

# 组学时代下机器学习方法在临床决策支持中的应用

赵学彤<sup>1,2</sup>, 杨亚东<sup>1,2</sup>, 渠鸿竹<sup>1,2</sup>, 方向东<sup>1,2</sup>

1. 中国科学院北京基因组研究所, 中国科学院基因组科学与信息重点实验室, 北京 100101

2. 中国科学院大学, 北京 100049

**摘要:** 随着组学技术的不断发展, 对于不同层次和类型的生物数据的获取方法日益成熟。在疾病诊治过程中会产生大量数据, 通过机器学习等人工智能方法解析复杂、多维、多尺度的疾病大数据, 构建临床决策支持工具, 辅助医生寻找快速且有效的疾病诊疗方案是非常必要的。在此过程中, 机器学习等人工智能方法的选择显得尤为重要。基于此, 本文首先从类型和算法角度对临床决策支持领域中常用的机器学习等方法进行简要综述, 分别介绍了支持向量机、逻辑回归、聚类算法、Bagging、随机森林和深度学习, 对机器学习等方法在临床决策支持中的应用做了相应总结和分类, 并对它们的优势和不足分别进行讨论和阐述, 为临床决策支持中机器学习等人工智能方法的选择提供有效参考。

**关键词:** 疾病; 机器学习; 人工智能; 临床决策支持

## Applications of machine learning in clinical decision support in the omic era

Xuetong Zhao<sup>1,2</sup>, Yadong Yang<sup>1,2</sup>, Hongzhu Qu<sup>1,2</sup>, Xiangdong Fang<sup>1,2</sup>

1. CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

2. University of Chinese Academy of Sciences, Beijing 100049, China

**Abstract:** With the development of the omic technologies, the acquisition approaches of various biological data on different levels and types are becoming more mature. As a large amount of data will be produced in the process of diagnosis and treatment of diseases, it is necessary to utilize the artificial intelligence such as machine learning to analyze complex, multi-dimensional and multi-scale data and to construct clinical decision support tools. It will provide a method to figure out rapid and effective programs in diagnosis and treatment. In this process, the choice of artificial intelligence seems to be particularly important, such as machine learning. The article reviews the type and algorithm of machine learning used in clinical decision support, such as support vector machines, logistic regression, clustering algorithms, Bagging, random

收稿日期: 2018-05-17; 修回日期: 2018-07-23

基金项目: 国家“精准医学研究”重点研发计划项目(编号: 2016YFC0901700, 2016YFC0901603, 2017YFC0907502, 2017YFC0908402, 2017YFC0907405)

资助[Supported by the National Key Research and Development Program of China (Nos. 2016YFC0901700, 2016YFC0901603, 2017YFC0907502, 2017YFC0908402, 2017YFC0907405)]

作者简介: 赵学彤, 博士研究生, 专业方向: 复杂疾病多组学数据整合与解析。E-mail: zhaoxuetong@big.ac.cn

通讯作者: 方向东, 博士, 研究员, 研究方向: 干细胞与复杂疾病临床组学与转化医学。E-mail: fangxd@big.ac.cn

DOI: 10.16288/j.ycz.18-139

网络出版时间: 2018/7/30 15:48:00

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20180730.1548.005.html>

forests and deep learning. The application of machine learning and other methods in clinical decision support has been summarized and classified. The advantages and disadvantages of machine learning are elaborated. It will provide a reference for the selection between machine learning and other artificial intelligence methods in clinical decision support.

**Keywords:** diseases; machine learning; artificial intelligence; clinical decision support

在高通量测序技术的快速发展下,涌现出了大量的多组学数据,如基因组、转录组、表观组、代谢组和蛋白组等,同时也出现了许多具有代表意义的国际项目,如 DNA 元件百科全书计划(The Encyclopedia of DNA Elements, ENCODE)<sup>[1]</sup>和国际人类基因组单体型图计划(Haplotype Map, HapMap)<sup>[2]</sup>等。随着数据的不断积累和基础研究的不断突破,人类对疾病的诊治迈入了精准医学时代。医生可以将患者的组学数据<sup>[3]</sup>、表型数据、临床诊疗数据、电子病历档案数据及影像数据等进行结合对疾病进行精准把控。基于此,众多优秀的临床决策支持系统纷纷涌现,使得以机器学习为代表的人工智能方法在基础医学研究领域迅速发展。利用机器学习等方法,从数据中生成模型,然后用大量数据来改善模型本身的性能,以达到对疾病预测和分类、用药指导、疾病诊断等目的,为临床决策支持提供技术基础。本文主要对临床决策支持领域中常用的机器学习方法进行综述,对机器学习方法的应用进行相应总结和分类,并提出不足,旨在为临床决策支持中机器学习等人工智能方法的选择提供参考。

## 1 机器学习的类型和算法

人工智能含有多个分支,包括计算机视觉、自然语言处理和机器学习等。计算机视觉<sup>[4]</sup>是通过自动提取和分析图像中的信息实现自动视觉理解。自然语言处理(natural-language processing, NLP)<sup>[5]</sup>属于计算机与人类(自然)语言之间的交互,特别是对计算机进行编程以成功处理大量自然语言数据。机器学习是人工智能的一个重要分支,一般是指利用统计方法求解最优化问题,通过学习输入数据的数据结构和其内在模式,选择对应的学习方式和训练方法以构建最优的数学模型,并不断调整模型参数,通过数学方法求解模型最优化后的反馈结果,以提高泛化能力防止发生过拟合。图 1 展示的是使用机器

学习方法进行数据处理的基本流程。

根据训练数据是否带有标记,机器学习分为监督学习、无监督学习、半监督学习和弱监督学习。监督学习一般是利用带标注的数据集学习出最优函数,再将新的数据输入此函数,根据函数预测出结果,常见的监督学习方法有回归分析和统计分类等<sup>[6]</sup>。无监督学习,其训练集不带标注,常见的无监督学习方法如聚类<sup>[7]</sup>。监督学习与无监督学习的结合体是半监督学习,半监督学习是在有监督的分类算法中加入无标签的数据来实现半监督分类,其目的是增强分类效果。弱监督学习是指标签含有的信息不完全、不确切或者不准确,但仍试图用这些较弱的标签构建模型。强化学习、集成学习和深度学习是近年来广受关注的机器学习方法。强化学习是指根据当前状态通过不断调整参数找到最优策略的一种方法,常用的算法有动态规划和蒙特卡罗方法等。集成学习是结合几个弱分类器,构建成一个强分类器的机器学习方法,集成学习的泛化误差小,并且不易发生过拟合,可用于处理不平衡分类问题和回归问题,但集成之后的效果未必会比单个分类器好。深度学习起源于人工神经网络,常见的深度学习如卷积神经网络和递归神经网络,其中使用了人工神经网络和人工神经单元的概念,是由多个处理层对数据进行高层抽象处理的算法<sup>[8]</sup>。

从机器学习算法看,常用的有监督学习算法有回归算法和分类算法,回归算法如线性回归、最小二乘法和贝叶斯回归等,分类算法如朴素贝叶斯、支持向量机(support vector machine, SVM)和最近邻算法等。回归和分类都是通过拟合得到最优函数再对输入的数据做出决策,但回归和分类在输入变量的类型、算法的目的和评价的方法上又有所区别:(1)回归的输出是定量的,是连续数据,而分类的输出是定性的,是离散数据;(2)回归的目的是寻找最优拟合,而分类则是为了找到决策边界;(3)回归的评价方法有拟合优度和误差平方和,分类则是靠精

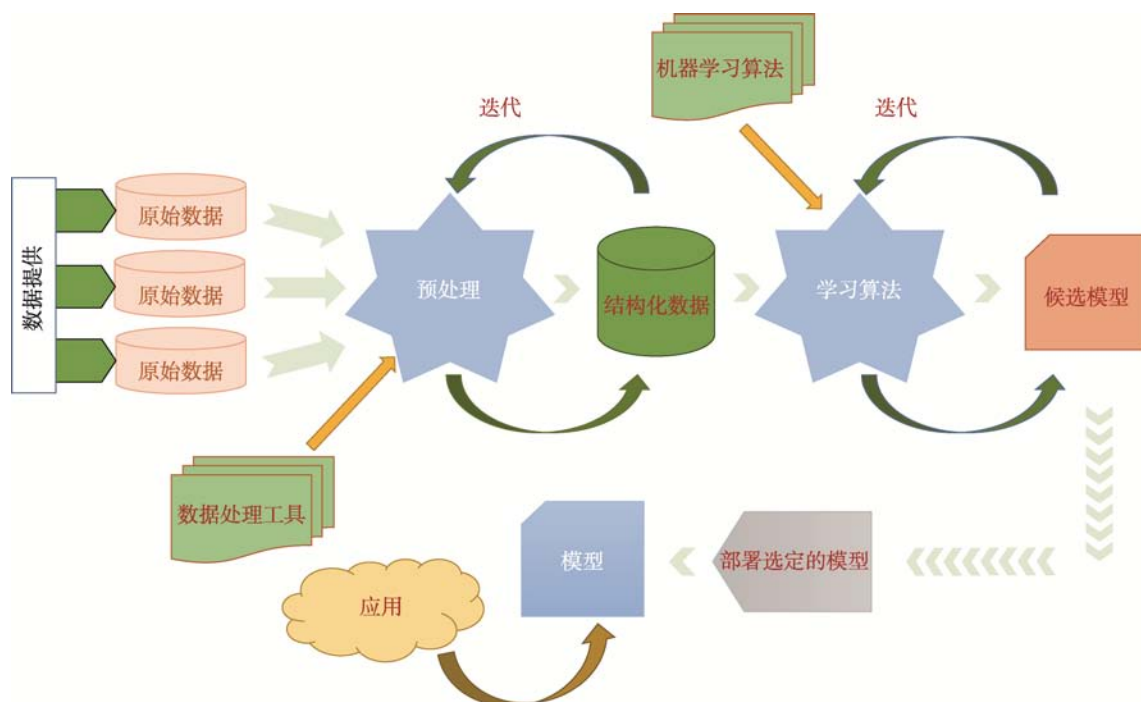


图1 使用机器学习方法进行数据处理的基本流程

Fig. 1 The basic flow of data processing using machine learning methods

度和混淆矩阵评价分类效果。常用的无监督学习算法有聚类,如层次聚类算法、 $K$ 均值算法和期望最大化算法(expectation maximization, EM)等。

应用机器学习等人工智能方法进行临床决策支持,一般是以疾病组和对照组的数据为输入,数据类型可以是基因表达数据<sup>[9, 10]</sup>、相关位点甲基化水平、单核苷酸多态性、插入缺失变异、结构变异信息或者医疗影像数据等其中的一种或几种。在进行数据挖掘之前,待输入的数据是以各种形式存在的,在一定程度上存在不一致性和不完整性,不能直接用于拟定的算法进行有意义的数据挖掘,所以要对数据进行前期预处理过滤不合格数据。随后,由于数据的样本量一般较少且变量较多,导致数据维度大,难以直接进行分析,所以会使用一些针对于特征选择减少变量的方法,如 LASSO (least absolute shrinkage and selection operator)<sup>[11]</sup>、岭回归(ridge regression)、桥回归(bridge regression)、弹性网络(elastic net)、主成分分析(principal component analysis, PCA)、偏最小二乘、等距映射、线性判别分析和独立成分分析等<sup>[12-16]</sup>剔除不相关或冗余的特征,从而实现减少特征个数、提高模型精确度并且有效

减少运行时间。之后,将筛选好的特征输入模型进行训练,构建临床决策支持模型。用于训练的数据称为训练集,同理有测试集和验证集。测试集和/或验证集的数据带入临床决策支持模型,对模型进行测试和验证。常见的评估模型好坏的指标是 AUC (area under curve)即 ROC 曲线(receiver operating characteristic Curve)下的面积<sup>[17]</sup>,面积值通常在 0.5 和 1.0 之间, AUC 越接近于 1.0,说明模型效果越好。在计算患者生存分析中,常用一致性指数,即 C-index (concordance index),计算 COX 模型预测值与真实值之间的区分度,用于表示肿瘤患者预后模型的预测精度<sup>[18]</sup>。

表 1 例举了目前应用于临床决策支持的几种常用机器学习方法。

## 2 机器学习等人工智能方法在临床决策支持中的研究进展

### 2.1 分类算法

常用于临床决策支持的分类算法主要有朴素贝

表 1 用于临床决策支持的机器学习方法

Table 1 Machine learning methods for clinical decision support

方法	描述	R 包	文献
SVM	支持向量机是监督学习模型，用于处理分类和回归分析问题。	e1071	[8, 19~23]
逻辑回归	逻辑回归常用于解决预测和判别问题，其因变量为二分类或多分类。	glm function	[24~26]
聚类分析	聚类是对一组对象进行分组，使得同一组中的对象彼此之间比其他组中的对象更相似。	mclust、cluster	[21, 27~29]
Bagging	Bagging 是一种集成学习方法，试图通过对随机样本的某些特征而不是整个特征集进行训练来减少集合中估计量之间的相关性。	ipred	[25, 30]
随机森林	随机森林是一种集成学习方法，常用于处理分类和回归问题。	randomforest、party	[31, 32]
深度学习	深度学习使用多层非线性处理单元的级联进行特征提取和转换，每个连续的图层使用前一层的输出作为输入。	nnet、AMORE、neuralnet、RSNNS	[33~36]

叶斯、SVM、决策树、逻辑回归和最近邻算法等，其中以 SVM 和逻辑回归方法最具有代表性。

### 2.1.1 SVM

SVM 是一种监督学习算法，常用于处理二分类预测、多分类预测和回归问题，其优化目标是找到一个最优分离超平面，使从数据点到超平面的最小距离之和最大化，距离超平面最近的训练样本则被称为支持向量。SVM 很容易发生核化，因此能够解决非线性分类问题，即将样本从原始样本空间通过核函数投影到高维特征空间，通过这个高维度特征空间使得原来线性不可分的数据变成线性可分。核函数的合适与否直接决定了支持向量机与核方法的性能好坏。常用核函数有线性核、多项式核、高斯核和 sigmoid 核，其中线性核和高斯核应用最为广泛。SVM 泛化能力优秀，结构化风险小，但其对空间消耗和时间消耗较大，当数据量增大时，SVM 的训练时间显著增加。SVM 方法已经成为近年来机器学习领域的常用方法，常应用于临床决策支持等精准医学数据挖掘问题。

在一项癌症检测和分类的研究中，Ramaswamy 等<sup>[8]</sup>使用从血小板中获取的 mRNA 进行多种肿瘤类型和健康供体的多类别癌症预测，使用 SVM / LOOCV 分类算法，多类癌症预测的结果显示出 71% 的平均准确度，证明血小板中的 mRNA 具有显著的多类别癌症鉴别能力。之后，他们选择了健康供体 ( $n=55$ ) 和在肺 ( $n=154$ )、脑 ( $n=114$ ) 以及肝脏 ( $n=127$ ) 中具有原发性或转移性肿瘤负荷的患者，并使用

SVM/LOOCV 算法确定肺、脑和肝脏是否存在癌症，分别得到 96%、91% 和 96% 的准确度。利用 SVM 方法，血小板中的 mRNA 能够较为准确地获知癌症所在位置，为癌症诊断应用提供了可靠的方法<sup>[21]</sup>。

CANScript 是一个基于 ROC 曲线下部分区域优化的支持向量机算法(即 SVMpAUC 学习算法)的功能性检测平台，可用于预测抗癌药物的临床疗效。SVMpAUC 是优化局部 AUC 的方法<sup>[23]</sup>，要求达到至少 75% 的特异性，即最多 25% 的假阳性率。研究人员使用 SVMpAUC 学习预测模型，用一组新数据测试 CANScript 模型，并使用临床数据加以验证，得到了很好的效果，在由 55 例患者组成的预测临床反应的测试集上达到 100% 的灵敏度和 91.67% 的特异性。CANScript 平台通过捕获肿瘤内异质性<sup>[23]</sup>，模拟患者的肿瘤微环境，衡量肿瘤在不同疗法下的应答，预测化疗方案的结果并提供有针对性的疾病诊断和治疗方案。其关键特点在于可以捕获肿瘤内异质性，辅助医生对疾病进行合理诊断。

### 2.1.2 逻辑回归

逻辑回归属于监督学习方法，常用于解决概率预测、二分类或多分类问题，其因变量的取值范围在 0~1 之间。逻辑回归最基本的学习算法是最大似然，通过最大似然推导损失函数，然后使用梯度下降、坐标轴下降等方法优化损失函数，其具有计算代价低、速度快、易于理解和实现等优点，但仅适用于解决线性问题。

在研究假基因的表达和临床相关的肿瘤亚型泛



癌分析中,为了评估假基因表达谱对于两种子宫内膜样腺癌组织学亚型的预测效力,研究人员使用多种机器学习算法,其中逻辑回归算法得到的假基因表达谱可以准确区分两种组织学亚型,其 AUC 达到了 0.892。在独立测试集上,逻辑回归算法展示出最佳性能,表明利用逻辑回归得到的假基因表达谱能够有效捕捉临床相关信息,获得有意义的肿瘤亚型,帮助医生和患者选择适当的临床治疗方案<sup>[24]</sup>。

在预测上皮性卵巢癌淋巴侵袭与患者生存情况相关性的研究中,使用 200 折交叉验证的逻辑回归、线性判别分析和支持向量机模型,利用组学标志物预测样本淋巴侵袭并建立预后模型。其中逻辑回归方法的 C-index 最大,可以很好地预测上皮性卵巢癌淋巴侵袭的分子特征,用于后续开发上皮性卵巢癌淋巴侵袭的预后模型<sup>[25]</sup>。

逻辑回归易于实现,并且对于线性可分的数据分类性能良好,适用于通过历史数据的表现对未来结果的发生概率进行预测。逻辑回归不仅能预测类别,还可以一并输出具体的概率值。

## 2.2 聚类分析

聚类是常见的无监督学习方法,指的是当训练样本无标记信息时,将训练样本划分为若干不相交的子集,每个子集被称为一个簇。划分依据的是组内距离最小化而组间距离最大化的原则。聚类的代表算法是  $K$  均值聚类、层级聚类、高斯混合聚类和密度聚类等。 $K$  均值聚类需要指定聚类数  $K$ ,对所有数据点寻找最近的随机初始化的  $K$  个聚类中心,采用贪心算法逐步更新聚类中心为所有类中点的均值,通过迭代优化求得最优解,但此算法聚类效果受  $K$  值约束,并且容易陷入局部最优。层次聚类分为“自下而上”和“自上而下”两种:“自下而上”是将每个样本各视为一类,根据两类之间的距离合并最近的类,再迭代计算各个类之间的距离,直到所有样本都聚到一类,可通过设置距离阈值决定是否继续合并类;“自上而下”则与之相反。层次聚类不需要指定最终聚类个数,但由于其距离度量的方法多种多样,聚类的效果受其影响较大,并且计算复杂度高。密度聚类是根据给定的距离半径和每类的最小个数,通过迭代更换类心形成各类,密度聚

类可以解决不规则形状的样本聚类问题,对于处理噪声数据有显著优势,但对于稀疏的高维数据聚类效果不理想。聚类分析已成为近几年来机器学习领域的研究热点,应用范围极其广阔,尤其生命科学领域应用聚类提供临床决策支持<sup>[28, 29]</sup>。

研究人员通过比较来自健康志愿者和分别来自非小细胞肺癌、结直肠癌、胶质母细胞瘤、胰腺癌、肝胆管癌和乳腺癌患者的血小板中的差异 mRNA,利用了无监督层级聚类的方法,成功区分出健康供体和患有特定类型癌症的患者,Fisher's 检验结果显示  $P < 0.0001$ ,证明了无监督层级聚类方法可以将健康供体和特定肿瘤类型患者区分开。研究人员利用此方法还得到了肿瘤特异性基因,这些基因可用于后续的特异性肿瘤模型的训练和验证<sup>[21, 27]</sup>。这为后续研究奠定了基础,也是疾病诊断决策过程中必不可少的环节之一。

## 2.3 集成学习

在临床决策支持中,集成学习方法也表现出色。常见的集成学习方法有:Bagging、随机森林、Voting、XGBoost、Stacking 和 Blending 等,但 Bagging 和随机森林方法使用率较高。下面主要介绍这两种方法。

### 2.3.1 Bagging

Bagging (bootstrap aggregating)是一种并行式的集成学习方法,其使用自助采样法训练基学习器,再将训练好的基学习器结合起来。Bagging 对于含有  $p$  个样本的数据集,先以随机的方式取出任意一个样本放入到采样集中,完成采样后,再把此样本放回到原始的数据集中,使得下次采样时此样本仍有可能被挑中,经过这样的  $p$  次随机采样操作,便可得到含有  $p$  个样本的采样集。在偏差-方差分解的角度上来看,Bagging 注重于降低方差,其在易受扰动学习器上效果将会更明显。

研究人员利用 2 186 例 TCGA 数据库中肺腺癌和鳞状细胞癌患者的癌组织图像和周围良性组织图像<sup>[37]</sup>,294 例来自斯坦福组织芯片数据库的肿瘤影像,结合每例肿瘤的级别和每例患者的存活时间等信息<sup>[38]</sup>,对肿瘤患者进行诊断预测。研究者使用了多种机器学习方法相互比较,Bagging 分类树可使得

到的癌症特异性特征用于准确区分恶性肿瘤与邻近致密正常组织。这些癌症特异性特征包括肿瘤细胞的大小和形状以及与相邻肿瘤细胞的空间关系等,模型的 AUC 达到了 0.87。Bagging 分类树还可以用于区分肺腺癌的组织病理学图像与肺鳞状细胞癌的图像,其 AUC 达到了 0.75<sup>[39]</sup>。利用 Bagging 分类树识别肿瘤影像评估肺癌组织切片比病理学家更加准确,对于区分肺癌中的腺癌与鳞癌有重要的临床意义,此方法也可以扩展到其他器官组织病理学图像分析中<sup>[30, 39]</sup>。

### 2.3.2 随机森林

随机森林(random forests)是 Bagging 的变体之一,是在以决策树构建 Bagging 的基础之上,在树的训练中引入了随机属性选择。一个随机森林模型集成了多棵决策树<sup>[40]</sup>。随机森林分类表现优秀、扩展性好且使用方法简单。

在肝癌的早期筛查模型中,研究人员利用肝癌患者和正常人血液样本中的 DNA 甲基化数据和生存数据,通过随机森林和 LASSO 等机器学习方法,得到了用于肝癌早期筛查、风险评估和预后监测的模型<sup>[31]</sup>。在随机森林分析中,利用 OOB (out of bag) 误差<sup>[41]</sup>作为最小化准则,从变量森林中进行变量消去,通过设置变量每次迭代的下降分数为 0.3,将变量从随机森林中缩减,最终从 450 000 个 DNA 甲基化位点中筛选出了 10 个甲基化位点作为生物标志物,进而使用逻辑回归构建了肝癌诊断预测模型,辅助医生进行临床决策。此模型还可用于区分不同的肝癌和肝脏疾病,如脂肪肝和酒精肝等,这些肝脏疾病都是肝癌的主要风险因素<sup>[31]</sup>。

此外,随机森林还可用于辅助医生进行慢性疾病的诊断及其风险评估。研究人员使用宏基因组测序方法,将 405 份粪便样本进行了微生物组成和生物信息学计算分析,包括宏基因组连锁群(metagenomic linkage groups, MLGs)聚类分析和随机森林,总共筛选得到 47 个 MLGs 标志物,这些标志物可用于区分冠心病患者和健康人,利用该标志物绘制出的 ROC 曲线可以用来评判标志物的分类能力<sup>[32]</sup>。同时研究人员将这些标志物与已有的标志物的分类能力进行了比较<sup>[42]</sup>,发现这 47 个生物标志物的分类能力

要高于公认的粥样硬化性心血管疾病标志物三甲胺的分类能力。这些新发现的生物标志物很可能发展成为未来冠心病的诊治标志物。研究者同时还使用随机森林研究了治疗粥样硬化性心血管疾病的药物<sup>[43]</sup>。在区分有无药物治疗的动脉粥样硬化性心血管疾病患者和健康对照中,随机森林分类器的分类效果最好;当两种药物一起分析时,随机森林分类器同样达到较高水平。结果表明,使用随机森林算法得到的显著富集在疾病组的宏基因连锁群,在一定程度上可以作为诊断和预防粥样硬化性心血管疾病的生物标志物<sup>[32]</sup>。

### 2.4 深度学习

深度学习是指通过基于非监督或有/半监督的特征学习和分层特征提取等算法获取特征的一种机器学习方法,其包含一个输入层、一个输出层和一个或多个隐含层,属于深层的神经网络。深度学习的输入层接收输入数据,将输入传递给第一个隐含层,隐含层会针对其输入进行数学运算,在隐含层内部进行传输,最后一个隐含层会将输出结果传递给输出层,输出层输出最终数据。其中每层都有若干的神经元,每两个神经元之间都有对应阈值,其决定输入数据的重要性。每个神经元之间又有一个激励函数,激励函数用于引入非线性因素,常见的激励函数有 sigmoid 函数、relu 函数和 tanh 函数等。深度学习至今已有数种学习框架,如 TensorFlow、Caffe、Deeplearning4j 和 Theano 等,这些框架被广泛应用于图像处理、自然语言处理和生物信息学等领域,取得很好的效果<sup>[44]</sup>。

常见的深度学习模型有卷积神经网络(convolutional neural network, CNN)、递归神经网络(recurrent neural network, RNN)、自编码器(auto-encoder, AE)、生成对抗网络(generative adversarial networks, GAN)和受限玻尔兹曼机(restricted boltzmann machines, RBM)等。CNN 是通过卷积、池化(pooling)和全连通过程将高维数据压缩,常用于处理结构化数据或网格状数据,例如图像数据;常见的 RNN 如长短时记忆网络(long short-term memory, LSTM),分为前向传播和反向传播,使用 3 个门控开关控制单元状态,适用于处理有序列问题,例如时间序列和语

音分析等;AE通过编码器和解码器将数据进行低维压缩,常用于提取数据特征和降低数据维度等;GAN包含两个训练网络,分别为生成网络和判别网络,使用无监督学习方法同时对其进行训练,分别捕捉样本数据的分布和样本来自训练集的概率,常用于处理图像数据和音频视频等;RBM属于生成式随机神经网络,包含可见单元和隐藏单元,分别用于描述数据和提取特征,常使用最大似然法学习目标,用于进行预测分析等。CNN、AE和DBN在癌症检测中最为常用,它们常用于分析图像数据(如X光片、CT图像)和分子数据(如基因突变、基因表达数据)等。

2017年2月,美国斯坦福大学的研究人员成功地利用深度学习进行了黑色素瘤皮肤病的诊断预测。他们的模型在学习了129 450张皮肤图像之后,对黑色素瘤皮肤病的诊断准确度超过94%,击败了21位专业的皮肤病学家<sup>[33]</sup>。另有研究人员利用CNN检测乳腺癌的发生,使用45 000多张乳房X光图像训练CNN模型,使模型的诊断准确度接近于人类专家的水平<sup>[34]</sup>。同样使用CNN模型,谷歌团队利用计算机视觉从100万像素的组织显微图像中检测并定位出100×100像素的肿瘤,其灵敏度高达92.4%,并且每张图片的假阳性结果平均只有8个,实现了乳腺癌检测领域的又一突破<sup>[35]</sup>。研究人员还使用多层CNN实现了从胸部CT扫描图像中识别肺结节,此模型的准确度达到了86.84%<sup>[36]</sup>。之后,研究人员还另外设计了3D-CNN,其能使计算机通过分析结肠镜拍摄到的视频找到肠息肉所在位置<sup>[45]</sup>。

研究者们还推出"DeepPatient"系统<sup>[46]</sup>,使用无监督深层特征学习方法,利用自然语言处理和3层去噪自编码器捕捉电子病历中的信息,随后使用随机森林分类器实现对每种疾病的诊断分类,实现了读入一份电子病历就可以预测出病历主人一年后的健康情况。研究人员使用76 214个来自不同临床领域和时间阶段的78种疾病的测试病人进行评估,在糖尿病、精神分裂症和多种癌症中的预测性能较好。在结直肠癌的诊断中,其准确率超过了88%,但对于一些复杂疾病,预测准确性还有待提高。这些结果表明将深度学习方法应用于电子健康档案数据中可辅助临床诊断预测。

Watson for Oncology<sup>[47]</sup>使用自然语言处理 and 高级认知算法收集了约1500万份医学资料,例如医学文献、指南、电子病历和患者数据等,其可在数分钟内生成患者的病历报告并实现数据可视化,辅助医生快速识别病历中的关键信息,提供相应证据并给出对应治疗方案供医生参考。Watson现已应用于淋巴瘤、黑色素瘤、胰腺癌、卵巢癌、脑癌、肺癌<sup>[48]</sup>、乳腺癌<sup>[49]</sup>和结直肠癌等疾病。

IDx-DR<sup>[50]</sup>系统利用人工智能技术读取了100万张眼部图像学习视网膜病变症状,再次输入视网膜图像即可检测糖尿病视网膜病变,支持眼科专家临床决策,对患者进行早期干预,甚至可独立诊断眼病,实现眼科疾病的实时检测。

另外,美国心脏病学院与美国心脏病协会(ACC/AHA)机构经过多年的研究,将胆固醇、高血压、吸烟、年龄和糖尿病等一系列因素列为心血管疾病的高风险因素,并推出了ACC/AHA诊断预测模型<sup>[51,52]</sup>。在这些建立的心血管病风险评估模型中,每个风险因素与心血管疾病之间的关系都是线性的,这简化了疾病与风险因素之间的关系。为寻找更为合适的机器学习算法,科研人员利用4种机器学习算法——随机森林、逻辑回归、梯度提升和神经网络,通过分析295 267份电子病历数据构建了关于心血管疾病的预测模型。研究人员将经过标准化后的验证数据输入到新建立的四个模型中,得到了与心血管疾病相关性较高的风险因子。之后,研究者将使用这四种机器学习算法建立的模型与ACC/AHA模型进行比较,分别预测82 989名居民10年后会患心血管疾病的人员名单,结果表明随机森林模型、逻辑回归模型、梯度提升机和神经网络模型的敏感度比ACC/AHA模型高出2.6%、4.4%、4.8%和4.8%<sup>[53]</sup>。然而,即使引入机器学习算法还是有三成左右的高风险居民没有被预测出来,这表明深度学习方法还需要进行大幅提升。

### 3 结语与展望

机器学习等人工智能方法在临床决策支持中的应用目前主要分为三个方面:基于医学影像数据<sup>[33-35]</sup>的图像识别、基于电子病历数据<sup>[46]</sup>的数据整



理和基于组学数据的数据挖掘<sup>[24, 27, 31]</sup>。基于医学影像数据,如利用获得的大量患者的 X 光片、视网膜图片、结节图片、皮肤病图片等输入到人工智能模型中,以训练其对病变组织进行筛查诊断的能力,之后对于新的数据,模型能够提供一定的诊断信息。基于电子病历数据,如利用深度学习加权的循环神经网络等捕捉病历信息,之后利用机器学习等人工智能方法对病历主人的患病情况做出一定预测,给出适宜的治疗方案。基于组学数据,如将筛选得到差异标志物进行特征选择,之后选择合适的机器学习模型,用于辅助医生进行疾病的诊断、预后模型的评估等。

从已有的数据角度来看,对于带标签的数据进行分类和回归分析,可使用有监督学习方法,例如对多种肿瘤类型和健康供体的多类别癌症进行辅助诊断<sup>[8]</sup>;对于不带标签的数据进行聚类,可以使用无监督学习方法,例如对患者的血小板中的差异 mRNA 使用无监督聚类方法,区分出是否为患者<sup>[27]</sup>;对于既含有带标签数据也含有不带标签数据时,可以使用半监督学习方法,例如在计算机辅助医学图像分析中,图像数据只有一部分是带有标签的;当使用单一机器学习方法对数据的建模效果不好时,可使用集成学习方法。在众多优秀的机器学习方法中,值得一提的是最近火热的深度学习,深度学习并不是一种独立的学习方法,它本身会用到有监督和无监督的学习方法来训练深度神经网络,如黑色素瘤的诊断预测<sup>[33]</sup>,疾病的类别预测<sup>[21]</sup>和患者的预后生存预测<sup>[31, 53]</sup>等。深度学习的盛行并不意味着机器学习的终结,相较于传统算法,深度学习效率高,模型可塑性高,还可以根据问题自动建立模型,具有一定的普适性。但其训练成本高,所需数据量大,在临床决策支持中遇到的并不都是大样本,当小样本量时深度学习效果不佳,容易发生过拟合,反而简单的机器学习可以解决问题;深度学习的模型训练时间长,验证模型的正确性较为复杂,随着深度的增加也会使得非凸的目标函数产生局部最优解;深度学习目前能较好的处理图像、语音数据,但缺乏一定的反馈机制,需要加入强化学习、迁移学习才能更好的解决相应问题。

在生物医学领域,机器学习等人工智能方法虽

然解决了很多问题,但疾病发生发展的过程十分复杂,目前还有许多问题需要我们进行不断探索和方法上的优化,比如模型的精确度有待提高、泛化能力还需要加强、模型的可解释性偏弱等,因此,利用人工智能方法实现临床决策支持任重道远。

## 参考文献(References):

- [1] Qu HZ, Fang XD. A brief review on the human encyclopedia of DNA Elements (ENCODE) project. *Genomics Prot Bioinform*, 2013, 11(3): 135–141. [DOI]
- [2] Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarroll SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghorji MJ, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. Integrating common and rare genetic variation in diverse human populations. *Nature*, 2010, 467(7311): 52–58. [DOI]
- [3] Xie BB, Yang YD, Ding N, Fang XD. Identification of disease targets for precision medicine by integrative analysis of multi-omics data. *Hereditas (Beijing)*, 2015, 37(7): 655–663.  
谢兵兵, 杨亚东, 丁楠, 方向东. 整合分析多组学数据筛选疾病靶点的精准医学策略. *遗传*, 2015, 37(7): 655–663. [DOI]
- [4] Goldberg Y. A primer on neural network models for natural language processing. *Comput Sci*, 2015, <https://arxiv.org/abs/1510.00726>. [DOI]
- [5] Szeliski R. Computer vision: algorithms and applications. Heidelberg: Springer-Verlag Berlin, 2010. [DOI]
- [6] Russell SJ, Norvig PN. Artificial intelligence: A modern approach. Prentice Hall. *Appl Mech Mater*, 1995, 263(5):



- 2829–2833. [DOI]
- [7] Kim YS, Street WN, Menczer F. Evolutionary model selection in unsupervised learning. *Intell Data Anal*, 2002, 6(6): 531–556. [DOI]
- [8] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci USA*, 2001, 98(26): 15149–15154. [DOI]
- [9] Zhang T, Yang YD, Fang XD. Transcriptome alternative splicing analysis applied to precision medical research. *J Dev Med*, 2016, 4(2): 78–84.  
张韬, 杨亚东, 方向东. 应用于精准医学研究的转录组可变剪接分析. 发育医学电子杂志, 2016, 4(2): 78–84. [DOI]
- [10] Li P, He N, Li YM, Fang XD. Progress in single cell transcriptome analysis and developmental biology. *J Dev Med*, 2017, 5(1): 28–34.  
李品, 贺宁, 李艳明, 方向东. 单细胞转录组数据分析与发育生物学研究进展. 发育医学电子杂志, 2017, 5(1): 28–34. [DOI]
- [11] Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J R Stat Soc Series B Stat Methodol*, 2010, 72(1): 3–25. [DOI]
- [12] Isabelle G, André E. An introduction to variable and feature selection. *J Mach Learn Res*, 2003, 3(7–8): 1157–1182. [DOI]
- [13] Awada W, Khoshgoftaar TM, Dittman D, Wald R, Napolitano A. A review of the stability of feature selection techniques for bioinformatics data. Proceedings of the 2012 IEEE 13th International Conference on Information Reuse and Integration, IRI 2012, 2012: 356–363. [DOI]
- [14] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2007, 23(19): 2507–2517. [DOI]
- [15] Jiang Y, Shi X, Zhao Q, Krauthammer M, Rothberg BEG, Ma S. Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. *Genomics*, 2016, 107(6): 223–230. [DOI]
- [16] Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Brief Bioinform*, 2008, 9(5): 392–403. [DOI]
- [17] Fogarty J, Baker RS, Hudson SE. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface*, 2005. [DOI]
- [18] Raykar V, Steck H, Krishnapuram B, Dehing-Oberije C, Lambin P. On ranking in survival analysis: bounds on the concordance index. *Adv Neural Inform Proc Sys*, 20: 1209–1216. [DOI]
- [19] Cortes C, Vapnik V. Support-vector networks. *Mach Learn*, 1995, 20(3): 273–297. [DOI]
- [20] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 2000, 16(10): 906–914. [DOI]
- [21] Best Myron G, Sol N, Kooi I, Tannous J, Westerman Bart A, Rustenburg F, Schellen P, Verschueren H, Post E, Koster J, Ylstra B, Ameziane N, Dorsman J, Smit Egbert F, Verheul Henk M, Noske David P, Reijneveld Jaap C, Nilsson RJonas A, Tannous Bakhos A, Wesseling P, Wurdinger T. RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Canc Cell*, 2015, 28(5): 666–676. [DOI]
- [22] Narasimhan H, Agarwal S. SVM pAUC tight: a new support vector method for optimizing partial AUC based on a tight convex upper bound. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013: 167–175. [DOI]
- [23] Majumder B, Baraneedharan U, Thiagarajan S, Radhakrishnan P, Narasimhan H, Dhandapani M, Brijwani N, Pinto DD, Prasath A, Shanthappa BU, Thayakumar A, Surendran R, Babu GK, Shenoy AM, Kuriakose MA, Bergthold G, Horowitz P, Loda M, Beroukhim R, Agarwal S, Sengupta S, Sundaram M, Majumder PK. Predicting clinical response to anticancer drugs using an ex vivo platform that captures tumour heterogeneity. *Nat Commun*, 2015, 6: 6169. [DOI]
- [24] Han L, Yuan Y, Zheng S, Yang Y, Li J, Edgerton ME, Diao L, Xu Y, Verhaak RGW, Liang H. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat Commun*, 2014, 5: 3963. [DOI]
- [25] Paik ES, Choi HJ, Kim TJ, Lee JW, Kim BG, Bae DS, Choi CH. Molecular signature for lymphatic invasion associated with survival of epithelial ovarian cancer. *Cancer Res and Treat*, 2017, 50(2): 461–473. [DOI]
- [26] Cai X, Han X, Zhang S, Luo Y, Chen Y, Ji L. Age at diagnosis and C-peptide level are associated with diabetic retinopathy in Chinese. *PLoS One*, 2014, 9(3): e91174. [DOI]
- [27] Best MG, Sol N, In 't Veld SGJG, Vancura A, Muller M, Niemeijer A-LN, Fejes AV, Tjon Kon Fat L-A, Huis In 't Veld AE, Leurs C, Le Large TY, Meijer LL, Kooi IE, Rustenburg F, Schellen P, Verschueren H, Post E,

- Wedekind LE, Bracht J, Esenkbrink M, Wils L, Favaro F, Schoonhoven JD, Tannous J, Meijers-Heijboer H, Kazemier G, Giovannetti E, Reijneveld JC, Idema S, Killestein J, Heger M, de Jager SC, Urbanus RT, Hoefer IE, Pasterkamp G, Mannhalter C, Gomez-Arroyo J, Bogaard H-J, Noske DP, Vandertop WP, van den Broek D, Ylstra B, Nilsson RJA, Wesseling P, Karachaliou N, Rosell R, Lee-Lewandrowski E, Lewandrowski KB, Tannous BA, de Langen AJ, Smit EF, van den Heuvel MM, Wurdinger T. Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets. *Cancer Cell*, 2017, 32(2): 238–252.e9. [DOI]
- [28] Yu ZW, Chen HT, You J, Liu JM, Wong HS, Han G, Li L. Adaptive fuzzy consensus clustering framework for clustering analysis of cancer data. *IEEE/ACM Trans Comput Biol Bioinform*, 2015, 12(4): 887–901. [DOI]
- [29] Mahata P. Exploratory consensus of hierarchical clusterings for melanoma and breast cancer. *IEEE/ACM Trans Comput Biol Bioinform*, 2010, 7(1): 138–152. [DOI]
- [30] Wahdan P, Saad A, Shoukry A. Automated breast tumour detection in ultrasound images using support vector machine and ensemble classification. *J Biomed Eng Biosci*, 2016, 3. [DOI]
- [31] Xu RH, Wei W, Krawczyk M, Wang W, Luo H, Flagg K, Yi S, Shi W, Quan Q, Li K, Zheng L, Zhang H, Caughey BA, Zhao Q, Hou J, Zhang R, Xu Y, Cai H, Li G, Hou R, Zhong Z, Lin D, Fu X, Zhu J, Duan Y, Yu M, Ying B, Zhang W, Wang J, Zhang E, Zhang C, Li O, Guo R, Carter H, Zhu JK, Hao X, Zhang K. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat Mater*, 2017, 16(11): 1155–1161. [DOI]
- [32] Jie Z, Xia H, Zhong SL, Feng Q, Li S, Liang S, Zhong H, Liu Z, Gao Y, Zhao H, Zhang D, Su Z, Fang Z, Lan Z, Li J, Xiao L, Li J, Li R, Li X, Li F, Ren H, Huang Y, Peng Y, Li G, Wen B, Dong B, Chen JY, Geng QS, Zhang ZW, Yang H, Wang J, Wang J, Zhang X, Madsen L, Brix S, Ning G, Xu X, Liu X, Hou Y, Jia H, He K, Kristiansen K. The gut microbiome in atherosclerotic cardiovascular disease. *Nat Commun*, 2017, 8(1): 845. [DOI]
- [33] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017, 542(7660): 686. [DOI]
- [34] Kooi T, Litjens G, Van GB, Gubern-Mérida A, Sánchez CI, Mann R, den HA, Karssemeijer N. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal*, 2017, 35: 303–312. [DOI]
- [35] Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, Venugopalan S, Timofeev A, Nelson PQ, Corrado GS, Hipp JD, Peng LL, Stumpe MC. Detecting cancer metastases on gigapixel pathology images. 2017, <https://arxiv.org/abs/1703.02442>. [DOI]
- [36] Shen W, Zhou M, Yang F, Yang C, Tian J. Multi-scale convolutional neural networks for lung nodule classification. *Inf Process Med Imaging*, 2015, 24: 588–599. [DOI]
- [37] Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 2014, 511(7511): 543–550. [DOI]
- [38] Marinelli RJ, Montgomery K, Liu CL, Shah NH, Prapong W, Nitzberg M, Zachariah ZK, Sherlock GJ, Natkunam Y, West RB, van de Rijn M, Brown PO, Ball CA. The stanford tissue microarray database. *Nucleic Acids Research*, 2008, 36(Database issue): D871–D877. [DOI]
- [39] Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, Snyder M. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun*, 2016, 7: 12474. [DOI]
- [40] Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinf*, 2008, 9: 307. [DOI]
- [41] Diaz-Uriarte R. GeneSrf and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinf*, 2007, 8(1): 328. [DOI]
- [42] EKrishna SM, Moxon JV, Golledge J. A review of the pathophysiology and potential biomarkers for peripheral artery disease. *Int J Mol Sci*, 2015, 16(5): 11294–22322. [DOI]
- [43] Haiser HJ, Gootenberg DB, Chatman K, Sirasani G, Balskus EP, Turnbaugh PJ. Predicting and manipulating cardiac drug inactivation by the human gut bacterium *eggerthella lenta*. *Science*, 2014, 341(6143): 295–298. [DOI]
- [44] Charalampous K, Gasteratos A. On-line deep learning method for action recognition. *Patt Anal Appl*, 2016, 19(2): 337–354. [DOI]
- [45] Dou Q, Chen H, Yu L, Zhao L, Qin J, Wang D, Mok V, Shi L, Heng PA. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Trans Med Imaging*, 2016, 35(5): 1182–1195. [DOI]
- [46] Riccardo M, Li L, Kidd BA, Dudley JT. Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep*, 2016, 6: 26094. [DOI]
- [47] S Somashekhar S P, Kumar R, Kumar A, Patil P, Rauthan A. 551PD validation study to assess performance of IBM cognitive computing system Watson for oncology with

- manipal multidisciplinary tumour board for 1000 consecutive cases: An Indian experience. *Ann Oncol*, 2016, 27(suppl 9), <https://doi.org/10.1093/annonc/mdw601.002>. [DOI]
- [48] Herath DH, Wilsoning D, Ramos E, Morstyn G. Assessing the natural language processing capabilities of IBM Watson for oncology using real australian lung cancer cases. 2016. [DOI]
- [49] Somashekhar SP, Sepúlveda MJ, Puglielli S, Norden AD, Shortliffe EH, Rohit CK, Rauthan A, Arun NK, Patil P, Rhee K, Ramya Y. Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann Oncol*, 2018, 29(2): 418–423. [DOI]
- [50] van der Heijden AA, Abramoff MD, Verbraak F, Van MH, Liem A, Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol*, 2018, 96(1): 63–68. [DOI]
- [51] Gibbons RJ, Balady GJ, Beasley JW, Bricker JT, Duvernoy WF, Froelicher VF, Mark DB, Marwick TH, McCallister BD, Thompson PD, Winters WL, Yanowitz FG, Ritchie JL, Gibbons RJ, Cheitlin MD, Eagle KA, Gardner TJ, Garson A, Lewis RP, O'Rourke RA, Ryan TJ. ACC/AHA Guidelines for Exercise Testing. A report of the American college of cardiology/American heart association task force on practice guidelines (committee on exercise testing). *J Am Coll Cardiol*, 1997, 30(1): 260–311. [DOI]
- [52] Eckel RH, Jakicic JM, Ard JD, de Jesus JM, Houston Miller N, Hubbard VS, Lee IM, Lichtenstein AH, Loria CM, Millen BE, Nonas CA, Sacks FM, Smith SC Jr, Svetkey LP, Wadden TA, Yanovski SZ, Kendall KA, Morgan LC, Trisolini MG, Velasco G, Wnek J, Anderson JL, Halperin JL, Albert NM, Bozkurt B, Brindis RG, Curtis LH, DeMets D, Hochman JS, Kovacs RJ, Ohman EM, Pressler SJ, Sellke FW, Shen WK, Smith SC Jr, Tomaselli GF. 2013 AHA/ACC guideline on lifestyle management to reduce cardiovascular risk: a report of the American college of cardiology/American heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 2014, 63(25): 2960–2984. [DOI]
- [53] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*, 2017, 12(4): e0174944. [DOI]

(责任编辑: 赵方庆)

## • 综合信息 •

# 第二届全国高校基因组学课程研讨会隆重召开

2018年8月7日至8日,由中国遗传学会和西安交通大学医学部联合举办的“第二届全国高校基因组学课程研讨会”在陕西省西安市隆重召开。会议由中国科学院院士杨焕明担任主席,来自中国科学院北京基因组研究所、南开大学、华中科技大学、厦门大学、东南大学、西南大学、西安交通大学、西北大学、南京医科大学、华南理工大学、内蒙古大学、重庆邮电大学、安徽师范大学、Journal of Genetics and Genomics编辑部、科学出版社、贝壳社、华大基因、中国遗传学会等20余家科研院所的40余名专家参与了本次会议。

会议围绕如何推进基因组学学科发展、加强本科生和研究生基因组学课程建设、以及《基因组学》教材的进一步改进议题等进行了深入的研讨与交流。会议认为,随着生命科学的发展,基因组学的内容越来越深于遗传学,设立基因组学课程是学科发展的趋势。针对《基因组学》教材的讨论,专家建议新版《基因组学》教材分设研究生版本和本科生版本,本科生版本应与高中《生物学》教材做好衔接,同时建议建立全国高校基因组学网络教学论坛。结合基因组学课程的教学讨论,与会教师在教学理念和方法创新等方面达成了诸多共识。本次会议对高校基因组学学科的发展和教学工作起到积极的推动作用。同时,会议呼吁广大遗传学和基因组学领域专家学者关注基因组学学科教育教学,为共同培养生命科学未来人才助力!