于1、大于1.5及大于2时样本表达量的聚类结果,可以发现logFC的绝对值越大,聚类效果越好;然

后,对癌症样本和癌旁样本的表达情况进行分析,获得差异表达的mRNA特征、lncRNA特征、miRNA特征及DNA甲基化特征等;最后,对这些特征在癌症样本中的表达情况进一步分析,获得癌症样本的预后情况(图3),其中图3a~c分别表示所有集合、训练集及测试集上癌症样本的预后情况,图中红线表示高预后风险的样本生存时间分布,绿线表示低预后风险的样本生存时间分布,可以发现基于表达值的预后分析在3个集合上都有着良好的预后效果。

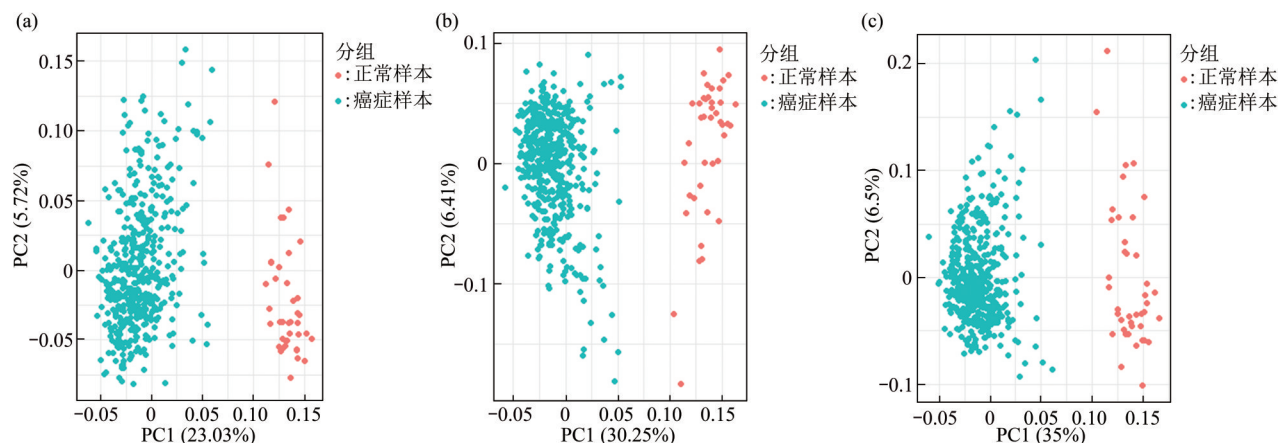


Fig. 2 Schematic diagram of the distribution of samples related to cancer survival prognosis

图2 癌症生存期预后样本分布示意图

(a) 样本筛选阈值为 $|\log FC| > 1$, (b) 样本筛选阈值为 $|\log FC| > 1.5$, (c) 样本筛选阈值为 $|\log FC| > 2$ 。

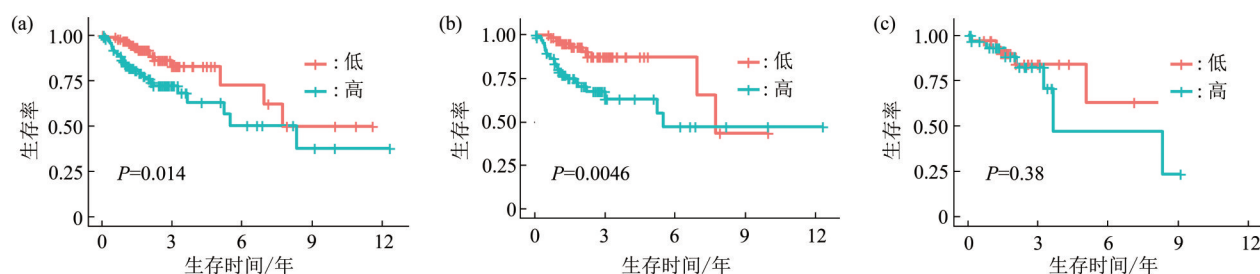


Fig. 3 Schematic diagram of cancer prognosis results

图3 癌症预后结果示意图

(a) 所有样本的癌症预后结果, (b) 训练集中样本的癌症预后结果, (c) 测试集中样本的癌症预后结果。

1.1.1 癌症预后分析

癌症预后与基因突变相关, 常见的癌症预后相关的突变基因有 TP53、KRAS、BRAF 和 PIK3CA 等^[4-7]。KRAS 的突变与胰腺癌的预后有关, 其是导致胰腺癌发生的主要事件, 且对胰腺癌的治疗具有重要意义^[4]。TP53 的突变与乳腺癌和其他一些癌症的不良预后有关, TP53 突变的确切位置与疾病的产生有关, 故 TP53 突变的位置可以为疾病的治疗提供指导^[5]; BRAF 的突变与甲状腺乳头癌 (PTC) 的预后有关, 其可以作为一个预测因子对 PTC 患者进行生存风险预测^[6]; PIK3CA 的突变与结肠癌的预后有关, 在接受结肠癌切除术的患者中, 具有 PIK3CA 突变患者的生存期较其他患者短^[7]。综上, 基因突变与癌症患者的预后情况相关, 对癌症预后相关的突变基因进行分析, 有助于癌症预后研究, 并可以为癌症患者的治疗提供一个

参考。

癌症预后受多种临床因素影响, 如患者的性别、年龄及其肿瘤分期等^[8-10]。在性别方面, 男性更容易患肺癌、肝癌和皮肤癌, 而女性更容易患甲状腺癌、乳腺癌和肾上腺皮质癌^[8]。在年龄方面, 癌症患者的存活率与其最初病理诊断时的年龄有关, 某些类型的癌症更容易在较年轻的群体中发生, 如肾上腺皮质癌、宫颈癌、脑部低级别胶质瘤、嗜铬细胞瘤、副神经节瘤、睾丸生殖细胞瘤和甲状腺癌等^[9-10]。在肿瘤分期方面, 分期值越低, 代表肿瘤处于越早期阶段, 预后情况越好; 而分期值越高, 则肿瘤处于越晚期阶段, 预后情况也会越差^[8]。

1.1.2 预后分析类型

随着测序技术与计算机技术的发展, 对癌症患者的预后情况进行自动化和个性化分析成为可

能^[11-14]。一系列开源数据库如TCGA和GEO等提供了大量医疗数据,这为人们提供了能构建更强大、更准确的模型以更有效预测癌症的机会。医疗数据具有数据量大、模式复杂、个体表达特异等特点。机器学习方法可以从一系列复杂的医疗数据集中挖掘重要的模式,极大推动了医疗信息化发展,使得智慧医疗受到广泛关注^[27-31]。癌症患者作为智慧医疗的重要组成部分,对其进行有效的智能预测十分必要。常见的基于机器学习的癌症预后方法及相关问题包括:

a. 癌症易感性的预测,即用机器学习方法预测癌症发生的可能性(第2节)。对某些常见癌症,如胃癌、肝癌、乳腺癌等的发生风险进行有效预测,得出癌症易感性的高低,为癌症的预防提供一个参考。

b. 癌症复发性的预测,即用机器学习方法预测癌症复发的可能性(第3节)。对癌症患者的术后复发风险进行有效预测,得出癌症复发性的风险,为癌症的治疗提供一个参考。

c. 癌症生存期的预测,即用机器学习方法预测癌症的生存结果(第4节)。对癌症患者的生存风险进行有效预测,得出癌症患者生存风险的高低,为癌症的治疗提供一个参考。

1.1.3 预后分析数据集

a. TCGA数据集。TCGA是一个开源的癌症数据集,包括level 1、level 2数据及level 3三种类型的数据。level 1和level 2数据属于测序数据,数据规模庞大、处理过程复杂且下载权限不易获取;level 3数据属于表达数据,数据规模较小、处理过程相对简单且可以免费获取。在基于机器学习方法的预后分析中,用到的数据类型以level 3数据居多。TCGA数据的下载地址为<https://portal.gdc.cancer.gov/>。

b. GEO数据集。GEO是一个开源的功能基因组学数据集,收录了序列数据和阵列数据两种,其中序列数据通过测序获得,数据规模庞大、处理过程复杂且下载权限不易获取,而阵列数据通过基因芯片获得,数据规模较小、处理过程相对简单且容易获取。在预后分析中,GEO数据可以作为TCGA数据的外部验证数据集,对预后模型进行有效性验证。GEO数据的下载地址为<https://portal.gdc.cancer.gov/>。

c. UCSC Xena数据集。UCSC Xena是一个开源的癌症全基因组数据集,收录了众多数据集中的

数据,包括TCGA、GETx以及target等,并对其进行了预处理。UCSC Xena数据集中的数据以表达数据为主,数据规模较小且处理过程简单。在预后分析中,UCSC Xena数据可以作为TCGA数据的外部验证数据集,对预后模型进行有效性验证。UCSC Xena数据集的下载地址为<http://xena.ucsc.edu/>。

d. GENCODE数据集。GENCODE是一个开源的基因注释数据集,收录了人类和小鼠的基因注释信息,在提取癌症患者的表达数据时需要用到此类信息。例如,在进行TCGA样本的mRNA和lncRNA表达数据提取时,选择的是GENCODE中的v19注释数据,这是由TCGA数据注释时的版本信息所决定的。GENCODE数据集的下载地址为<https://www.gencodegenes.org/>。

1.2 机器学习分析过程概述

机器学习(ML)是人工智能的一个重要分支,它可以对已有的样本数据进行学习,从而得出推理的结论^[32-34]。ML过程包括两个阶段:第一阶段是学习阶段,根据给定数据集对系统中的未知依赖关系进行估计;第二阶段是推理阶段,根据预测的依赖关系对系统的输出进行估计。ML方法分为有监督学习和无监督学习两种,有监督学习方法的输入数据包含标签,而无监督学习方法的输入数据不包含标签。ML的目标是生成一个用于预测、分类、评估及其他类似任务的模型。模型生成前,需要进行数据集划分。模型生成后,需要进行性能评估与误差分析。

数据集划分过程中应保证训练集和测试集的样本足够大且相互独立,测试集的标签信息已知。常用数据集划分方法包括:a. 留出法,将数据集划分为两个互斥且分布尽量一致的集合;b. 随机抽样法,将留出法重复多次,随机选择训练样例和测试样例;c. 交叉验证法,将数据集划分为 k 个互斥的集合;d. 自助法,从数据集中有放回采样 m 次,获得包含 m 个样本的数据作为训练集,那些未被抽到的数据作为测试集。

性能评估需确定以下几个参数: TP ,实际情况与预后结果均为正样本; TN ,实际为正样本,预后结果为负样本; FP ,实际为负样本,预后结果为正样本; FN ,实际情况与预后结果均为负样本。根据以上4个参数,可以得到下面的性能评估指标:

a. TPR ,即真阳性率,也叫灵敏度,用于评估

预测出的正样本占有所有正样本的比例, 定义如下:

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

b. *FPR*, 即假阳性率, 用于评估预测出的正样本占有所有负样本的比例, 定义如下:

$$FPR = \frac{FP}{FP + TN} \quad (2)$$

c. *TNR*, 即真阴性率, 也叫特异性, 用于评估预测出的负样本占有所有负样本的比例, 定义如下:

$$TNR = \frac{TN}{FP + TN} \quad (3)$$

d. *ACC*, 即准确度, 用于评估正确分类的样本数与总样本数的比值, 定义如下:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

e. *AUC*, 用于评估 ROC 曲线下的面积, ROC 曲线的横纵坐标分别为 *FPR* 和 *TPR*。

ML 在生成分类模型时, 会产生训练和泛化两种错误, 训练错误是指训练集上的误分类错误, 而泛化错误是指测试集上的预期错误。ML 中常用的泛化误差分析方法为偏差-方差分解法, 即将泛化误差分解为偏差、方差、噪声之和, 对这 3 种误差分别进行估计。

1.3 机器学习分析方法概述

ML 方法可以从一系列复杂的数据集中挖掘重要的模式, 在预后分析中发挥着重要作用, 以下对几个常用预后相关的 ML 分析方法进行介绍。

1.3.1 回归分析方法

预后相关的回归分析方法包括 COX 比例风险回归分析和 LASSO 回归分析等^[35-36]。COX 回归以癌症患者的生存时间和生存状态为预后变量, 分析预后特征对患者预后变量的影响。LASSO 回归是一种常用的正则化回归方法, 通过构造一个惩罚函数来对患者的预后情况进行有效筛选, 得到更为精炼的预后特征, 从而对癌症患者的预后情况进行准确判断^[37]。

COX 回归分析又分为单因素 COX 回归分析和多因素 COX 回归分析, 单因素 COX 回归用于确定单个预后特征对预后变量的影响, 多因素 COX 回归用于确定多个预后特征协同作用对预后变量产生的影响。预后相关的 COX 回归分析过程包括: a. 对癌症数据进行初步分析, 获得各个特征的表达值; b. 用偏似然函数对预后特征进行参数估计; c. 用 Wald 检验方法进行假设检验, 获得显著的预后特征; d. 根据预后特征及对应的回归系数, 构建

风险预后模型, 并用 KM 方法对模型的预后能力进行评估。

LASSO 回归分析用以减小模型的过拟合。由方差权衡可知, 建立回归模型时的一个重要的任务是进行变量选择。传统的变量选择方法存在灵活性差且模型方差高的缺点, 基于正则化回归的 LASSO 回归方法应运而生。预后相关的 LASSO 回归分析过程包括: a. 获得预后特征的表达值; b. 加入正则项对预后特征进行回归惩罚; c. 将冗余预后变量的系数变为 0; d. 筛选回归系数不为 0 的特征进行预后分析。

1.3.2 神经网络分析方法

神经网络是对生物神经系统的交互反应进行模拟, 由具有适应性的简单单元组成广泛并行的互联网络。神经网络各层之间通过权值连接, 通过激活函数对输出进行控制。预后相关的神经网络分析方法包括人工神经网络 (ANN)、卷积神经网络 (CNN)、深度神经网络 (DNN)、循环神经网络 (RNN) 和自编码器 (autoencoder) 等^[38-42]。这些方法的输入为预后属性, 通过权值矩阵和激活函数对预后属性进行处理, 得到患者的预后结果并输出。其基本结构包括: a. 输入层, 对预后特征进行表示; b. 权值矩阵, 位于输入层与隐藏层之间, 对对预后特征进行加权处理; c. 激活函数, 位于隐藏层与输出层之间, 用于计算并输出最终的预后结果; d. 输出层, 输出癌症的预后结果。

1.3.3 支持向量机分析方法

SVM 根据患者的预后属性对他们进行分类, 最终将患者的预后结果分为预后好和预后差两类^[43]。超平面被认为是两种预后结果间的决策边界, 两种预后类别处于超平面的两侧。对于线性可分的癌症预后数据集来说, 这样的超平面可能有很多个, 但是使得几何间隔最大的超平面是唯一的。预后相关的 SVM 分类过程包括: a. 获得预后特征的表达值; b. 训练获得合适的超平面, 将训练集中的样本分为预后好和预后差两类; c. 根据训练好的模型, 对癌症样本进行预后分析。

1.3.4 其他分析方法

决策树 (DT) 方法、随机森林 (RF) 方法、K 近邻 (KNN) 方法、半监督学习的协同训练 (SSL co-training) 方法与相似网络融合 (SNF) 方法也可用于癌症预后分析^[44-50]。DT 方法根据预后属性与节点分裂算法建立决策树模型, 预后结果由叶子节点的值确定。RF 方法对多个 DT 的预后结果

进行集成,做出最终的预后判断。KNN方法对离癌症样本最近的K个邻居样本的预后结果进行分析,得出该样本的预后结果。SSL co-training方法用有标记预后样本对无标记预后样本进行标记,得出其预后结果。SNF方法对不同的预后特征进行计算并融合,获得统一的相似性网络并进行预后分析。

2 基于机器学习的癌症易感性预后

基于ML的癌症易感性预测,是指用ML方法对某些常见癌症,如胃癌、肝癌、乳腺癌等的发生风险进行有效预测,得出癌症易感性的高低,为癌症的预防提供一个参考。

2.1 机器学习与癌症易感性预测

ML方法可以对癌症的易感性进行预测。某些常见的癌症,如乳腺癌、胃癌和肝癌等在特定家族中有着较大的发病率^[51-53],有必要对这些癌症的易感性进行有效预测。癌症易感性预测具有重要意义:a.有助于揭示癌症的发生、发展及恶化机制^[54];b.有助于癌症患者选择恰当的治疗路线及方法,如携带肿瘤易感基因的患者更容易在某些药物的影响下发生第二肿瘤,所以应尽量避免这类药物的使用^[55];c.有助于携带遗传易感基因的患者选择合适的防治方案^[56]。随着对癌症易感性的深

入研究及ML技术的广泛使用,基于ML的癌症易感性预测方法逐渐受到关注。该方法对癌症患者的易感性数据进行学习,获得有效的ML模型,从而对癌症患者的易感性情况进行预测。癌症易感性预测是癌症预后的重要组成部分,本文对其相关研究进行了深入分析。

2.2 基于机器学习的癌症易感性预测研究现状

本文对PubMed上收录的过去20年间发表的癌症易感性预测相关研究进行统计,检索的关键词是“machine learning in cancer susceptibility”(表1)。可以发现,在过去的20年中,用ML方法对癌症易感性进行预测的相关研究数量逐渐增长,且增长速率在近5年显著提升。其中,2002~2006年间基于ML的癌症易感性预测的相关研究总数为4,2007~2011年此数值增长到了10,2012~2016年此数值增长到了42,2017年至今(截至2021年6月22日)相关研究的数量达到了141。可以发现,基于ML的癌症易感性预测在过去的20年中逐渐受到研究者的关注,尤其是在过去5年中,此类研究的数量明显提升。具体来说,2017~2021年相关研究的数量是2002~2006年相关研究数量的35倍之多,是2007~2011年相关研究数量的14倍之多,是2012~2016年间相关研究数量的3倍之多。综上,ML在癌症易感性预测方面的研究逐渐成为研究热点。

Table 1 Research status of cancer susceptibility prediction based on machine learning (2002–2021)

表1 基于机器学习的癌症易感性预测研究情况 (2002~2021)

时间	第一年论文数	第二年论文数	第三年论文数	第四年论文数	第五年论文数	论文总数
2002~2006	0	1	3	0	0	4
2007~2011	3	2	1	2	2	10
2012~2016	2	7	3	11	19	42
2017~2021	18	29	45	31	>18	>141

2.3 基于机器学习的癌症易感性预测比较分析

进一步对基于ML的癌症易感性预测的相关研究进行比较分析(表2)。可以发现,CNN、RNN、独立成分分析(ICA)、K均值聚类(K-means)、RF、SVM、DT、朴素贝叶斯(naïve Bayes)、SSL co-training及多分类方法在预测癌症易感性方面都有着良好的性能^[57-65]。CNN和RNN属于有监督的神经网络分类方法(神经网络方法在癌症的诊断和预后中有着广泛的应用^[66-68]),可以对癌症患者的MRI图形进行分析,获得合适的神经网络模型,从而对患者的易感性进行预测。DT、RF、SVM和

naïve Bayes属于有监督学习的分类方法,其中DT和RF属于树结构分类方法,且RF属于集成学习范畴,其集成多个DT的预后结果以做出最终的预后判断。ICA和K-means属于无监督学习分类方法,K-means通过对癌症患者的预后信息进行聚类以获得癌症易感性预测结果,ICA通过识别癌症样本的属性特征并从中提取有用的特征进行预易感性预测。Co-training属于半监督学习方法,通过对有易感性标记的癌症样本和无易感性标记的癌症样本进行处理,从而获得癌症样本的易感性预测结果。

Table 2 Comparative analysis of cancer susceptibility prediction based on machine learning
表2 基于机器学习的癌症易感性预测比较分析

机器学习方法	癌症类型	数据来源	预测性能	发表时间	PubMed
CNN	色素性皮损	ISIC数据集	$AUC \approx 0.9$	2021	33423009 ^[57]
SVM	高级胶质瘤	MRI成像	灵敏度/特异性/准确度= 100%/91.67%/94.7%	2021	34047805 ^[58]
RF	胶质母细胞瘤	MRI成像	准确度=0.950±0.091	2019	31448207 ^[59]
RNN	胶质瘤	MRI成像	灵敏度/特异性/准确度= 92.6%/93.1%/92.8%	2019	31127834 ^[60]
ICA, K-means	胶质瘤	MRI成像	准确度=86.7%	2019	31176951 ^[61]
多分类方法	20种常见癌症	TCGA数据集, 1000 Genome数据集	$AUC=0.75\sim0.96$	2018	29358382 ^[62]
SVM	乳腺癌	Utah数据集, Ontario数据集	$AUC=0.73\sim0.76$	2015	26538066 ^[63]
SSL co-training	乳腺癌	SEER数据集	$AUC=0.81$	2014	25080202 ^[64]
SVM, DT, naïve Bayes	乳腺癌	SNP数据	准确度=69%	2004	15102677 ^[65]

癌症易感性预测涉及多种癌症：色素性皮损、胶质瘤、乳腺癌及常见的20种癌症。色素性皮损是一种常见疾病，目前已有基于此疾病的图像数据库及检测相关的研究^[69-70]。胶质瘤是最常见的脑部原发性肿瘤，它拥有多种亚型且预后情况各不相同^[71-72]。乳腺癌是一种常见的具有高发病率和致死率的癌症，位居女性恶性肿瘤的首位，超过90%的乳腺癌患者在诊断时未发生转移，故对于乳腺癌患者，其主要治疗目标是根除肿瘤和预防复发^[73]。常见的20种癌症为TCGA数据集中包含的癌症，如膀胱尿路上皮癌、结肠癌、头颈鳞状细胞癌、胃癌等。

癌症易感性预测的数据来源包括临床数据和公共数据两种。临床数据通过收集癌症患者的医疗信息获得，收集渠道一般为医院或医学研究机构。公共数据通过从公共数据集下载获得，如ISIC数据集、TCGA数据集、1000 Genome数据集、Utah数据集、Ontario数据集和SEER数据集等。这两种易感性预测数据的类型包括表达数据、MRI成像数据和SNP数据等。其中，表达数据包括mRNA表达数据、miRNA表达数据、lncRNA表达数据和DNA甲基化数据等，MRI成像数据是癌症患者的病理图像数据，SNP数据则是癌症患者相关基因的单位点变异数据。

癌症易感性预测的性能由其涉及的癌症类型、ML方法、分析的数据类型共同决定。基于CNN对ISIC数据集中色素性皮损患者进行易感性分析时， AUC 值达到了0.9^[57]。基于SVM、RF、RNN、ICA和K-means对胶质瘤患者MRI图像数据进行分

析时，准确度接近或超过90%，甚至会达到95%^[58-61]。基于多分类方法对TCGA数据集和1000 Genome数据集中20种常见癌症进行分析时， AUC 在0.75~0.96之间^[62]。基于SVM、SSL co-training、DT和naïve Bayes对乳腺癌患者SNP数据进行分析时， AUC 在0.73~0.81之间^[63-65]。综上，ML在癌症易感性预测方面有着良好的性能，其对色素性皮损和胶质瘤的易感性预测有着较高的准确度，对乳腺癌的易感性预测性能较其他癌症差一些，但也在可接受的范围内。总的来说，ML方法在癌症易感性预测方面表现优良，尤其是近几年提出的预测方法，其准确度逐渐提高。

癌症易感性预测可以为临床医生提供指导，在智慧医疗的癌症预后方面有着极大价值。癌症易感性预测的特点包括：预测数据丰富、预测分析过程自动化实现、预测结果准确性高。癌症易感性预测的优势是分析方法简单、数据获取容易、预测的准确度较高。癌症易感性预测不足为只对常见癌症进行了预测，缺少罕见疾病的易感性预测相关研究。

3 基于机器学习的癌症复发性预后

基于ML的癌症复发性预测，是指用机器学习方法预测癌症复发的可能性，对癌症患者的术后复发风险进行有效预测，得出癌症复发性的高低，为癌症的治疗提供一个参考。

3.1 机器学习与癌症复发性预测

ML方法可以对癌症的复发性进行预测。尽管

癌症在早期是可以预防和治愈的,但绝大多数患者的确诊时期较晚且治疗后容易复发,癌症复发性预测有助于寻求特定的方法进行癌症治疗,改善癌症预后情况。随着对癌症复发性的深入研究及机器学习技术的广泛使用,基于ML的癌症复发性预测方法逐渐受到关注。该方法对癌症患者的复发性数据进行学习,获得有效的机器学习模型,从而对癌症患者的复发性情况进行预测。癌症复发性预测是癌症预后的重要组成部分,本文对其相关研究进行了深入分析。

3.2 基于机器学习的癌症复发性预测研究现状

对PubMed上收录的过去20年间发表的癌症复发性预测相关研究进行统计,检索的关键词是“machine learning in cancer recurrence”(表3)。可

以发现,基于ML的癌症复发性预测相关研究在过去20年里呈现逐年增长趋势,增长速率在过去的5年中有了显著提升。其中,2017~2021(截至2021年6月22日)发表的基于ML的癌症复发性预测研究数是2012~2016年此类研究数的7倍,是2007~2011年间此类研究数目的26倍,是2002~2006年间此类研究数目的133倍。此外,与基于ML的癌症易感性预测的相关研究相比,基于ML的癌症复发性预测的研究数目明显较多,其近20年的总研究数约为癌症易感性预测相关研究的2.4倍,近5年的总研究数约为癌症易感性预测相关研究的2.8倍。总的来说,ML在癌症复发性预测方面的研究受到研究者的广泛关注,越来越多的预测方法被提了出来。

Table 3 Research status of cancer recurrence prediction based on machine learning (2002~2021)

表3 基于机器学习的癌症复发性预测研究情况 (2002~2021)

时间	第一年论文数	第二年论文数	第三年论文数	第四年论文数	第五年论文数	论文总数
2002~2006	0	2	0	0	1	3
2007~2011	3	1	4	2	5	15
2012~2016	2	5	9	14	26	56
2017~2021	26	65	105	139	>65	>400

3.3 基于机器学习的癌症复发性预测比较分析

进一步对基于ML的癌症复发性预测的相关研究进行分析(表4)。可以发现,癌症复发性预测相关的ML算法包括Autoencoder、SNF、KNN、RF、DNN、SVM、LASSO回归、COX回归、逻辑回归和ANN等^[49, 74-82]。其中,Autoencoder、DNN和ANN属于神经网络分类方法,它们均属于有监督学习方法,根据已有的癌症复发性数据对癌症患者复发性进行预测。LASSO回归、COX回归和逻辑回归均属于回归分析方法,LASSO回归分析用于筛选癌症复发性特征,COX回归分析用于确定预后特征和癌症复发性之间的相关性,逻辑回归用于对癌症复发性相关样本进行分类。SNF将各个癌症复发性相关的相似性网络构造成一个统一的网络,这个网络包含了各个网络中特征数据的共同点,可以用于癌症样本的复发性分析。KNN是有监督学习方法,根据离待预测样本最近的K个邻居的复发情况对该样本的复发性进行预测。RF是树结构的有监督学习方法,根据癌症复发性数据构造相关的森林结构,用于确定待测样本的复发性。上述方法在癌症易感性预测方面均有其对应的使用场

景,总的来说,SVM的应用最为广泛,其次是LASSO和COX,最后是一些其他的ML算法。

癌症复发性预测涉及多种癌症,包括前列腺癌、结直肠癌、鼻咽癌、肝癌、胶质母细胞瘤和肺癌等,都是一些致死率较高的疾病。前列腺癌是男性群体中第二大最常见的恶性肿瘤,也是全球第5大死亡原因^[83],对前列腺癌病因和致病风险因素的了解有助于对易患该疾病的高危男性进行预测,开发有效的筛查和预防方法。结直肠癌是世界范围内常见的恶性肿瘤,拥有较高的发病率和致死率,在确诊的结直肠癌患者中,晚期诊断比例较大,这对结直肠癌的防控提出了挑战^[84]。鼻咽癌是指发生于鼻咽腔顶部和侧壁的恶性肿瘤,在过去10年中,鼻咽癌的发病率和死亡率有所下降,这可能是生活方式的改变与医学治疗共同作用的结果^[85]。原发性肝癌是中国常见的恶性肿瘤,多种治疗方案如传统中医和现代西医已被广泛用于提高肝癌患者的生活质量、延缓癌症发展时间、延长患者的生存时间^[86]。胶质母细胞瘤属于最常见的脑部恶性肿瘤,属于胶质瘤中的一种,其治疗方法包括手术、放疗和化疗等^[87]。肺癌的发生和环境因素有

Table 4 Comparative analysis of cancer recurrence prediction based on machine learning
表4 基于机器学习的癌症复发性预测比较分析

机器学习方法	癌症类型	数据来源	预测性能	发表时间	PubMed
Autoencoder、SNF	前列腺癌	TCGA数据集	$AUC = 0.789$	2021	34064004 ^[49]
KNN、RF、DNN	前列腺癌	多中心国家数据库	$AUC=0.77$ (KNN), $AUC=0.74$ (RF), $AUC=0.84$ (DNN)	2021	33643910 ^[74]
SVM	结直肠癌	GEO数据集	$AUC=0.703$	2020	32150672 ^[75]
LASSO、COX	前列腺癌	TCGA数据集	$AUC=0.733$	2020	32117427 ^[76]
KNN、逻辑回归、RF	前列腺癌	圣约瑟夫医院收集的数据	$AUC=0.903$ (KNN), $AUC=0.924$ (RF), $AUC=0.940$ (逻辑回归)	2019	29969172 ^[77]
ANN、KNN、SVM	鼻咽癌	MRI图像	准确度=81.2% (ANN), 准确度=77.5% (KNN), 准确度=73.2% (SVM)	2018	30622931 ^[78]
LASSO、SVM、COX	肝癌	中山大学第一附属医院收集的数据集, TCGA数据集	一致性指数≈0.7	2017	28068175 ^[79]
SVM	胶质母细胞瘤	MRI图像	$AUC=0.84$, 灵敏度=91%, 特异性=93%	2016	26813856 ^[80]
SVM	肺癌	CT图像	准确度=81.3%~82.7%	2016	27429433 ^[81]
LASSO、SVM	肺癌	CT图像	$AUC=0.8\pm0.01$	2015	25979058 ^[82]

很大的相关性, 如吸烟和环境污染等会导致肺癌的发病率上升, 肺癌筛选有助于高风险患者的识别和相关预测工具的开发^[88]。

癌症复发性预测涉及到的数据来源包括公共数据和临床数据两种。公共数据是指从公开数据集获取的数据, 表4中涉及到的癌症复发性预测相关的公共数据集包括TCGA数据集、多中心国家数据库提供的数据集、GEO数据集。临床数据是指从医院或医学研究机构收集的数据, 表4涉及到的癌症复发性相关的临床数据集包括圣约瑟夫医院收集的数据集与中山大学第一附属医院收集的数据集。上述数据集包含的数据类型主要分为两种, 一是表达数据, 如TCGA和GEO数据集提供的表达数据; 二是癌症图像数据, 如MRI图像数据和CT图像数据。

癌症复发性预测的性能由其涉及的癌症类型、ML方法和分析的数据类型共同决定。如表4所示, 在预测前列腺癌的复发性时, 涉及到的ML方法有Autoencoder、SNF、KNN、RF、DNN、LASSO、COX和逻辑回归, 其预测的 AUC 在0.703~0.940之间, 最高 AUC 值是由逻辑回归方法得到的; 在预测鼻咽癌的复发性时, 涉及到的ML方法有ANN、KNN和SVM, 其预测的准确性分别为0.812、0.775和0.732; 在预测肝癌的复发性时, 涉及到的ML方法有LASSO、SVM和COX, 预测的一致性指数约为0.7; 在预测胶质母细胞瘤时, 用到的ML方法为SVM, 其 AUC 为0.84; 在预测肺癌的复发

性时, 用到的ML方法为SVM和LASSO, 预测的 AUC 在0.79~0.84之间。可以发现, 前列腺癌易感性预测的准确性较其他癌症高, 且准确性的差异主要是由预测的癌症类型和数据集的来源决定, 使用的ML方法对准确性预测的影响较前两者小。此外, 通过表4可以发现, 癌症复发性预测中最常用的分析方法是SVM, 然后是神经网络方法与回归分析方法, 最后其他的ML方法。综上, 癌症复发性预测的性能由癌症类型、数据来源和使用的分析方法共同决定, 癌症易感性预测在某些癌症如前列腺癌上有着十分准确的预测性能, 但总的来说, 其准确性较癌症易感性预测略逊一筹。

癌症复发性预测有助于寻求特定的方法进行癌症治疗, 有助于癌症预后情况的改善。癌症复发性预测的特点包括: 预测数据丰富、预测分析过程自动化实现、预测结果的准确性在可接受的范围内。癌症复发性预测的优势是分析方法简单且数据获取容易。癌症复发性预测不足为只对常见癌症进行了预测, 缺少罕见疾病的复发性预测相关研究, 且预测模型的准确性有待进一步提高。

4 基于机器学习的癌症生存期预测

基于ML的癌症生存期预测, 是指用机器学习方法预测癌症的生存结果, 对癌症患者的生存风险进行有效预测, 得出癌症患者生存风险的高低, 为癌症的治疗提供一个参考。

4.1 机器学习与癌症生存期预测

ML方法可以对癌症的生存期进行预测。癌症生存期预测具有重要意义：a.可以有效避免过度治疗，根据患者的预后情况为其提供适当和个性化的治疗方案；b.可以避免医疗资源的浪费，使得有限的医疗资源发挥更重大的治疗作用；c.可以改善癌症的预后，根据预后结果选择合适的治疗方案，使得预后情况得以提高。随着对癌症生存期的深入研究及ML技术的广泛使用，基于ML的癌症生存期预测方法逐渐受到关注。该方法对癌症患者的生存期数据进行学习，获得有效的ML模型，从而对癌症患者的生存情况进行预测。癌症生存期预测是癌症预后中的重要组成部分，本文对其相关研究进行了深入分析。

4.2 基于机器学习的癌症生存期预测研究现状

ML方法可以对癌症的生存期进行预测，对

PubMed上收录的过去20年间发表的癌症生存期预测相关研究进行统计，检索的关键词是“machine learning in cancer survival”（表5）。可以发现，此类研究在过去的20年中一直被研究人员所关注，且关注度在最近10年显著提升。在最近5年中，ML在癌症生存期预测方面的研究有了爆发式的增长。截至2021年6月22日，近5年此类研究的数目达到1442项之多，是2012~2016年间此类研究数的8.7倍，是2007~2011年间此类研究数的28.8倍，是2002~2006年间此类研究数的84.8倍。此外，ML在癌症生存期预测方面的研究明显较其在癌症易感性预测和癌症复发预测方面的研究数多，其在近20年的总研究数分别是前两者的3.5倍和8.5倍，而在近5年的总研究数分别是前两者的10.2倍和3.6倍。综上，ML在癌症生存期预测方面有着广泛的应用，且越来越受到研究者的关注。

Table 5 Research status of cancer survival prediction based on machine learning (2002–2021)

表5 基于机器学习的癌症生存期预测研究情况 (2002–2021)

时间	第一年论文数	第二年论文数	第三年论文数	第四年论文数	第五年论文数	论文总数
2002~2006	1	3	1	4	8	17
2007~2011	7	6	11	9	17	50
2012~2016	11	15	34	47	59	166
2017~2021	111	180	344	496	>311	>1442

4.3 基于机器学习的癌症生存期预测比较分析

进一步对基于ML的癌症生存期预测的相关研究进行分析（表6）可以发现，癌症生存期预测涉及到的机器学习算法包括SVM、COX回归、LASSO回归、梯度提升机（GBM）、RF、广义线性模型（GLM）、QuPath、逻辑回归、DNN、RSF、概率神经网络（PNN）、多层感知机（MLP）、径向基函数神经网络（RBFNN）及K-means等^[89-98]。SVM属于有监督学习方法，在进行生存期预测时，对已有的生存期预测数据进行学习，获得相关的学习模型，从而对新的癌症患者进行生存期预测。COX回归、LASSO回归、逻辑回归和GLM均属于回归分析方法，COX用于确定预后特征和癌症生存期之间的相关性，LASSO用于筛选癌症生存期相关的预后特征，逻辑回归用于对癌症生存期相关的样本进行分类，GLM用于构建预后特征与癌症生存期之间的线性模型。GBM属于集成学习算法，是基于梯度下降算法得到提升树模型。RF与RSF属于树结构的有监督的分类算法，

根据癌症生存期数据构造RF/RSF结构用于确定待测样本的生存期。MLP、DNN、PNN和RBFNN属于神经网络分类方法，它们根据已有的癌症生存期数据对新的癌症患者生存期进行预测。K-means属于无监督学习方法，通过对癌症患者的生存期数据进行聚类以获得生存期相关的预后结果。在上述癌症生存期相关的ML方法中，SVM、LASSO和COX有着最广泛的应用，其次是RF和神经网络，最后是其他ML方法。

癌症生存期预测涉及多种癌症，包括宫颈癌、肾透明细胞癌、尿路上皮癌、乳腺癌、大肠癌及TCGA中的癌症等，这些癌症大都有着较高的发病率和致死率。宫颈癌是常见的女性恶性肿瘤，具有较高的发病率和致死率，每年都有超过50万妇女被诊断出患有宫颈癌，且每年都会有超过30万人因宫颈癌死亡^[99]。肾透明细胞癌是一种常见的肾脏恶性肿瘤，拥有较高的发病率，有研究表明免疫检查点抑制剂可延长一部分转移性透明细胞肾细胞癌患者的生存期^[100]。尿路上皮癌是起源于尿路上

皮的一种多源性恶性肿瘤,可以分为非浸润性乳头状肿瘤、扁平病变和浸润性癌^[101]。大肠癌具体分为结肠癌和结直肠癌,其发病与生活方式和遗传等相关,在人群中有着较高的发病率和致死率^[102-103]。

Table 6 Comparative analysis of cancer survival prediction based on machine learning
表6 基于机器学习的癌症生存期预测比较分析

机器学习方法	癌症类型	数据来源	预测性能	发表时间	PubMed
SVM	宫颈癌	TCGA数据集	$AUC \approx 0.9515$	2021	34134623 ^[89]
COX+Lasso+SVM	肾透明细胞癌	Firebrowse数据集	准确度=92.3%, 灵敏度=92.7%, 特异性=91.9%, $AUC \approx 0.78$	2021	34058987 ^[90]
GBM, RF, COX, GLM	尿路上皮癌	IMvigor210数据集, IMvigor211数据集	$AUC=0.71$ (GBM), $AUC=0.7$ (RF), $AUC=0.69$ (CoxBoost), $AUC=0.69$ (GLM)	2021	33919237 ^[91]
QuPath	乳腺癌	TNBC数据集	风险指数=0.6~2.6	2020	33322174 ^[92]
ANN	泛癌	TCGA数据集集	一致性指数 $\approx 0.6 \sim 0.8$	2020	33311484 ^[93]
LASSO、SVM、RF、逻辑回归	乳腺癌	DCE-MRI图像	$AUC \approx 0.9$	2020	33289845 ^[94]
DNN	TCGA中的癌症	TCGA数据集	一致性指数 ≈ 0.6	2019	31865908 ^[95]
DNN、COX	宫颈癌	宫颈癌数据集	一致性指数=0.695~0.795	2019	30468921 ^[96]
RSF、COX	大肠癌	SEER数据集	准确度=0.898 (RSF), 准确度=0.732 (COX)	2018	30409119 ^[97]
PNN、MLP、SVM、RBFNN、K-means	宫颈癌	波兰热舒夫州立医院收集的数据	准确度=0.892, 灵敏度=0.975, $AUC=0.818$	2017	29233120 ^[98]

癌症生存期预测涉及到的数据来源包括公共数据与临床数据两种。公共数据是指从公开的数据集获取的数据,临床数据是指从医院或医学研究机构收集的数据。表6中涉及到癌症生存期预测的公共数据集有TCGA数据集、Firebrowse数据集、IMvigor210数据集、IMvigor211数据集、TNBC数据集、宫颈癌数据集和SEER数据集,涉及到癌症生存期相关的临床数据集包括波兰热舒夫州立医院收集的数据集。上述数据集涉及到的数据类型主要分为两种,一是表达数据,如TCGA数据集提供的表达数据;二是癌症图像数据,如DCE-MRI图像数据。

癌症生存期预测的性能由其涉及的癌症类型、ML方法和分析的数据类型共同决定。如表6所示,在预测宫颈癌的生存期时,涉及到的ML方法有SVM、DNN、COX、PNN、MLP和SVM,其预测的 AUC 在0.818~0.9515之间,最高 AUC 值是由SVM方法得到的,其预测的C-index在0.695~0.795之间,是由DNN和COX得到的;在预测肾透明细胞癌的生存期时,涉及到的ML方法有COX、Lasso和SVM,其 AUC 约为0.78;在预测尿路上皮癌的生存期时,涉及到的ML方法有GBM、RF、

COX和GLM,其预测的 AUC 在0.69~0.71之间,最高 AUC 值是由GBM方法得到的;在预测乳腺癌的生存期时,涉及到的ML方法有QuPath、ANN、LASSO、逻辑回归、SVM和RF,其预测的 AUC 值约为0.9;在预测大肠癌的生存期时,涉及到的ML方法有RSF和COX,其预测的C-index在0.695~0.795之间。通过上述分析可知,ML在癌症生存期预测方面有着良好的性能,且平均来说,要较ML对癌症复发性预测的性能更好。

癌症生存期预测可以有效避免过度治疗及医疗资源的浪费,且有助于癌症预后情况的改善。癌症生存期预测的特点包括:预测数据丰富、预测分析过程自动化实现、预测结果的准确性在可接受的范围内。癌症生存期预测的优势是分析方法简单且数据获取容易。癌症生存期预测不足为只对常见癌症进行了预测,缺少罕见疾病的生存期预测相关研究,且预测模型的准确性有待进一步提高。

5 总结与展望

癌症预后与多种因素相关,包括基因突变与患者临床特征等。常见预后相关的突变基因包括

TP53、KRAS、BRAF 和 PIK3CA 等,对癌症预后相关的突变基因进行分析,有助于癌症预后研究,并可以为癌症患者的治疗提供一个参考。癌症预后受多种临床特征影响,包括患者的性别、年龄及其肿瘤分期等。随着测序技术和计算机技术的发展,对癌症患者的预后情况进行自动化和个性化分析成为可能。癌症数据具有数据量大、模式复杂及个体表达特异等特点。ML 作为人工智能的一个分支,可以从一系列复杂的数据集中挖掘重要的模式,故可以应用到癌症数据的预后分析中。基于 ML 的癌症预后分析方法分为: a. 癌症易感性预测,即预测癌症发生的可能性; b. 癌症复发性预测,即预测癌症治疗后的复发可能; c. 癌症生存期预测,即预测癌症患者的生存结果。

癌症预后涉及多种 ML 方法。癌症易感性预测涉及到的 ML 方法包括 CNN、RNN、ICA、K-means、RF、SVM、DT、naïve Bayes、SSL co-training 及多分类方法等。癌症复发性预测涉及到的 ML 方法包括 Autoencoder、SNF、KNN、RF、DNN、SVM、LASSO 回归、COX 回归、逻辑回归及 ANN 等。癌症生存期预测涉及到的 ML 方法包括 SVM、COX、LASSO、GBM、RF、GLM、QuPath、逻辑回归、DNN、RSF、PNN、MLP、RBFNN 及 K-means 等。上述提到的方法中,最常用的是 COX 和 LASSO 回归分析方法,然后是 SVM 分类方法,其次是一些神经网络方法,最后是剩余的其他方法。癌症预后涉及到的数据来源有两种,公共数据集提供的数据和医院/医学研究机构收集的临床数据,公共数据样本量多但涉及到的癌症类型不全,临床数据样本量较少但记录的临床信息丰富。癌症预后涉及到的数据类型主要包括两种,表达数据和癌症图像数据。癌症预后性能由其涉及的预后种类、癌症类型、机器学习方法和分析的数据类型共同决定。总的来说,预测性能从高到低依次为:癌症易感性预测、癌症生存期预测、癌症复发性预测。癌症预后相关研究在过去 20 年间逐渐受到研究者的关注,尤其是近 5 年,基于 ML 的癌症预后研究方法迅速增长,研究数从多到少依次为:癌症生存期预测、癌症复发性预测、癌症易感性预测。

基于以上研究现状,在未来的基于 ML 的癌症预后分析中,应从以下几个方面进行探索与完善: a. 癌症预后分析覆盖的癌症类型不全面,当前研究大都基于一些常见癌症进行预后分析,对于其他具

有高致死率的癌症,也有必要将其纳入预后分析的范畴; b. 癌症易感性和复发性预后相关研究较癌症生存期预后相关研究少,有待对癌症易感性和复发性进行深入研究,为癌症的预防和治疗提供参考; c. 癌症预后数据信息利用不充分,目前用到的数据类型主要包括癌症的表达数据和癌症切片的图像数据,现有预后方法均是基于其中一种数据类型进行预后分析,但表达数据和图像数据在预后分析中各有优劣,在未来的研究中,应当对两种数据进行综合分析,使得预后判断更加准确; d. 癌症预后性能仍有可观的提升空间,尤其是癌症的生存期和复发性预测,未来的预后研究中,应从使用的 ML 方法出发,训练合适的预后模型,进行有效的预后分析。

参 考 文 献

- [1] Siegel R L, Miller K D, Jemal A. Cancer statistics, 2015. *CA Cancer J Clin*, 2015, **65**(1): 5-29
- [2] Bray F, Ferlay J, Soerjomataram I, *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, 2018, **68**(6): 394-424
- [3] Miller K D, Nogueira L, Mariotto A B, *et al.* Cancer treatment and survivorship statistics, 2019. *CA Cancer J Clin*, 2019, **69**(5): 363-385
- [4] Buscail L, Bournet B, Cordelier P. Role of oncogenic KRAS in the diagnosis, prognosis and treatment of pancreatic cancer. *Nat Rev Gastroenterol Hepatol*, 2020, **17**(3): 153-168
- [5] Petitjean A, Achatz M, Borresen-Dale A, *et al.* TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene*, 2007, **26**(15): 2157-2165
- [6] Xing M, Westra W H, Tufano R P, *et al.* BRAF mutation predicts a poorer clinical prognosis for papillary thyroid cancer. *J Clin Endocrinol Metab*, 2005, **90**(12): 6373-6379
- [7] Ogino S, Noshio K, Kirkner G J, *et al.* PIK3CA mutation is associated with poor prognosis among patients with curatively resected colon cancer. *J Clin Oncol*, 2009, **27**(9): 1477-1484
- [8] Fajkovic H, Halpern J A, Cha E K, *et al.* Impact of gender on bladder cancer incidence, staging, and prognosis. *World J Urol*, 2011, **29**(4): 457-463
- [9] Huang X, Shu C, Chen L, *et al.* Impact of sex, body mass index and initial pathologic diagnosis age on the incidence and prognosis of different types of cancer. *Oncol Rep*, 2018, **40**(3): 1359-1369
- [10] Salmerón D, Chirlaque M, Izarzugaza M I, *et al.* Lung cancer prognosis in Spain: the role of histology, age and sex. *Respir Med*, 2012, **106**(9): 1301-1308
- [11] Cruz J A, Wishart D S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*, 2006, **2**: 117693510600200030

- [12] Ferroni P, Zanzotto F M, Riondino S, *et al.* Breast cancer prognosis using a machine learning approach. *Cancers*, 2019, **11**(3): 328
- [13] Huang S, Yang J, Fong S, *et al.* Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. *Cancer Lett*, 2020, **471**: 61-71
- [14] Yue W, Wang Z, Chen H, *et al.* Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, 2018, **2**(2): 13
- [15] 上官艺, 王孟, 王春娟, 等. 基于机器学习的缺血性卒中功能预后预测模型研究. *中国卒中杂志*, 2021, **16**(9): 895-900
Shangguan Y, Wang M, Wang C J, *et al.* *Chinese Journal of Stroke*, 2021, **16**(9): 895-900
- [16] 殷子博. 基于决策树算法的癌症合成致死基因组合的预测及预后分析[D]. 南京: 南京邮电大学, 2020
Yin Z B. Prediction and Prognostic Analysis of Cancer Synthetic Lethal Gene Pairs Based on Decision Tree Algorithm [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2020
- [17] Dimitriou N, Arandjelović O, Harrison D J, *et al.* A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis. *NPJ Digit Med*, 2018, **1**: 52
- [18] Chen Y C, Ke W C, Chiu H W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Comput Biol Med*, 2014, **48**: 1-7
- [19] Murthy H N, Meenakshi M. ANN, SVM and KNN classifiers for prognosis of cardiac ischemia-a comparison. *Bonfring International Journal of Research in Communication Engineering*, 2015, **5**(2): 7-11
- [20] Ludwig J A, Weinstein J N. Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer*, 2005, **5**(11): 845-856
- [21] Brenner D E, Normolle D P. Biomarkers for cancer risk, early detection, and prognosis: the validation conundrum. *Cancer Epidemiol Biomarkers Prev*, 2007, **16**(10): 1918-1920
- [22] Kourou K, Exarchos T P, Exarchos K P, *et al.* Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*, 2015, **13**: 8-17
- [23] Zhu W, Xie L, Han J, *et al.* The application of deep learning in cancer prognosis prediction. *Cancers*, 2020, **12**(3): 603
- [24] Gao M, Guo Y, Xiao Y, *et al.* Comprehensive analyses of correlation and survival reveal informative lncRNA prognostic signatures in colon cancer. *World J Surg Oncol*, 2021, **19**(1): 104
- [25] Boeri C, Chiappa C, Galli F, *et al.* Machine learning techniques in breast cancer prognosis prediction: a primary evaluation. *Cancer Med*, 2020, **9**(9): 3234-3243
- [26] 段明月. 决策树模型在预测乳腺癌5年生存状况研究中的应用[D]. 长春: 吉林大学, 2020
Duan M Y. Application of Decision Tree Model in Predicting 5-year Survival of Breast Cancer[D]. Changchun: Jilin University, 2020
- [27] Muse E D, Barrett P M, Steinhubl S R, *et al.* Towards a smart medical home. *Lancet*, 2017, **389**(10067): 358
- [28] Winter A, Stäubert S, Ammon D, *et al.* Smart medical information technology for healthcare (SMITH). *Methods Inf Med*, 2018, **57**(S 01): e92-e105
- [29] Abdel-Basset M, Manogaran G, Gamal A, *et al.* A group decision making framework based on neutrosophic TOPSIS approach for smart medical device selection. *J Med Syst*, 2019, **43**(2): 38
- [30] Choi M K, Park O K, Choi C, *et al.* Cephalopod-inspired miniaturized suction cups for smart medical skin. *Adv Healthc Mater*, 2016, **5**(1): 80-87
- [31] Jiang D, Li W, Lv H. An energy-efficient cooperative multicast routing in multi-hop wireless networks for smart medical applications. *Neurocomputing*, 2017, **220**: 160-169
- [32] Bishop C M, Nasrabadi N M. *Pattern Recognition and Machine Learning*[M]. New York: Springer, 2006: 12-17
- [33] Jordan M I, Mitchell T M. Machine learning: trends, perspectives, and prospects. *Science*, 2015, **349**(6245): 255-260
- [34] Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of machine learning*. Massachusetts: MIT Press, 2018: 2-6
- [35] Harrell F E. *Cox proportional hazards regression model*//Frank E H Jr. *Regression Modeling Strategies*. Switzerland: Springer, Cham, 2015: 475-519
- [36] Kidd A C, McGettrick M, Tsim S, *et al.* Survival prediction in mesothelioma using a scalable Lasso regression model: instructions for use and initial performance using clinical predictors. *BMJ Open Respir Res*, 2018, **5**(1): e000240
- [37] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J Royal Statistical Society: Series B (Statistical Methodology)*, 2011, **73**(3): 273-282
- [38] Walczak S, Velanovich V. Improving prognosis and reducing decision regret for pancreatic cancer treatment using artificial neural networks. *Decis Support Syst*, 2018, **106**: 110-118
- [39] Mukherjee P, Zhou M, Lee E, *et al.* A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets. *Nat Mach Intell*, 2020, **2**(5): 274-282
- [40] Sun D, Wang M, Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans Comput Biol Bioinform*, 2018, **16**(3): 841-850
- [41] Sathyabama K, Saruladha K, Hemalatha M. *Smart Intelligent Computing and Applications*. Singapore: Springer Singapore, 2019: 315-323
- [42] Huang Z, Johnson T S, Han Z, *et al.* Deep learning-based cancer survival prognosis from RNA-seq data: approaches and evaluations. *BMC Med Genomics*, 2020, **13**(5): 41
- [43] Zhou J, Li L, Wang L, *et al.* Establishment of a SVM classifier to predict recurrence of ovarian cancer. *Mol Med Rep*, 2018, **18**(4): 3589-3598
- [44] Pradeep K, Naveen N. Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4. 5 and Naive Bayes algorithms for healthcare analytics. *Procedia Computer Science*, 2018, **132**: 412-420
- [45] Brims F J, Meniawy T M, Duffus I, *et al.* A novel clinical prediction

- model for prognosis in malignant pleural mesothelioma using decision tree analysis. *J Thoracic Oncol*, 2016, **11**(4): 573-582
- [46] Wang Q, Zhou Y, Ding W, *et al.* Random forest with self-paced bootstrap learning in lung cancer prognosis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2020, **16**(1s): 1-12
- [47] Jensen C, Carl J, Boesen L, *et al.* Assessment of prostate cancer prognostic Gleason grade group using zonal - specific features extracted from biparametric MRI using a KNN classifier. *J Appl Clin Med Phys*, 2019, **20**(2): 146-153
- [48] Kim J, Shin H. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *J Am Med Inform Asso*, 2013, **20**(4): 613-618
- [49] Wang T H, Lee C Y, Lee T Y, *et al.* Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion. *Cancers*, 2021, **13**(11): 2528
- [50] Liberti L, Lavor C, Maculan N, *et al.* Euclidean distance geometry and applications. *SIAM Rev*, 2014, **56**(1): 3-69
- [51] Bradbury A R, Olopade O I. Genetic susceptibility to breast cancer. *Rev Endocr Metab Disord*, 2007, **8**(3): 255-267
- [52] Mu L N, Lu Q Y, Yu S Z, *et al.* Green tea drinking and multigenetic index on the risk of stomach cancer in a Chinese population. *Int J Cancer*, 2005, **116**(6): 972-983
- [53] Chen J G, Zhang S W. Liver cancer epidemic in China: past, present and future. *Seminars in Cancer Biology*, 2011, **21**(1): 59-69
- [54] Breyer J P, Dorset D C, Clark T A, *et al.* An expressed retrogene of the master embryonic stem cell gene POU5F1 is associated with prostate cancer susceptibility. *Am J Hum Genet*, 2014, **94**(3): 395-404
- [55] Bao B Y, Pao J B, Huang C N, *et al.* Significant associations of prostate cancer susceptibility variants with survival in patients treated with androgen - deprivation therapy. *Int J Cancer*, 2012, **130**(4): 876-884
- [56] Mittal B, Tulsyan S, Kumar S, *et al.* Cytochrome P450 in cancer susceptibility and treatment. *Adv Clin Chem*, 2015, **71**: 77-139
- [57] Maron R C, Haggemüller S, Von Kalle C, *et al.* Robustness of convolutional neural networks in recognition of pigmented skin lesions. *Eur J Cancer*, 2021, **145**: 81-91
- [58] Siakallis L, Sudre C H, Mulholland P, *et al.* Longitudinal structural and perfusion MRI enhanced by machine learning outperforms standalone modalities and radiological expertise in high-grade glioma surveillance. *Neuroradiology*, 2021, **63**(12): 2047-2056
- [59] Jeong J, Wang L, Ji B, *et al.* Machine-learning based classification of glioblastoma using delta-radiomic features derived from dynamic susceptibility contrast enhanced magnetic resonance images: introduction. *Quant Imaging Med Surg*, 2019, **9**(7): 1201-1213
- [60] Choi K S, Choi S H, Jeong B. Prediction of IDH genotype in gliomas with dynamic susceptibility contrast perfusion MR imaging using an explainable recurrent neural network. *Neuro Oncol*, 2019, **21**(9): 1197-1209
- [61] Ji B, Wang S, Liu Z, *et al.* Revealing hemodynamic heterogeneity of gliomas based on signal profile features of dynamic susceptibility contrast-enhanced MRI. *NeuroImage*, 2019, **23**: 101864
- [62] Kim B-J, Kim S-H. Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. *Proc Natl Acad Sci USA*, 2018, **115**(6): 1322-1327
- [63] Piccolo S R, Andrulis I L, Cohen A L, *et al.* Gene-expression patterns in peripheral blood classify familial breast cancer susceptibility. *BMC Med Genomics*, 2015, **8**: 72
- [64] Shin H, Nam Y. A coupling approach of a predictor and a descriptor for breast cancer prognosis. *BMC Med Genomics*, 2014, **7**(Suppl 1): S4
- [65] Listgarten J, Damaraju S, Poulin B, *et al.* Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clin Cancer Res*, 2004, **10**(8): 2725-2737
- [66] Horie Y, Yoshio T, Aoyama K, *et al.* Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointest Endosc*, 2019, **89**(1): 25-32
- [67] Çınar M, Engin M, Engin E Z, *et al.* Early prostate cancer diagnosis by using artificial neural networks and support vector machines. *Expert Systems with Applications*, 2009, **36**(3): 6357-6361
- [68] Sherbet G V, Woo W L, Dlay S. Application of artificial intelligence-based technology in cancer management: a commentary on the deployment of artificial neural networks. *Anticancer Res*, 2018, **38**(12): 6607-6613
- [69] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*, 2018, **5**: 180161
- [70] Soenksen L R, Kassis T, Conover S T, *et al.* Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Sci Transl Med*, 2021, **13**(581): eabb3652
- [71] Chen R, Smith-Cohn M, Cohen A L, *et al.* Glioma subclassifications and their clinical significance. *Neurotherapeutics*, 2017, **14**(2): 284-297
- [72] Modrek A S, Bayin N S, Placantonakis D G. Brain stem cells as the cell of origin in glioma. *World J Stem Cells*, 2014, **6**(1): 43-52
- [73] Waks A G, Winer E P. Breast cancer treatment: a review. *JAMA*, 2019, **321**(3): 288-300
- [74] Sargos P, Leduc N, Giraud N, *et al.* Deep neural networks outperform the CAPRA score in predicting biochemical recurrence after prostatectomy. *Front Oncol*, 2021, **10**: 3237
- [75] Chen Q, Gao P, Song Y, *et al.* Predicting the effect of 5-fluorouracil-based adjuvant chemotherapy on colorectal cancer recurrence: a model using gene expression profiles. *Cancer Med*, 2020, **9**(9): 3043-3056
- [76] Cheng B, He Q, Cheng Y, *et al.* A three-gene classifier associated with microRNA-mediated regulation predicts prostate cancer recurrence after radical prostatectomy. *Front Genet*, 2020, **10**: 1402
- [77] Wong N C, Lam C, Patterson L, *et al.* Use of machine learning to predict early biochemical recurrence after robot-assisted

- prostatectomy. *BJU Int*, 2019, **123**(1): 51-57
- [78] Li S, Wang K, Hou Z, *et al.* Use of radiomics combined with machine learning method in the recurrence patterns after intensity-modulated radiotherapy for nasopharyngeal carcinoma: a preliminary study. *Front Oncol*, 2018, **8**: 648
- [79] Qiu J, Peng B, Tang Y, *et al.* CpG methylation signature predicts recurrence in early-stage hepatocellular carcinoma: results from a multicenter study. *J Clin Oncol*, 2017, **35**(7): 734-742
- [80] Akbari H, Macyszyn L, Da X, *et al.* Imaging surrogates of infiltration obtained via multiparametric imaging pattern analysis predict subsequent location of recurrence of glioblastoma. *Neurosurgery*, 2016, **78**(4): 572-580
- [81] Cirujeda P, Cid Y D, Müller H, *et al.* A 3-D Riesz-covariance texture model for prediction of nodule recurrence in lung CT. *IEEE Trans Med Imaging*, 2016, **35**(12): 2620-2630
- [82] Depeursinge A, Yanagawa M, Leung A N, *et al.* Predicting adenocarcinoma recurrence using computational texture models of nodule components in lung CT. *Med Phys*, 2015, **42**(4): 2054-2063
- [83] Rawla P. Epidemiology of prostate cancer. *World J Oncol*, 2019, **10**(2): 63-89
- [84] Mármol I, Sánchez-De-Diego C, Pradilla Dieste A, *et al.* Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. *Int J Mol Sci*, 2017, **18**(1): 197
- [85] Chua M L, Wee J T, Hui E P, *et al.* Nasopharyngeal carcinoma. *Lancet*, 2016, **387**(10022): 1012-1024
- [86] Ma L, Wang B, Long Y, *et al.* Effect of traditional Chinese medicine combined with Western therapy on primary hepatic carcinoma: a systematic review with meta-analysis. *Front Med*, 2017, **11**(2): 191-202
- [87] Wirsching H G, Galanis E, Weller M. Glioblastoma. *Handbook Clin Neurol*, 2016, **134**: 381-397
- [88] Barta J A, Powell C A, Wisnivesky J P. Global epidemiology of lung cancer. *Ann Global Health*, 2019, **85**(1): 8
- [89] Ding D, Lang T, Zou D, *et al.* Machine learning-based prediction of survival prognosis in cervical cancer. *BMC Bioinformatics*, 2021, **22**(1): 331
- [90] Dessie E Y, Tsai J J, Chang J G, *et al.* A novel miRNA-based classification model of risks and stages for clear cell renal cell carcinoma patients. *BMC Bioinformatics*, 2021, **22**(Suppl 10): 270
- [91] Abuhelwa A Y, Kichenadasse G, Mckinnon R A, *et al.* Machine learning for prediction of survival outcomes with immune-checkpoint inhibitors in urothelial cancer. *Cancers*, 2021, **13**(9): 2001
- [92] Millar E K, Browne L H, Beretov J, *et al.* Tumour stroma ratio assessment using digital image analysis predicts survival in triple negative and luminal breast cancer. *Cancers*, 2020, **12**(12): 3749
- [93] Qiu Y L, Zheng H, Devos A, *et al.* A meta-learning approach for genomic survival analysis. *Nat Commun*, 2020, **11**(1): 6350
- [94] Yu Y, Tan Y, Xie C, *et al.* Development and validation of a preoperative magnetic resonance imaging radiomics-based signature to predict axillary lymph node metastasis and disease-free survival in patients with early-stage breast cancer. *JAMA Network Open*, 2020, **3**(12): e2028086
- [95] Hao J, Kim Y, Mallavarapu T, *et al.* Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Med Genomics*, 2019, **12**(Suppl 10): 189
- [96] Matsuo K, Purushotham S, Jiang B, *et al.* Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *Am J Obstet Gynecol*, 2019, **220**(4): 381.e1-381.e14
- [97] Tian Y, Li J, Zhou T, *et al.* Spatially varying effects of predictors for the survival prediction of nonmetastatic colorectal cancer. *BMC Cancer*, 2018, **18**(1): 1084
- [98] Obrzut B, Kusy M, Semczuk A, *et al.* Prediction of 5-year overall survival in cervical cancer patients treated with radical hysterectomy using computational intelligence methods. *BMC Cancer*, 2017, **17**(1): 840
- [99] Cohen P A, Jhingran A, Oaknin A, *et al.* Cervical cancer. *Lancet*, 2019, **393**(10167): 169-182
- [100] Miao D, Margolis C A, Gao W, *et al.* Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science*, 2018, **359**(6377): 801-806
- [101] Rouprêt M, Babjuk M, Burger M, *et al.* European association of urology guidelines on upper urinary tract urothelial carcinoma: 2020 update. *Eur Urol*, 2021, **79**(1): 62-79
- [102] Siegel R L, Miller K D, Goding Sauer A, *et al.* Colorectal cancer statistics, 2020. *CA Cancer J Clin*, 2020, **70**(3): 145-164
- [103] Siegel R L, Miller K D, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin*, 2020, **70**(1): 7-30

Artificial Intelligence–based Prediction for Cancer Susceptibility, Recurrence and Survival*

GAO Mei-Hong, SHANG Xue-Qun**

(School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract Cancer has a high incidence and mortality and is a major threat to human health. Cancer prognosis analysis can effectively avoid excessive treatment and waste of medical resources and provide a scientific basis for medical staff and their families to make medical decisions. It has become a necessary condition for cancer research. With the rapid development of artificial intelligence technology and medical informatization in recent years, smart medicine has received widespread attention. It has become possible to analyze the prognosis of cancer patients automatically. As an essential part of smart medicine, cancer patients need to conduct effective intelligent prognostic analysis. This article reviews the existing machine learning-based cancer prognosis methods. Firstly, it provides an overview of machine learning and cancer prognosis, introduces cancer prognosis and related machine learning methods, and analyzes the application of machine learning in cancer prognosis. Then, it summarizes cancer prognosis methods based on machine learning, including cancer susceptibility prediction, cancer recurrence prediction, and cancer survival prediction, and sorts out their research status, cancer types and data sets involved, and machine learning methods used and prediction performance. Finally, the cancer prognosis methods are summarized and prospected, and the aspects that should be explored and improved are proposed: (1) include other high-fatal cancers in the prognostic analysis; (2) comprehensive analysis of cancer expression data and image data to improve prognostic performance; (3) optimize the prognostic model to improve prognostic performance.

Key words cancer prognosis analysis, artificial intelligence, smart medical, susceptibility prediction, recurrence prediction, survival prediction

DOI: 10.16476/j.pibb.2021.0334

* This work was supported by grants from The National Natural Science Foundation of China (61772426, U1811262).

** Corresponding author.

Tel: 86-13484907397, E-mail: shang@nwpu.edu.cn

Received: October 30, 2021 Accepted: February 8, 2022