

多个 AI 疾病预测模型竟由可疑数据训练而成

本报讯 一项近日公布于预印本服务器 medRxiv 的研究显示, 一些可疑的数据集正被用于训练那些预测人们患病风险的人工智能(AI)模型, 其中一些模型已投入临床应用。

据《自然》报道, 澳大利亚昆士兰科技大学的 Adrian Barnett 团队发现, 124 篇经过同行评审的论文均报告使用两个开放获取健康数据集中的一个是来训练机器学习模型, 但没有提供这些数据的来源。

相关分析揭示了许多不可能来自真人数据的异常之处。这让 Barnett 团队怀疑, 这些数据可能是伪造的。

“基于不明来源数据的预测模型在临床决策中毫无价值, 它们本质上是不可靠的。”澳大利亚乔治全球健康研究院的 Soumyadeep Bhaumik 表示, 如果这些工具未使用真实世界的的数据, 则很可能做出错误的预测, 导致临床医生做出不当的决策, 例如开出不必要的治疗处方, 或在需要时不开处方。

目前至少有两个模型已在印度尼西亚和西班牙的医院使用, 其中一个还出现在 2024 年提交的一份医疗器械专利申请中。还有两个模型是公开的网络安全工具, 允许人们通过上传个人信息



图片来源: Marko Nikolic/Alamy

来评估自身的风险等级。

Bhaumik 表示, 研究机构和资助方必须要求研究人员披露用于训练医疗应用 AI 模型的数据来源, 期刊也应拒收未披露数据来源的论文。Barnett 提醒, 研究中标记的问题数据集应立即“删除”, 以防被后续研究使用。

这项研究调查的两个数据集均来自 Kaggle——一个供开发者获取数据集以构建机器学习模型的平台。

第一个数据集名为“中风预测数据集”, 上传说明为“用于预测中风事件的 11 项临床特征”, 包含了 5110 人的健康信息, 涉及心脏病史、婚姻状况、平均血糖水平和身体质量指数(BMI)

等风险因素数据。

然而, 当将平均血糖水平与参与者标识进行对比后, 研究人员发现了多处异常。Barnett 指出, 其中一处异常是数据缺失极少, 这与真实数据形成了鲜明对比。真实数据通常存在缺失, 因为部分参与者会错过随访、退出研究或去世。“在现实世界中收集的任何数据集都不可能是完整无缺的。”

共有 104 项研究使用该数据集构建中风预测模型, 其中包括一个被印度尼西亚一家医院采用的模型, 以及一个在少数人身上进行了测试的模型。该中风数据集由西班牙数据科学家 Federico Soriano Palacios 上传, 下载量

已超过 28.8 万次。

第二个数据集名为“糖尿病预测数据集”, 是“一个利用医疗和人口统计数据预测糖尿病的综合数据集”, 包含了 10 万人的信息, 涵盖 BMI、吸烟史、血糖水平等。但 Barnett 团队发现, 所有参与者的数据中只有 18 个离散的血糖值。鉴于人群存在巨大差异, 这种情况是不可能的。该团队还发现了数千个似乎重复的数值。该糖尿病数据集由印度数据工程师 Mohammed Mustafa 上传。

共有 21 项研究使用该数据集构建了糖尿病预测模型, 但截至目前, 这些模型均未应用于临床。值得一提的是, 有一项研究同时使用了上述两个数据集。

Palacios 和 Mustafa 迄今仍未对数据来源作出回应, Kaggle 也拒绝就该平台是否会对这些数据集进行调查或采取相关措施发表评论。

《科学报告》主编 Rafal Marszalek 表示, 他的团队正在审查 Barnett 团队标记的 5 篇论文, 并将在适当情况下采取相应措施。 (文乐乐)

相关论文信息:

<https://doi.org/10.64898/2026.02.24.26347028>

大脑内部有一款“计时器”

本报讯 大脑如何判断该在何时醒来? 研究人员在小鼠大脑中发现了一种化学信号, 能够充当大脑内部的计时器, 记录每次睡眠的时长和被打断的次数, 并预测动物在任何特定时刻醒来的可能性。这种信号有望为测量“睡眠债”, 也就是累积的睡眠缺失提供一种新的方法。相关成果近日公布于预印本平台 bioRxiv。

“睡眠不足会带来负面影响, 造成十分严重的后果。”美国加州大学洛杉矶分校的 Ketema Paul 表示, 这类生物标志物未来或许有助于开发出检测人体是否睡眠不足的方法。这种检测十分实用, 能够保障高风险岗位从业者维持良好的警觉性, 例如卡车司机、医院急诊室工作人员。

以往的睡眠研究大多聚焦在可以让人瞬间清醒的大脑神经回路, 或是

连续多日睡眠不佳后的“睡眠债”累积上。但介于两者之间的时间尺度——大脑如何记录在一次连续睡眠中度过的分钟或小时, 一直是个谜。

为寻找在睡眠过程中发生渐进变化的分子信号, 美国圣路易斯华盛顿大学的陈瑶团队在小鼠大脑中使用一种特殊荧光传感器, 实时观测了蛋白激酶 A(PKA) 的作用。这种酶同样存在于人体中, 此前已被证实和觉醒状态有关。PKA 会在脑细胞表面的蛋白质上添加化学标记。

研究人员发现, 这些标记的存在遵循一个清晰且可预测的规律: 小鼠清醒时, 蛋白质标记维持高水平且稳定; 一旦小鼠进入睡眠, 蛋白质标记的水平就开始缓慢而持续地下降。这种下降贯穿了整个睡眠过程。每当小鼠出现微觉醒或睡眠短暂中断时, 与 PKA 相关的信号就会升高, 随后再下降。

陈瑶表示, 这意味着蛋白质标记可以同时记录睡眠总时长和被打断的频率。研究人员发现, 由于 PKA 信号可以反映这些躁动不安的时刻, 因此监测它比追踪小鼠的睡眠时长更能准确预测小鼠即将醒来的可能性。

陈瑶说, 该标记并不是闹钟, 无法触发小鼠苏醒的确切时间, 而这可能是由突然的噪声引起的。但它反映了一种增强的生理觉醒状态。

研究人员故意强迫小鼠长时间保持清醒, 之后再让它们入睡。结果显示, 此时 PKA 信号会下降到比正常睡眠更低的水平, 但下降的速率保持不变。由于缺觉小鼠睡得更多、不安的时间更少, 因此在醒来之前, 蛋白质标记有更多不受干扰的时间来下降。陈瑶认为, 这说明 PKA 信号衡量的是累积的睡眠需求的缓解程度, 而不是简单

地充当计时器。

瑞士洛桑大学的 Anita Lüthi 指出, 长久以来, 睡眠研究大多依靠弥散电信号, 比如脑电图记录的信号来解读人类睡眠期间的大脑活动。她认为, 任何能够揭示睡眠过程中分子层面变化的研究都意义重大, 但这并不意味着 PKA 信号可以真正衡量睡眠需求。“蛋白质标记水平反映的也许是动物的警觉程度, 或者预示着在向快速眼动睡眠过渡。”Lüthi 补充说, 心率等其他生理指标同样会呈现出类似的平稳下降趋势, 表明即将进入快速眼动睡眠。

陈瑶团队计划开展后续实验, 实时调控 PKA 的活性, 探究蛋白质标记水平能否直接影响“睡眠债”。 (王方)

相关论文信息:

<https://doi.org/10.64898/2026.07.03.736427>