doi:10.3969/j.issn.0253-9608.2018.05.004

人工智能在疾病预测中的应用

徐亮†，阮晓雯，李弦，洪博然，肖京††

平安科技(深圳)有限公司，广东 深圳 518057

摘要 首先介绍了人工智能在疾病预测中的应用现状和发展，包括公共卫生防控和个人疾病筛查及健康管理两个方面。接着分析了传统疾病预防方法存在的弊端，从数据源和技术方面综述了人工智能给疾病预测带来的突破和发展。最后，列举了人工智能技术在疾病预测中的应用。

关键词 疾病预测；人工智能；卫生防控；大数据

近年来，得益于人工智能技术的突破性进展以及数据来源的不断丰富和积累，人工智能不断运用在新的产业中，从制造业扩展到家庭、娱乐、教育、军事、医疗以及金融等专业服务领域，助力传统行业智能升级或智能转型。其中，人工智能在医疗领域的应用尤其受到重视和关注，这是因为医疗健康是国家及民众重点关注的民生问题。党的十九大报告更是提出“实施健康中国战略”，从国家战略层面上推动卫生健康事业实现高质量发展。1949年以来，中国主要疾病谱变化显著，死亡率高的疾病由以传染病为主过渡到以慢性病为主。中国传统的以治疗为主的诊疗模式也将会随着国家疾病谱变化而改变，未来以预防为主的诊疗模式可能更加贴合国情民情。现阶段我国的公共卫生工作仍然集中在疾病事中和事后的应急处理，而目前发达国家的公共卫生工作已发展到以疾病预防为主。我国的疾病预测和疾病预防工作急切而紧迫。

1 疾病预防的现状

疾病预防从宏观和微观层面可以分为公共卫生防控和个人疾病筛查及健康管理。以重大传染病疫情为主的突发公共卫生事件不仅危害人民的生命财产安全，还极易造成社会影响，影响群众日常生活的方方面面，甚至阻滞经济的发展。以流感为例，根据世界卫生组织最新估计[1]，全球每年5%～10%的成人和20%～30%的儿童会罹患流感，流感的季节性流行会导致全球300万～500万重症病例和25万～50万死亡病例。建立和发展传染病预测预警技术，提高预测预警的及时性和准确性，对于传染病控制工作意义重大。目前各国政府实际采用的传染病疫情预警系统主要依赖传统监测手段，包括各级医疗机构、疾病预防控制中心和流感样病例监测哨点医院协作，由医疗机构诊断并报告流感临床诊断病例和确诊病例。现有的流感监测体系存在诸多弊病：定时抽样、每周汇总的数据获取方式，数据结果

†深圳市“孔雀计划”专家，现任平安科技智能引擎部副总工程师，研究方向：人工智能在智慧城市、智慧医疗的研发及应用。E-mail: xlpaul@126.com

††通信作者，国家“千人计划”特聘专家、广东省珠江人才、深圳市“孔雀计划”A类人才、高级工程师（教授级），现任平安集团执委、平安集团首席科学家，研究方向：人工智能与互联网大数据分析挖掘相关领域。E-mail: xiaojing661@pingan.com.cn

相对滞后；监测手段耗费大量人力物力，遍及全国的监测网络中任一节点产生的差错都将影响数据的准确性，且各实验室检测和逐级上报的过程繁琐；该监测手段获取的数据来源单一，无其他来源数据的比对修正。

在个人疾病筛查和健康管理方面，随着城市化和人口老龄化进程加快，诸如心脑血管疾病、慢性阻塞性肺疾病、恶性肿瘤、中风、糖尿病等原先被视为发达国家独有现象的慢性非传染性疾病已急剧改变中国人口的疾病谱。我国每年死于慢性非传染性疾病和伤损的人数近年来迅速上升。慢性病导致的医疗负担给个人、家庭以及整个国家的医疗保障体系带来了沉重的经济和社会负担。数据显示[2]，目前我国的慢性病患者已超过3亿人，慢性病致死人数已占到我国因病死亡人数的80 %，导致的医疗负担已占到总医疗负担的70 %。同时，糖尿病等慢性病已呈现年轻化发展趋势，严重影响到居民的生活质量和身体健康。由于慢性病的症状一般不明显，患者大多无法在患病初期及时发现并进行医疗干预，往往发现时已是中晚期，不仅治疗难度增加，患者的疾病经济负担也随之升高。有效的慢性病管理，对慢性病相关的危险因素进行筛查，可以及早发现疾病的发展趋势，一方面帮助患病高危人群提高疾病意识，从而做到早发现、早诊断、早治疗，另一方面帮助政府干预、控制危险因素，降低民众的患病风险。传统的慢性病人群筛查主要依据历史统计结果，其筛查因素范围及力度有限，因此亟需高质量的慢性病管理体系。

2 基于人工智能技术的疾病预测研究现状

目前，人工智能的浪潮汹涌澎湃，在视觉图像识别、语音识别、文本处理等诸多方面人工智能已经达到或超越人类水平。大数据和人工智能技术的发展也为疾病预测带来了新突破。

2.1 大数据源

随着互联网和物联网技术的发展和普及，人们生活中的行为和状态很可能转化成数据记录，而这些电子数据，尤其是互联网数据都具有覆盖群体大、实时性高的特点，对于疾病防控具有较大的利用价值。通过监测网络数据源发现公共健康事件的相关研究数量增多，尤其是搜索服务提供商等持有大量用户行为数据的公司在这一方向上做出了许多卓有成效的尝试。

2008年，Google公司开发了“谷歌流感趋势”(Google Flu Trends，GFT)软件，利用Google巨大的用户搜索数据(认为网络用户及其家人出现流感相关症状时可能采取搜索相关的关键词的行为)，提前1~2周准确预测了美国流感样病例百分比的变化趋势，由此在学术界掀起了利用互联网数据预测流感的研究浪潮[3]。尽管GFT在后期预测中出现较大偏差[4]，但越来越多的研究表明搜索数据可以作为流感预测的有效因子之一[5-7]。在传染病流行季节，人们除了通过搜索引擎关注传染病的暴发情况以及应对措施外，还有可能会在社交网络平台上发表有关自己或家人朋友患病情况的言论。2011年，Signorini等[8]以美国境内发表的含有流感相关关键词的每周Twitter量的占比作为预测因子，采用支持向量机回归(support vector regression，SVR)模型算法建立了美国全国及某一地区的流感样病例百分比的实时跟踪预测模型，交叉验证的32周预测结果平均误差不超过0.4 %。2013年， Li等[9]利用Twitter数据建立了流感暴发的早期预警模型。他们采用分类算法对Twitter数据进行自动过滤，留取与流感相关的记录，再通过无监督算法结合流感的空间时间信息进行预测，发现预测结果与真实数据的相关系数达到0.97。在我国，研究人员尝试使用中文搜索引擎百度的搜索数据[5]以及新浪微博等社交媒体的数据[10]构建流感预测模型，验证利用互联网舆情数据预测我国流感的可行性。

不断兴起的互联网应用也持续为疾病防控，特别是传染病的监控和预测提供了新思路。比如自发性报告流感的网络监测系统(如美国的Flu Near You 、澳大利亚的Flutracking)[11]以及近年来用户量激增的在线健康咨询及管理的移动互联网应用平台，其与疾病相关的导医初诊及预约挂号数可以直接反映用户的患病情况，且超前于

医院就诊记录。此外，各互联网医疗平台的药物出售统计量也可反映疾病的流行形势。这些数据结合人工智能算法都被尝试用于传染病等公共卫生事件的预测预警建模[12]，且具有较好的预测效果。

除了新兴的互联网数据源，医疗相关的传统数据转换成结构化或非结构化的电子数据后，随着人工智能技术的突破在疾病预测中同样发挥着重要作用。借助于先进的人工智能算法，研究者使用可穿戴设备或远程医疗设备实时记录的患者生命体征数据[13]、患者的电子病历[14]、体检数据[15]、医学影像(超声/CT/核磁)[16]，乃至患者的语音数据[17]，建立了个人患病风险评估模型，自动筛查疾病相关的危险因素。从2011年起，大量的研究者开始利用可穿戴设备或远程医疗设备记录的用户生命体征数据进行慢阻肺和哮喘患病风险的预测，并不断对预测模型进行优化改进，目前预测准确率为94 %[13]。华中科技大学的Chen等[15]利用结构化的医院数据包括个人属性(性别、年龄、身高体重等)、生活习惯(吸烟与否)、检查结果(血常规等)和非结构化的个人患病史及历史医嘱等文本数据，基于改进的卷积神经网络对个体脑梗患病风险进行预测，预测准确率达到94.8%。波士顿大学的Theodora等[14]采用改进的人工智能算法，基于电子病历预测了心脏病以及糖尿病两种慢性病的患病风险。

随着人工智能算法的改进及GPU对计算能力的提升，从大数据层面，充分利用多源、复杂、更全面的疾病相关数据已然成为了疾病预测的趋势。丰富的特征数据源增加了疾病监控和筛查的维度，对传统数据源提供了有力的补充，也为人工智能技术在疾病预测中的应用提供了充足的“燃料”。

2.2 人工智能技术

近年来，人工智能技术的突破一方面离不开算法性能更优、灵活度高的机器学习算法的开发，更主要的是归功于深度学习技术的成熟。2006年，Geoffrey Hinton提出深层神经网络逐层训练的高效算法，让当时计算条件下的神经网络模型训练成为了可能，同时通过深度神经网络模型得到的优异的实验结果让人们开始重新关注人工智能。之后，深度神经网络模型成为了人工智能领域的重要前沿阵地，深度学习算法模型也经历了一个快速迭代的周期，深度信念网络(deep belief network)、稀疏编码(sparse coding)、循环神经网络(recursive neural network，RNN)、卷积神经网络(convolutional neural network，CNN)等各种新的算法模型被不断提出。利用深度学习模型，人工智能在图像识别、语音识别及自然语言处理等领域都达到了令人满意的识别精度，有些领域甚至赶超人类。

公共卫生事件的预测预警主要是预测未来时间点某一个城市或地区居民传染病如流感的患病率，而针对个体的疾病风险预测是预测个体在未来设定的时间窗口内是否会患某种疾病或者患病的概率。在人工智能领域，这些预测场景则会转换成回归预测或分类建模问题，利用人工智能技术进行疾病预测建模的主要技术点如下：

(1) 数据预处理。用于疾病预测的输入数据，比如电子病历经常存在字段缺失或者数据异常的情况，导致特征无法提取或者给建模造成噪声，因此需要对输入数据进行去噪、缺失值填充等预处理。缺失值填充方法除了常用的均值填充、中位数填充等，有研究针对该问题提出的隐藏因子模型进行缺失值自动填充，有助于疾病预测精度的提升[15]。

(2)特征选择。在疾病预测应用中，用于传染病预测的特征因子可能涵盖天气、舆情、人口等多源数据。在疾病风险预测中，每位患者的数据涉及病情主诉、诊断、生活习惯等，往往有上百维，而真实电子病历的数据甚至有上千维。因此在使用机器学习算法进行建模时，为了避免冗余的无意义的特征给模型引入噪声，降低模型拟合的精度，需要选择有意义的、相关的特征作为模型的输入。疾病预测中使用的特征选择算法类别包括过滤法(方差及相关系数检验)、封装法(前向特征选择等)以及嵌入式法(树模型等)。在使用深度学习算法进行建模时，深度学

习网络将原始特征进行多层变换，把原始特征映射到新的空间中，因此不需要另外加入特征选择模块。

(3)模型选择。用于挖掘序列本身相关性规律的时间序列模型自回归积分滑动平均模型(autoregressive integrated moving average model，ARIMA)是经典的传染病患病率预测模型，用数学模型近似描述序列的变化，对于短期趋势的预测准确率较高[18]。逻辑回归模型(logistic regression, LR)由于可解释性强被广泛应用在疾病预测中。2001年新兴的集成学习算法——随机森林(random forest)及其后续的改进算法，由于兼具可解释性且能够进一步提高预测精度，被应用在越来越多的疾病预测研究中[19]。此外，SVR回归[8]、Lasso回归以及组合模型[11]等预测算法也被尝试用于传染病患病率及个人疾病风险预测模型中。近年来，由于深度学习算法在处理高维复杂的结构化数据以及非结构数据时表现出优秀的算法性能，已有一些研究利用深度学习算法建立疾病预测模型，采用卷积神经网络(CNN)[15]、循环神经网络(RNN)[20]对电子病历数据、医学图像以及语音数据进行分析，预测个人患病风险。

2.3 应用突破

先进的人工智能算法也给疾病预测带来了新的发现和突破。2017年4月，英国诺丁汉大学流行病学家Stephen Weng博士团队将机器学习算法应用于电子病历的常规数据分析，发现与当前的心脏病预测方法相比，机器学习算法不仅可以更准确地预测心脏病发病的风险，还可以降低假阳性患者的数量。该团队利用随机森林、逻辑回归、梯度提升(gradient boosting)和神经网络4种人工智能算法预测人类患心血管疾病的风险，“摸索”出传统模型结果中未出现的如房颤、种族差异等重要风险因子[21]。

根据科学期刊《自然》的报道[22]，2017年2月北卡罗来纳大学的精神病学家Heather Hazlett带领团队利用深度学习算法，开发了可预测12个月大的儿童在2岁时是否会患上自闭症的人工智能系统。采用的人工智能算法通过不断“学习”脑部数据自动判断婴儿的大脑生长速度是否异常，以此来获得自闭症的早期线索。这种预测方法具有 81%的准确率与88%的灵敏度。这意味着医生可以借助这套算法在疾病发生的早期，筛选出会患病的儿童，提前进行介入治疗以达到更好的治疗效果。

2018年，IBM 研究团队利用机器学习预测人类罹患精神疾病的风险。IBM团队用人工智能算法分别对59名受试者的语言模式进行了追踪和分析。受试者参加了一项访谈测试，访谈的记录依据词性不同被逐个拆解，然后对句子的连贯性进行评分。机器算法则根据他们的语言模式判断哪些人有罹患精神疾病的风险。受试者中有19人在两年内患上了精神疾病，其余40人则一切正常，算法预测的准确率高达83%。这套算法还能够区分近期罹患精神疾病的人群与正常人群的语言模式，而且准确率达到72%。研究发现，具有高患病风险的人说话时较少使用物主代词，说出的句子也不那么连贯[17]。

经典的机器学习和统计方法普遍采用基于向量的表示方法，通过特征选择提取最有预测能力的特征。最新的深度学习方法从输入数据中自动学习特征，对原始数据进行多层变换，把原始特征映射到新的空间，虽提高了预测精度，但同时也降低了模型的可解释性。

3 人工智能在疾病预测中的应用落地

3.1 人工智能在公共卫生防控方面的疾病预测应用

2017年，平安集团与重庆市疾病预防控制中心的联合研发课题组，利用“互联网+医疗健康”大数据前沿技术，首次提出“宏观+微观”的深度智能疾病预测方法，实现了提前一周预测某一地区流感和手足口病的患病率。该模型整合了上万维度数据因子，同时结合本地疾病防控实际业务经验和专家知识，采用多种人工智能算法的组合，使疾病预测能够达到时效性更强、精度更高、范围更广、输出更稳定、可扩展性更强的要求，充分体现了多维数据来源的业务应用优势

和实践价值。

该流感预测模型在宏观或地区层面，通过整合全国上百个城市的环境气象因子(环境/天气/季节)、人口信息(人口/流动/结构)、地区生活行为、医疗习惯、就诊行为等一系列宏观因子，对历史数据进行尝试挖掘，分析时间序列。在微观层面，通过整合全方位、多维度的预测因子和信息来预测疾病发生风险。这些信息包括信息高度相关，但频度较低、分布较稀疏的医疗健康因子(体检/就诊/告知等)，也包括信息间接相关，但信息频度和深度较高的个人行为因子(财务/职业/生活等)、互联网数据因子(舆情/行为/LBS等)等。通过精准评估个人层面风险并汇总到宏观层面，该方法能够深入挖掘宏观层面无法统计的细颗粒度的信息，从而提升预测精度。最终采用模型融合的方法，将深度学习和人工智能方法，如时间序列模型、树模型等进行组合，提高预测准确度。该流感预测模型目前已在重庆市上线应用，在重庆长达3年的历史静态数据及上线后动态数据的验证中，预测平均误差率都不超过10 %。

基于人工智能技术的传染病预测，将帮助政府部门及时监控疫情和合理分配医学资源，并指导民众进行疾病预防，提升疾病事前预防的成功率，有效降低国家疾病预测与防控工作的成本。

3.2 人工智能在慢性病筛查中的应用

2017年，平安集团与重庆市卫生计生委联合开展大数据在慢阻肺筛查与防控方面的应用研究，研发的慢阻肺危险因素筛查模型准确率达到92 %。应用慢阻肺危险因素筛查模型，可大幅减少城市医疗管理部门的筛查成本，提高筛查效率；同时利用早期筛查和早期干预，可显著减少患者疾病的经济负担。

2018年8月，平安集团在上海黄浦区某药店正式上线个人智能疾病预测系统，完成了人工智能在个人疾病风险预测中应用的落地实施。顾客在完成血压、心率等物理设备检测时，就可以同步进行12类常见的糖尿病及其并发症、心脑血管疾病、高血压、慢性肾病、慢性阻塞性肺疾病等慢性病的智能风险精准预测。该系统基于大数据并采用人工智能和机器学习技术建立而成，从大量特征中挖掘疾病风险因子，进行风险因素分析，并融合专家知识，针对精准人群提供个性化的预防干预建议。糖尿病筛查等模型的准确率在90%以上，灵敏度较通用模型提高了50%以上。

2018年2月美国食品药品监督管理局(FDA)批准了一项人工智能成果——Cognoa公司用于检测儿童自闭症的人工智能平台，这也是FDA监管许可的首个用于自闭症筛查的II类诊断医疗设备。通过分析家长提供的儿童自然行为信息和视频，Cognoa的应用程序使用机器学习算法来评估该儿童是否正在以正确的速度发展，并评估他们的行为健康状况。该应用已经通过临床验证，可以在早期识别儿童的自闭症，其准确率超过80%。

人工智能技术在个人疾病筛查和健康管理中的应用能够帮助患病高危人群的高效筛选，及早发现疾病的发展趋势，提高疾病防控意识。通过患病因素分析获得定制化的健康信息服务，比如个人健康顾问、预防治疗措施以及求医用药指导等等，也是未来人工智能在疾病预测领域应用落地的重要方向。

4 结论与展望

人工智能技术的发展使得疾病预测智能化和精确化成为可能，人工智能在疾病预测中的应用近年来也取得了较大的突破。然而，人工智能技术在疾病预测中的预测精度还有待进一步提高。一方面，如何处理多模态的医疗数据，充分利用结构化数据、文本、影像和流数据(心率、血氧、呼吸等)等综合信息进行疾病预测建模，提高预测模型的精度和泛化能力是接下来很重要的技术挑战。另一方面，由于医学领域的特殊性，对预测模型的可解释性具有较高要求。然而，目前由数据驱动的人工智能疾病预测模型，其预测原理较难回溯到医疗领域知识。如何有效地融合医学领域知识和机器学习方法，构建可解释性强的预测模型还有待深入研究。

(2018年9月17日收稿)■

参考文献

[1] WORLD HEALTH ORGANAZATION. Influenza (Seasonal): Fact sheet no.211. (2018-01-31)[2018-09-05]. URL: http://www.who.int/mediacentre/factsheets/fs211/en/.

[2] 前瞻产业研究院. 2018-2023年中国大健康产为战略规划和企业战略咨询报告[R/OL]. [2018-09-16]. https://bg.qianzhan.com/report/detail/1801091028309482.html.

[3] GINSBERG J, MOHEBBI M H, PATEL R S, et al. Detecting influenza epidemics using search engine query data [J]. Nature, 2009, 457(7232): 1012.

[4] LAZER D, KENNEDY R, KING G, et al. The parable of Google flu: traps in big data analysis [J]. Science, 2014, 343(6176): 1203.

[5] YUAN Q, NSOESIE E O, LY B, et al. Monitoring influenza epidemics in china with search query from baidu [J]. Plos One, 2013, 8(5): e64323.

[6] HYEKYUNG W, YOUNGTAE C, EUNYOUNG S, et al. Estimating influenza outbreaks using both search engine query data and social media data in South Korea [J]. Journal of Medical Internet Research, 2016, 18(7): e177.

[7] ZHANG Y, BAMBRICK H, MENGERSEN K, et al. Using Google trends and ambient temperature to predict seasonal influenza outbreaks [J]. Environment International, 2018, 117(2018): 284.

[8] ALESSIO S, MARIA S A, POLGREEN P M. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic [J]. Plos One, 2011, 6(5): e19467.

[9] LI J, CARDIE C. Early stage influenza detection from Twitter [J]. Computer Science, 2013. arXiv: 1309.7340.

[10] 黄江妙. 基于社交网络的流感监控和预测算法[D]. 上海: 华东师范大学, 2015 .

[11] LU F S, HOU S, BALTRUSAITIS K, et al. Accurate influenza monitoring and forecasting using novel internet data streams: a case study in the Boston metropolis [J]. Jmir Public Health Surveill, 2018, 4(1): e4.

[12] LIU T Y, SANDERS J L, TSUI F C, et al. Association of over-the-counter pharmaceutical sales with influenza-like-illnesses to patient volume in an urgent care setting [J]. Plos One, 2013, 8(3): e59273.

[13] SANCHEZ-MORILLO D, FERNANDEZ-GRANERO M A, LEON-JIMENEZ A. Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: A systematic review [J]. Chron Respir Dis, 2016, 13(3): 264-283.

[14] BRISIMI T S, XU T, WANG T, et al. Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach [J]. Proceedings of the IEEE, 2018, 106(4): 690-707.

[15] CHEN M, HAO Y, KAI H, et al. Disease prediction by machine learning over big data from healthcare communities [J]. IEEE Access, 2017, 5(99): 8869-8879.

[16] GILLIES R J, KINAHAN P E, HRICAK H. Radiomics: images are more than pictures, they are data [J]. Radiology, 2016, 278(2): 563-577.

[17] CORCORAN C M, CARRILLO F, FERNÁNDEZ-SLEZAK D, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis [J]. World Psychiatry Official Journal of the World Psychiatric Association, 2018, 17(1): 67.

[18] SMOLEN H J. PRM102—development of an influenza outbreak forecasting model using time series analysis methods [J]. Value in Health, 2014, 17(7): A561-A561.

[19] SOUNAK C, MOHAMMED K, MIHAIL P. Predicting disease risks from highly imbalanced data using random forest [J]. Bmc Medical Informatics & Decision Making, 2011, 11(1): 51.

[20] RAZAVIAN N, MARCUS J, SONTAG D. Multi-task prediction of disease onsets from longitudinal lab tests [J]. Machine Learning, 2016. arXiv:1608.00647 [cs.LG].

[21] WENG S F, REPS J, KAI J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? [J]. Plos One, 2017, 12(4): e0174944.

[22] HAZLETT H C, GU H, MUNSELL B C, et al. Early brain development in infants at high risk for autism spectrum disorder [J]. Nature, 2017, 542(7641): 348-351.

Use of artificial intelligence in disease prediction

XU Liang, RUAN Xiaowen, LI Xian, HONG Boran, XIAO Jing

Ping An Technology(Shenzhen) Co., Ltd, Shenzhen 518057, Guangdong Province China

Abstract This work introduces the application status and prospect of artificial intelligence in disease prediction comprising two aspects: public health prevention and control, personal disease screening and health management. The drawbacks of traditional methods for disease prevention and control are analyzed. The breakthroughs and developments of disease prediction brought by artificial intelligence are summarized in view of data sources and techniques. Finally, this work gives some examples of the productions of disease prediction by artificial intelligence.

Key words disease prediction, artificial intelligence, health prevention and control, big data

（编辑：段艳芳）