临床研究

基于机器学习的注意力缺陷多动障碍风险预测研究

赵健翔， 吴振起， 王雪峰， 王子， 褚亚奇， 游毅

基金项目：沈阳市科技计划项目(22-321-32-19)；辽宁省科学技术计划项目(2022-YGJC-81)

作者单位：110847 沈阳，辽宁中医药大学 2021 级中医儿科学专业研究生(赵健翔)；110034 沈阳，辽宁中医药大学附属第二医院(吴振起)；110032 沈阳，辽宁中医药大学附属医院儿科(王雪峰，王子)；110016 沈阳，中国科学院沈阳自动化研究所机器人学国家重点实验室(褚亚奇)；100089 北京，北京深睿博联科技有限责任公司研发中心科研合作部(游毅)

作者简介：赵健翔(1997—)，男，辽宁中医药大学 2021 级硕士研究生在读。研究方向：中医药防治感染性疾病

通信作者：吴振起，E-mail：zhenqiwu@163.com

【摘要】 目的 探讨基于机器学习算法对儿童注意力缺陷多动障碍(ADHD)预测的可行性。**方法** 回顾性分析我院于 2022 年 11 月至 2023 年 8 月儿科门诊就诊患者 358 例，其中 ADHD 患儿 119 例，非 ADHD 患儿 239 例，以人口学基本信息、儿童个人生活情况、母亲孕期情况、家庭生活情况及遗传因素等 31 个变量作为危险因素，采用单因素分析筛选出具有明显差异的变量，然后分别建立决策树模型、随机森林模型、自适应提升算法及 K 近邻算法模型，采用受试者工作特征(ROC)曲线的面积(AUC)、特异度、准确性、F1 分数及 ROC 曲线等进行模型预测效能评估。**结果** 4 种机器学习算法建立的 ADHD 的预测模型以随机森林算法最优，其 AUC 为 0.955，特异度、准确性、F1 分值分别为 0.903、0.898、0.853；同时，根据随机森林模型筛选出的前五位特征变量为：教育方式、情绪稳定情况、每日看电子产品时长、学习困难情况、近期反复呼吸道感染。**结论** 初步构建出基于机器学习算法建立儿童 ADHD 的预测模型，该模型对 ADHD 有良好的预测能力。

【关键词】 注意力缺陷多动障碍； 危险因素； 机器学习； 预测模型； 儿童

doi：10.3969/j.issn.1674-3865.2024.02.007

【中图分类号】 R749.94 **【文献标识码】** A **【文章编号】** 1674-3865(2024)02-0130-07

Risk prediction of attention deficit hyperactivity disorder based on machine learning *ZHAO Jianxiang, WU Zhenqi, WANG Xuefeng, WANG Zi, CHU Yaqi, YOU Yi. The Liaoning University of TCM, Shenyang 110847, China*

【Abstract】 Objective To explore the feasibility of predicting attention deficit hyperactivity disorder (ADHD) in children based on machine learning algorithm. **Methods** A total of 358 patients treated in the pediatric outpatient department of our hospital from November 2022 to August 2023 were retrospectively analyzed, and 119 patients were finally included in the ADHD group and 239 patients in the non-ADHD group. Totally 31 variables, including basic demographic information, children's personal life situation, mother's pregnancy situation, family life situation and genetic factors, were taken as risk factors. Single factor analysis was used to select variables with obvious differences, and then the decision tree(DT) model, random forest(RF) model, adaptive enhancement algorithm(Adaboost) and K-nearest neighbor algorithm(KNN) models were established respectively. AUC, specificity, accuracy, F1 score and ROC curve were used to evaluate the model prediction efficiency. **Results** Random forest algorithm was the best predictive model for ADHD, with AUC being 0.955, and specificity, accuracy and F1 scores being 0.903, 0.898 and 0.853, respectively. Meanwhile, the top five characteristic variables screened according to the random forest model were: education style, emotional stability, daily time spent playing with electronic products, learning difficulties, and recent recurrent respiratory infections. **Conclusion** A prediction model of child ADHD based on machine learning algorithm is established, which has good prediction ability for ADHD.

【Keywords】 Attention deficit hyperactivity disorder; Risk factors; Machine learning; Prediction model; Children

注意力缺陷多动障碍(attention deficit hyperactivity disorder,ADHD)是当今儿童时期最为常见的一类神经行为障碍性疾病,该疾病的主要特征为发育不良的注意力不集中和多动、冲动,进而导致学习、社交和情感功能多个领域的功能障碍[1]。截至 2022 年,国外 ADHD 患病率在儿童和青少年中约占 5%[2]。而国内患病率为 6.26%,影响着 2 300 万儿童身心健康发展[3]。ADHD 治疗不及时发展至成年后,则会增加躯体和精神共病以及低生活质量、社交障碍、职业成就低下和危险行为的风险[4]。目前对于 ADHD 的确诊高度依赖于对感知行为的主观评价,而缺少客观的生物学指标体系,早发现、早评估和诊断是治疗该病的重要前提[5]。近些年来,机器学习被广泛地应用于精神障碍类疾病当中,通过算法和统计学方法建立有效的结果预测,同时提供客观且准确的指标[6-7]。本研究基于 ADHD 采用随机森林、决策树、自适应提升算法及 K 近邻算法建立机器学习模型,以期对该疾病提供更为准确的发病风险预测,进而有针对性对预防及诊断提供帮助。

1 资料与方法

1.1 研究对象

回顾性分析辽宁中医药大学儿科门诊于 2022 年 11 月至 2023 年 8 月就诊患者 358 例,其中 ADHD 患儿 119 例,非 ADHD 患儿 239 例。

本研究经辽宁中医药大学附属医院伦理委员会批准[2023072FS(KT)-031-02]。同时对患者的姓名等个人信息进行了去隐私化处理。

1.2 诊断标准

参照美国精神医学学会《精神疾病诊断与统计手册》第 5 版(DSM-Ⅴ)[8]中 ADHD 的诊断标准,并结合临床表现进行半定量式访谈综合诊断后明确诊断。

1.3 纳入标准

(1)符合 ADHD 的诊断标准;(2)年龄 3～14 岁;(3)患儿家属均知情同意。

1.4 排除标准

(1)自闭症等其他精神障碍、神经系统疾病;(2)长期或近期服用治疗 ADHD 相关药物。

1.5 研究变量

研究结局变量为就诊时是否存在 ADHD 的症状,收集以下几类数据用于建模:(1)人口学基本信息:性别、年龄;(2)儿童个人生活情况:平时情绪是否稳定,学习是否困难,与同学朋友关系,喂养方式(母乳喂养、奶粉喂养、混合喂养),是否挑食或偏食,饮食类型(荤菜、素菜、均衡饮食)等;(3)母亲孕期情况:母亲受孕年龄(≥35 岁),妊娠期是否肥胖(孕期体质量指数≥28),孕期情绪是否低落,孕期不良生活习惯(吸烟/被动吸烟、酗酒、服用药物),孕期合并疾病(糖尿病、高血压、心脏病、低甲状腺素血症、甲状腺功能减退)等;(4)家庭生活情况:父母之间关系,家庭月收入情况,对孩子的教育方式(说服、责骂/打骂),抚养方式(单亲、双亲、隔代人或其他人抚养),父母文化程度(初中及以下、高中或大专、大专及以上)等;(5)遗传因素:父母是否患有 ADHD、孩子的同胞兄弟姐妹患有 ADHD、家族是否有精神疾病史等。

1.6 数据处理

(1)数据清理:对所有变量进行筛选,缺失值超过 5%的被排除在外,并且对内容过于一致的选项进行剔除。最终共纳入 31 个变量。

(2)模型开发:本文通过深睿医疗多模态科研平台(https://keyan.deepwise.com)构建决策树、随机森林、自适应提升算法、K 近邻算法 4 种机器学习模型,将本次研究纳入的 358 例患者按照 7∶3 的比例随机分配为训练集和验证集进行建模,并比较各模型的预测效能,选出最优模型。

(3)模型效能评价:通过训练集和验证集的受试者工作特征(receiver operating characteristic,ROC)曲线的面积(area under the curve,AUC)、特异度、准确性、F1 分数及 ROC 曲线等对于模型预测效能综合评估。其中 AUC 值越高,说明该模型预测结果越好;F1 分值用于综合反映整体的指标。其取值在 0～1,越接近 1,效果越好。

1.7 统计学方法

统计分析采用深睿医疗多模态科研平台,对数据进行清洗分析。符合正态分布的计量资料以($\bar{x}\pm s$)表示,组间比较采用独立样本 t 检验;不符合正态分布的计量资料以中位数 $M(P_{25},P_{75})$ 表示,应用 Wilcoxon 独立秩和检验比较两组间差异。计数资料采用 χ^2 检验。$P<0.05$ 表示差异有统计学意义。

2 结果

2.1 基本情况特征比较

两组患者在性别、情绪稳定情况、学习困难情况、与同学朋友关系、挑食或偏食的习惯,饮食类型、

每日看电子产品时长(≥2 h)、入睡困难情况、近期是否有呼吸道感染、是否有过敏性疾病、母亲孕期情绪是否低落、孕期是否被动吸烟等 20 个参数比较差异有统计学差异($P<0.05$),见表 1。

表 1 患者基本情况特征比较

因素	总例数($n=358$)	ADHD 患儿($n=119$)	非 ADHD 患儿($n=239$)	χ^2	P
性别				30.693	<0.01
男	200(55.9)	91(76.5)	109(45.6)		
女	158(44.1)	28(23.5)	130(54.4)		
情绪稳定情况				116.885	<0.01
好	197(55.0)	36(30.3)	161(67.3)		
一般	88(24.6)	20(16.8)	68(28.5)		
差	73(20.4)	63(52.9)	10(4.2)		
学习困难情况				70.419	<0.01
是	127(35.5)	78(65.5)	49(20.5)		
否	231(64.5)	41(34.5)	190(79.5)		
与同学朋友关系				46.370	<0.01
好	227(63.4)	60(50.5)	167(69.9)		
一般	102(28.5)	33(27.7)	69(28.9)		
差	29(8.1)	26(21.8)	3(1.2)		
挑食或偏食的习惯				23.058	<0.01
是	206(57.5)	89(74.8)	117(49.0)		
否	152(42.5)	30(25.2)	122(51.0)		
饮食类型				36.437	<0.01
荤菜	136(38.0)	71(59.7)	65(27.2)		
素菜	27(7.5)	8(6.7)	19(7.9)		
均衡饮食	195(54.5)	40(33.6)	155(64.9)		
每日看电子产品时长(≥2 h)				65.323	<0.01
是	160(44.7)	89(74.8)	71(29.7)		
否	198(55.3)	30(25.2)	168(70.3)		
入睡情况				40.793	<0.01
好	202(56.4)	56(47.1)	146(61.1)		
一般	97(27.1)	22(18.5)	75(31.4)		
差	59(16.5)	41(34.4)	18(7.5)		
近期反复呼吸道感染				43.871	<0.01
是	113(31.6)	65(54.6)	48(20.1)		
否	245(68.4)	54(45.4)	191(79.9)		
有过敏性疾病				29.048	<0.01
是	159(44.4)	75(63.0)	84(35.1)		
否	199(55.6)	44(37.0)	155(64.9)		
母亲孕期情绪低落				30.765	<0.01
是	94(26.3)	53(44.5)	41(17.2)		
否	264(73.7)	66(55.5)	198(82.8)		
母亲孕期被动吸烟				40.023	<0.01
是	38(10.6)	30(25.2)	8(3.3)		
否	320(89.4)	89(74.8)	231(96.7)		

续表 1

因素	总例数（n=358）	ADHD 患儿（n=119）	非 ADHD 患儿（n=239）	χ^2	P
胎儿脐带绕颈				21.386	<0.01
是	48(13.4)	30(25.2)	18(7.5)		
否	310(86.6)	89(74.8)	221(92.5)		
是否有流产史				5.345	0.021
是	107(29.9)	45(37.8)	62(25.9)		
否	251(70.1)	74(62.2)	177(74.1)		
父母之间关系				42.507	<0.01
和睦	277(77.4)	73(61.4)	204(85.4)		
一般	43(12.0)	16(13.4)	27(11.3)		
欠佳	38(10.6)	30(25.2)	8(3.3)		
对孩子的教养方式				180.121	<0.01
说服	241(67.3)	24(20.2)	217(90.8)		
责骂或打骂	117(32.7)	95(79.8)	22(9.2)		
对孩子的抚养方式				30.133	<0.01
双亲	298(83.2)	82(68.9)	216(90.4)		
单亲	11(3.1)	4(3.4)	7(2.9)		
隔代人或其他人	49(13.7)	33(27.7)	16(6.7)		
父亲的文化程度				14.643	0.001
初中及以下	57(15.9)	31(26.1)	26(10.9)		
高中或中专	48(13.4)	17(14.3)	31(13.0)		
大专及以上	253(70.7)	71(59.6)	182(76.1)		
母亲的文化程度				7.873	0.041
初中及以下	48(13.4)	23(19.3)	25(10.4)		
高中或中专	60(16.8)	23(19.3)	37(15.5)		
大专及以上	250(69.8)	73(61.4)	177(74.1)		
父母患有多动症				42.996	<0.01
是	28(7.8)	25(21.0)	3(1.3)		
否	330(92.2)	94(79.0)	236(98.7)		

2.2　各模型预测性能评价

将表 1 中 P<0.05 的参数纳入机器学习算法中构架模型。4 种模型在训练集和验证集中准确性均>0.6，AUC 均>0.8。在训练集中随机森林模型的 AUC、特异度、准确性、F1 分值均排在第一，AUC 为 0.988，特异度、准确性、F1 分值分别为 0.988、0.948、0.917；在验证集中，从 AUC 上看，随机森林预测结果最高，为 0.955；从特异度上看，K 近邻算法表现最佳，为 0.944；从准确性和 F1 分值上看，自适应提升算法略高于随机森林模型，为 0.907、0.865；综合考量以上指标，结果显示随机森林表现效果最好，预测效能最优。4 种机器学习模型预测的 ROC 曲线对比见图 1；模型预测性能对比见表 2。

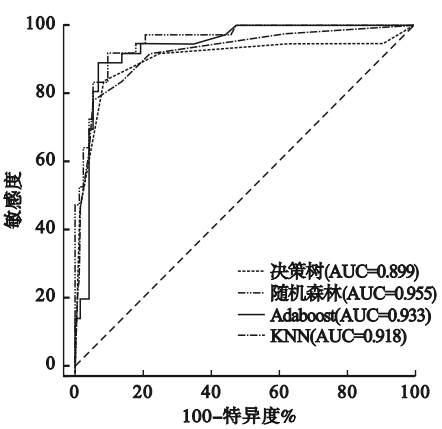

图 1　各模型 ROC 曲线

表 2 各模型对 ADHD 发生的预测性能比较

模型		AUC	特异度	准确性	F1 分值
决策树	训练集	0.926	0.748	0.812	0.768
	验证集	0.899	0.750	0.805	0.758
随机森林	训练集	0.988	0.988	0.948	0.917
	验证集	0.955	0.903	0.898	0.853
自适应提升算法	训练集	0.935	0.928	0.880	0.812
	验证集	0.933	0.916	0.907	0.865
K 近邻算法	训练集	0.942	0.898	0.888	0.809
	验证集	0.918	0.944	0.888	0.823

2.3 最优模型的特征重要性排名

在随机森林模型中，根据其重要性排名前五的特征变量为：教育方式、情绪稳定情况、每日看电子产品时长、学习困难情况、近期反复呼吸道感染，见图 2。在所有的特征变量中，最重要、贡献度最高的是父母对孩子的教育方式，这提示父母在孩子的教育方式上与 ADHD 的发病可能存在着更为紧密的联系。

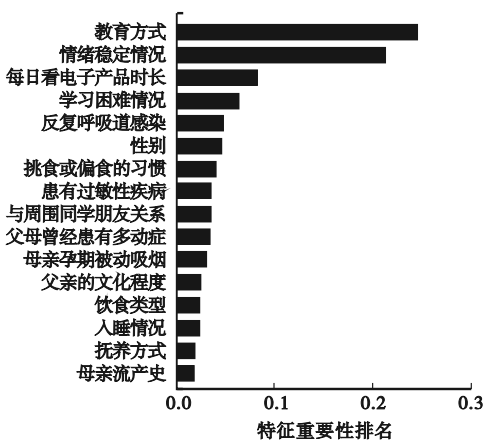

图 2 随机森林模型中重要特征排名

3 讨论

ADHD 是儿童最常见的精神行为障碍类疾病之一，目前该患病率达到 5.91%[9]，约有 78% 的患者在儿童时期出现的疾病症状会延续到成年。给个人、家庭、社会带来巨大影响[10]。目前国内对于该病的诊断主要根据 DSM-Ⅳ 中 ADHD 的诊断量表。但这些量表缺乏临床大样本验证，具体诊断还需结合实际临床经验进行全面评估[11]。目前机器学习被广泛应用于精神障碍类疾病的研究，如对于急性期精神分裂症[12]、阿尔茨海默病[13]、抑郁症[14]的预测等。当前国内通过机器学习诊断 ADHD 的研究尚少，因此，本次研究通过收集临床相关数据资料，建立决策树、随机森林、自适应提升算法、K 近邻算法 4 种机器学习算法对 ADHD 进行预测，并通过比较参数比较，筛选出诊断 ADHD 的最佳预测模型。

以上 4 种机器模型当中，随机森林、自适应提升算法为集成学习模型，是弱分类器学习模型[15]。在实际应用中发现，决策树和 K 近邻算法在处理大样本、高维度数据时，易出现过拟合的情况[16]。随机森林主要采用 Bagging 方法，具有良好的泛化能力，可降低过拟合的风险[17]。自适应提升算法则是通过组合弱分类器而得到强分类器，非常适合于处理二分类问题[18]。Haque 等[19]在应用机器学习模型对 ADHD 患儿发病进行预测中发现随机森林预测效果最佳，准确率为 91%，精确度为 94%，特异性为 99%。与本研究结果相似：随机森林模型 AUC 为 0.955、特异度为 0.903、准确性为 0.898、F1 分值为 0.853，其对于阳性结果的预测效果最优。

通过随机森林模型对纳入的变量进行排序，排在前五位的分别是父母对孩子的教养方式、孩子的情绪稳定情况、孩子每日看电子产品的时长、学习困难情况、孩子近期反复呼吸道感染。研究发现，社会心理因素在 ADHD 发病中起到重要的作用，目前 ADHD 的治疗方法主要是对患儿进行心理干预，对孩子责骂或打骂是诱发 ADHD 的独立危险因素[20]。这是因为儿童阶段是对父母最依赖的时期，父母教育方式不当，给孩子的心理造成创伤，容易出现行为异常的举动[21]。同时国外研究证实，良好说服教育方式长大的孩子在学习中更为成功；而暴力式的教育更容易加剧 ADHD 发病[22]。有研究表明，48%～54% 的 ADHD 儿童都存在着情绪调节障碍，多是在情绪紧张或沮丧的情况下出现，因此采用情绪控制训练的方法来控制情绪稳定[23-24]。近来发现，5-羟色胺转运体相关启动子区域（serotonin transporter gene linked polymorphic region，5-HTTLPR）与 5-羟色胺的转录调控及运输密切相关。5-HTTLPR

通过对脑执行控制网络和脑默认模式网络两大情绪控制网络进行改变，使携带 5-HTTLPR s 等位基因个体在遭受到压力下更加敏感，而 5-HTTLPR ll 纯合体无此精神行为的变化[25]。儿童及青少年由于在神经上的脆弱性，对电子产品更为敏感。ADHD 患儿自身抑制力较差，会花费更长的时间在电子产品上，从而加重症状[26-27]。在电子产品中获取到的及时奖励，导致患儿兴奋使机体内多巴胺释放增多，出现愉快感[28]。研究发现，ADHD 儿童的学习成绩、被重视程度均低于同龄正常儿童。学龄期儿童的心理行为处于塑造初期，儿童学习兴趣、求知欲和父母亲的教育方式存在显著关系，如父母关系欠佳、不良教育方式易促使儿童出现学习困难症状出现[29]。有研究发现 ADHD 儿童的反复上呼吸道感染发生率为非 ADHD 儿童的 1.769 倍，可能由于过大的压力导致患儿机体免疫功能紊乱，进而增加反复呼吸道感染的易感性[30]。也有研究者发现可能是通过炎症导致细胞因子释放增加穿过血脑屏障影响前额叶皮质的功能，诱发 ADHD 发生[31]。

本次研究联合多个特征变量构建了随机森林模型，对儿童 ADHD 的发病进行预测，且表现出良好的预测效能，但也存在以下的局限性：(1)本次研究可能忽略了其他的重要危险因素。例如，没有对患儿的脑电或临床上的检验指标进行关联，未来的研究应努力将这一类数据添补上。(2)本次研究为单中心研究，尚未进行外部验证，之后我们也将利用外部数据与本地数据进行比对，提高模型性能，使模型广泛使用。

4　结论

本研究基于 ADHD 相关危险因素构建了 4 种儿童 ADHD 预测模型，其中随机森林模型表现最好，其 AUC 为 0.955，该模型可以帮助儿科医师早期辨别 ADHD 患儿潜在群体，进而采取及时有效的干预措施，及早诊断和预防 ADHD。但该模型仍需进一步进行外部验证以及补充完善更多指标变量。

参考文献

[1] Cortese S, Coghill D. Twenty years of research on attention-deficit/hyperactivity disorder(ADHD): looking back, looking forward[J]. Evid Based Ment Health, 2018, 21(4): 173-176.

[2] Drechsler R, Brem S, Brandeis D, et al. ADHD: Current concepts and treatments in children and adolescents[J]. Neuropediatrics, 2020, 51(5): 315-335.

[3] Fan X, Ma Y, Cai J, et al. Do parents of children with ADHD know the disease? Results from a Cross-Sectional Survey in Zhejiang, China[J]. Children(Basel), 2022, 9(11): 1775.

[4] Salvi V, Ribuoli E, Servasi M, et al. ADHD and bipolar disorder in adulthood: clinical and treatment implications[J]. Medicina(Kaunas), 2021, 57(5): 466.

[5] 张晓华，崔永华，闫俊娟，等. 儿童注意缺陷多动障碍的评估与诊断[J]. 中国实用儿科杂志，2023，38(8)：584-587.

[6] Kautzky A, Vanicek T, Philippe C, et al. Machine learning classification of ADHD and HC by multimodal serotonergic data[J]. Transl Psychiatry, 2020, 10(1): 104.

[7] Pereira-Sanchez V, Castellanos FX. Neuroimaging in attention-deficit/hyperactivity disorder[J]. Curr Opin Psychiatry, 2021, 34(2): 105-111.

[8] American Psychiatric Association. Diagnostic and statistical manual of mental disorders(DSM-5)[M]. 5th ed. Arlington, VA: American Psychiatric Association, 2013: 81.

[9] Huang Y, Zheng S, Xu C, et al. Attention-deficit hyperactivity disorder in elementary school students in Shantou, China: prevalence, subtypes, and influencing factors[J]. Neuropsychiatr Dis Treat, 2017, 13: 785-792.

[10] Zheng Y, Pingault JB, Unger JB, et al. Genetic and environmental influences on attention-deficit/hyperactivity disorder symptoms in Chinese adolescents: a longitudinal twin study[J]. Eur Child Adolesc Psychiatry, 2020, 29(2): 205-216.

[11] 许明慧，韩梦蝶，赵雪，等. 注意力缺陷多动障碍评估工具的研究进展[J]. 护理研究，2023，37(2)：289-292.

[12] 仲捷，朱虹，郑思思，等. 基于机器学习的急性期精神分裂症心理理论能力对社会功能的预测作用[J]. 首都医科大学学报，2023，44(4)：596-601.

[13] 丛慧文，徐雅琪，王爱民，等. XGBoost 算法在轻度认知障碍人群阿尔兹海默病发病预测中的应用[J]. 郑州大学学报(医学版)，2022，57(6)：751-756.

[14] 聂卉，吴晓燕. 结合梯度提升树算法与可解释机器学习模型 SHAP 的抑郁症影响因素研究[J/OL]. 数据分析与知识发现，1-17[2024-03-02]. http://kns.cnki.net/kcms/ detail/10.1478.G2.20230504.1700.006.html.

[15] Chatpreecha P, Usanavasin S. Design of a collaborative knowledge framework for personalised attention deficit hyperactivity disorder(ADHD) treatments[J]. Children(Basel), 2023, 10(8): 1288.

[16] Garcia-Argibay M, Zhang-James Y, Cortese S, et al. Predicting childhood and adolescent attention-deficit/hyperactivity disorder onset: a nationwide deep learning approach[J]. Mol Psychiatry, 2023, 28(3): 1232-1239.

[17] Kaur A, Kahlon KS. Accurate identification of ADHD among adults using real-time activity data[J]. Brain Sci, 2022, 12(7): 831.

[18] 苏枫，张少衡，陈楠楠，等. 基于机器学习分类判断算法构建心力衰竭疾病分期模型[J]. 中国组织工程研究，2014，18(49)：7938-7942.

[19] Haque UM, Kabir E, Khanam R. Early detection of paediatric and adolescent obsessive-compulsive, separation anxiety and attention deficit hyperactivity disorder using machine learning algorithms[J]. Health Inf Sci Syst, 2023, 11(1): 31.

[20] 李赛，樊秋月，刘旭华，等. 儿童注意缺陷多动障碍相关危险因素的研究进展[J]. 中国中西医结合儿科学，2023，15(5)：410-415.