人工智能医疗器械创新合作平台
ARTIFICIAL INTELLIGENT MEDICAL DEVICE INNOVATION AND COOPERATION PLATFORM

抑郁症数字疗法产品算法性能评估标准

人工智能医疗器械创新合作平台数字疗法工作组

2025 年 10 月

目录

第一章：抑郁症常规诊断与治疗现状 1
1.1 抑郁症诊断现状 1
1.2 抑郁症治疗现状 3
第二章：抑郁症数字疗法产品发展进展 5
2.1 产品形态与发展演进 5
2.2 核心技术应用 7
第三章 抑郁症数字疗法产品算法性能测试 9
3.1 测试数据 9
3.2 测试过程 12
3.3 测试结果 13
总结 15

表目录

表 1　数据集构建 11
表 2　数据预处理标准化方法 12
表 3　测试指标 14
表 4　算法性能多维度评估标准 14

本文件为北京市科技计划项目《精神疾病数字疗法产品测试验证技术服务平台构建及应用》的研究成果，可作为产品设计开发和成品检验的技术参考，不具有强制性。

第一章：抑郁症常规诊断与治疗现状

1.1 抑郁症诊断现状

抑郁症的诊断至今仍是一项复杂的临床挑战，其核心依赖于症状学的评估而非客观的生物学标志物。这种依赖主观报告的模式源于抑郁症的异质性，即其症状表现因人而异，且常与其他精神障碍（如焦虑或创伤后应激障碍）重叠，导致诊断准确性受到限制。目前，缺乏可靠的生物标志物（如血液检测或神经影像学指标）来客观量化抑郁状态，这使得诊断过程高度依赖临床医生的经验和患者自我描述的一致性。这种主观性不仅增加了误诊风险，还可能导致治疗延迟，尤其在高负担医疗环境中更为突出。因此，开发辅助诊断工具已成为改善抑郁症管理的关键方向。

1.1.1 抑郁症主流诊断标准与工具

目前，全球范围内主要采用两大诊断标准体系：**一是《精神障碍诊断与统计手册》第五版（DSM-5）：**由美国精神医学学会（APA）发布，提供了抑郁症的诊断症状清单（如情绪低落、兴趣丧失、睡眠改变等），要求满足至少 5 项症状，且持续两周以上，并导致社会功能受损。**二是《国际疾病分类》第十一次修订本（ICD-11）：**由世界卫生组织（WHO）发布，其诊断标准与 DSM-5 类似，是全球疾病和

死亡统计的通用标准。

在临床实践中，医生通常结合结构化或半结构化的访谈工具来辅助诊断：**一是汉密尔顿抑郁量表（HAMD-17/HAMD-24）：** clinician-reported 量表，是评估抑郁严重程度的“金标准”工具之一，但其对焦虑/躯体化症状权重较高。**二是蒙哥马利-阿斯伯格抑郁评定量表（MADRS）：** 更侧重于抑郁的核心心理症状，在欧洲应用广泛，被认为对治疗后的变化更敏感。**三是患者健康问卷（PHQ-9）：** 基于 DSM-5 标准的 9 项自评量表，因其简洁高效，被广泛应用于初级保健机构和社区中进行筛查和严重程度评估。**四是贝克抑郁量表（BDI-II）：**一款经典的自评量表，涵盖 21 个症状－态度类别，具有良好的信效度。

1.1.2 当前诊断模式的核心挑战

高度主观性与一致性难题: 诊断严重依赖患者的主观描述和医生的临床经验。不同医生对同一患者的诊断一致性（Kappa 值）通常仅在 0.6-0.7 之间，存在较大的主观判断空间。

“诊断延迟”与“隐形”患者群体: 从个体首次出现症状到最终被确诊，平均延迟时间长达 2 至 4 年。大量患者因病耻感、认知不足或医疗资源可及性问题而未能就医，成为“隐形”患者，错过了早期干预的黄金窗口。

共病情况复杂: 抑郁症常与焦虑障碍、强迫症、躯体形式障碍及物质滥用等共病，症状相互交织，给准确诊断和鉴别诊断带来巨大困难。

评估的静态与片段化：传统评估通常在特定时间点进行（如门诊问诊），难以捕捉患者情绪和症状在日常生活环境中的动态波动，无法反映症状的昼夜节律、情境触发因素等丰富信息。

医疗资源分布不均：全球范围内，尤其是中低收入国家和偏远地区，精神科医生和心理治疗师严重短缺，使得专业诊断服务可及性极低。

当前诊断模式的挑战不仅限于主观性，还涉及文化和社会因素。例如，在不同文化背景下，患者对抑郁症状的表达方式可能存在差异，某些群体可能更倾向于报告躯体症状而非情绪问题，这进一步降低了诊断工具（如 PHQ-9）的跨文化效度。此外，医疗系统的碎片化导致诊断信息难以在初级保健和精神专科之间无缝流转，使得早期干预机会流失。未来，结合数字健康记录和人工智能分析可能有助于实现更连续的监测，但需解决数据标准化和隐私问题。

1.2 抑郁症治疗现状

尽管存在有效的治疗方法，但在现实世界的执行过程中仍面临诸多瓶颈。这些瓶颈主要源于医疗资源分配不均、治疗依从性低以及个体差异大等因素。例如，心理治疗和药物治疗虽被临床指南广泛推荐，但实际应用中往往因成本、可及性和社会污名化而受阻。此外，治疗效果的评估通常基于短期临床观察，缺乏长期随访数据，这限制了我们对复发预防和个性化调整的理解。数字疗法的兴起旨在通过技术手段弥补这些差距，但需确保其算法性能经过严格验证，以整合到现有医疗体系中。

1.2.1 主流治疗手段

心理治疗：特别是认知行为疗法（CBT）和人际疗法（IPT），被众多临床指南推荐为轻中度抑郁症的一线治疗方案。其优势在于无药物副作用、复发率较低，但需要患者具备一定的领悟力和执行能力。

药物治疗：选择性 5-羟色胺再摄取抑制剂（SSRIs）和 5-羟色胺和去甲肾上腺素再摄取抑制剂（SNRIs）是常用的抗抑郁药物。它们通过调节大脑神经递质来缓解症状，但存在起效慢（通常 2-4 周）、副作用（如恶心、失眠、性功能障碍）以及个体疗效差异大等问题。

物理治疗：对于难治性抑郁症，改良电休克疗法（MECT）和重复经颅磁刺激（rTMS）等也是有效的选择。

1.2.2 当前治疗模式的核心挑战

治疗可及性障碍：心理治疗师资源稀缺，导致预约等待时间长、费用高昂。药物治疗虽相对可及，但定期复诊开药对于许多患者而言仍是不小负担。

治疗依从性低下：药物治疗中，约 50%的患者在 6 个月内会自行停药，主要原因包括副作用难以耐受、感觉无效或病耻感。心理治疗的高脱落率也是一个普遍问题。

治疗反应的异质性：抑郁症具有高度的异质性，这意味着并非所有患者都对同一治疗方案产生反应。目前缺乏可靠的生物标志物来指导治疗选择，导致医生和患者往往需要进行漫长的“试错”过程，增加了患者的痛苦和治疗成本。

复发率高：抑郁症复发率高达 50%以上。传统的随访模式难以实

现对患者的长期、连续监测，无法及时发现复发的早期迹象并实施干预。

治疗模式的挑战还体现在资源分配上。全球范围内，中低收入国家的精神卫生服务覆盖率不足20%，而高收入国家也面临心理治疗师短缺的问题。这导致许多患者只能依赖药物治疗，但其副作用和个体耐受性差异常引发依从性问题。数字疗法可通过远程交付降低门槛，但需确保其内容基于循证医学，并避免过度依赖自动化而忽视人性化支持。同时，考虑到患者在治疗反应上存在的异质性，开发能够预测个体疗效、从而匹配最佳干预方案的预测模型，是一个重要的探索方向。这通常需要整合多模态数据并持续优化算法，以期实现更精准的个性化治疗。

第二章：抑郁症数字疗法产品发展进展

数字疗法指由软件驱动，基于循证医学证据的干预方案，用于预防、治疗或管理疾病，改善疾病预后。

2.1 产品形态与发展演进

数字疗法指由软件驱动，基于循证医学证据的干预方案，用于预防、治疗或管理疾病，改善疾病预后。这类产品不同于传统健康应用，其核心在于通过临床验证的算法提供个性化干预，并往往需要监管审批（如FDA或NMPA）。在抑郁症领域，数字疗法的发展反映了从信息传递到智能适应的演进，旨在解决传统治疗的可及性和依从性瓶颈。随着人工智能技术的成熟，数字疗法正从辅助工具转向核心治疗

组件，但其有效性高度依赖于算法性能的稳健性。

数字疗法通过“认知情绪训练”技术，用户需识别不同情绪面孔并进行记忆训练，以提升情绪信息处理能力，针对传统抗抑郁药仅部分有效的人群，作为辅助治疗方案。抑郁症 DTx 产品经历了从简单工具到智能平台的演进过程:

第一代：数字化内容平台（信息提供阶段）形态: 提供心理教育文章、音频、视频和简单的情绪追踪工具。特点：本质是“数字化的书本”，缺乏个性化和交互性，用户体验单一。代表：早期的心理健康 APP 如 Pacifica（提供 CBT 思想和情绪追踪）。第一代产品的局限性在于其静态性，无法根据用户反馈动态调整内容，导致参与度较低。例如，简单的情绪追踪工具往往因用户遗忘或厌倦而弃用，这凸显了需要更智能的交互机制。

第二代：结构化数字疗法（标准化干预阶段）形态: 将成熟的循证心理学疗法（如 CBT、ACT 接纳承诺疗法、正念冥想）拆解为结构化的、循序渐进的课程和互动练习，引导用户每周完成特定任务。特点：具备了治疗框架，但干预路径相对固定，个性化程度有限。代表：Pear Therapeutics 的 Somryst（FDA 批准的首个 DTx 产品，用于失眠的 CBT-I，其模式可迁移至抑郁焦虑）。第二代产品通过结构化课程提升了规范性，但标准化路径可能无法适应个体的情绪波动或共病情况。例如，固定时间表的 CBT 练习可能不适合症状急性期的患者，这促使了第三代产品对实时适应的需求。

第三代：AI 驱动的自适应数字疗法（个性化干预阶段）形态:

当前发展的前沿。核心是利用人工智能（AI）和机器学习（ML）算法，实现感知－评估－干预－优化的闭环。特点：**个性化评估：**利用多模态数据（语音、文本、行为、生理）持续、被动地评估用户状态，构建“数字表型（Digital Phenotype）”。**个性化干预：**基于算法模型，动态推荐和调整干预内容、强度、类型和时机，实现“千人千面”的精准治疗。**预测性预警：**通过时序模型预测复发风险和症状恶化趋势，实现前瞻性干预。**增强参与度：**利用游戏化、聊天机器人（Chatbot）等技术提升用户粘性和依从性。代表产品：Mindstrong Health（通过分析智能手机使用模式评估认知功能与情绪）、Woebot Health（基于AI的CBT聊天机器人）。第三代产品的优势在于其闭环设计，能够通过持续数据采集优化干预。然而，这也带来了算法透明度和伦理挑战，如何确保用户数据安全并避免算法偏见。未来方向可能包括融合穿戴设备数据和临床医生反馈，以增强可靠性。

美国约有 30%的抑郁症患者对药物治疗反应不佳。抑郁症 DTx 产品为这部分患者提供新选择。目前，该产品已在英国上市，作为首个欧洲市场推出的处方级数字疗法，需医生或治疗师转诊。因此，数字疗法仍面临说服医生、保险机构及患者持续使用的难题，市场接受度是未来成功的关键。

2.2 核心技术应用

第三代抑郁症数字疗法产品的效能与先进性，根本上由其核心算法引擎的性能所决定。该算法引擎构成了一个从感知、分析到干预的闭环系统，通过处理多模态数据实现动态评估与个性化适应。其技术

应用主要围绕两个关键层面展开：一是对用户状态的精准识别与评估，二是基于评估结果的智能化干预与推荐。这两个层面相辅相成，共同构成了产品实现个性化治疗目标的技术基础。

2.2.1 症状识别与评估算法

语音情感计算（Vocal Biomarkers）：通过智能手机麦克风采集语音片段，提取声学特征（如基频、能量、语速、停顿、抖动率、共振峰），利用深度学习模型（如 CNN, LSTM）识别与抑郁、焦虑相关的语音模式变化（如语调平坦、语速减慢）。

自然语言处理（NLP）与文本分析：一是情感分析，分析用户在日记、与聊天机器人对话中的文本，判断情绪效价（积极/消极）和唤醒度。二是语义分析，识别认知扭曲（如“全或无”思维、过度概括）、自杀风险用语和叙事结构的变化。

行为特征分析（Behavioral Biomarkers）：通过智能手机传感器和使用日志被动监测用户行为，无需主动输入。活动水平：GPS 轨迹、步数反映社会退缩和运动减少。睡眠模式：屏幕开启时间、手机操作时间反映睡眠－觉醒节律紊乱。社交互动：通话记录、短信/社交媒体活动频率反映社交回避。手机使用模式：打字速度、滑屏速度的细微变化可能反映精神运动性迟滞或激越。

2.2.2 干预与推荐算法

个性化推荐系统：基于协同过滤、内容过滤或强化学习算法，为用户推荐最匹配的干预模块（如特定的 CBT 技巧、正念练习、行为激活任务）。

适应性学习路径：根据用户对前序任务的完成度、效果反馈和当前状态，实时调整后续课程的难度、顺序和呈现方式。

数字表型与预测模型：构建用户的个人基线，并利用时间序列分析（如 ARIMA 模型）或生存分析模型，预测未来短期内情绪恶化的概率，及时触发预警和轻量级干预。

核心技术应用的成功取决于多模态数据的整合质量。例如，语音情感计算需在嘈杂环境中保持准确性，而行为特征分析可能受设备差异影响。因此，建议算法测试涵盖多样化的真实场景，以确保泛化能力。此外，自然语言处理模型需训练于多语言数据集，以支持全球化部署，但这需要大规模标注数据，涉及高昂成本。

第三章 抑郁症数字疗法产品算法性能测试

为支持抑郁症数字疗法产品的安全性、有效性和可靠性，建议对其核心算法进行标准化、可重复的性能验证。算法性能测试是评估数字疗法产品能否从实验研究走向临床应用的一项重要工作，其目的在于系统性地评估算法在预设任务上的表现，并识别其局限性与风险。一个较为完整的测试流程通常建议涵盖测试数据的准备与管理、测试过程的执行与监控、以及测试结果的综合分析与解读。

3.1 测试数据

测试数据的质量、规模及代表性是算法性能评估的基石，直接决定了验证结果的可信度与泛化能力。一个构建良好的测试数据集应满

足来源多样、标注规范、划分合理及预处理标准化的要求。

3.1.1 数据来源与采集

建议测试数据尽可能反映产品目标应用场景的真实情况。主要数据来源包括多中心临床研究、真实世界研究以及经过严格筛选的公开数据集。多中心临床研究数据因其具有统一的入排标准和规范的操作流程，通常被视为高质量的数据来源，有助于控制混杂变量，但可能存在与真实世界环境的差异。真实世界数据则能更好地体现实际使用中的复杂性和多样性，但需特别注意数据质量的均一性和伦理合规性。在使用公开数据集进行基准测试时，宜明确其与产品目标人群和任务的相关性。数据的采集过程应获得伦理委员会批准和参与者的知情同意，对数据进行去标识化处理。

3.1.2 数据标注与金标准

算法预测结果的准确性需通过与临床金标准进行比对来验证。因此，数据标注的准确性至关重要。为获得高质量的标注结果，可参考的一种标注方法是：由至少两名经验丰富的精神科主治医师，依据国际公认的半结构化访谈工具(如 SCID-5 或 MINI)达成的一致性诊断。对于抑郁严重程度评估算法，金标准则建议采用临床医生评定的量表分数，如 HAMD-17 或 MADRS。为确保标注质量，建议计算标注者间信度。Kappa 系数是一种用于评估分类任务中，两名或多名评估者之间一致性（信度）的统计指标，计算 Kappa 系数有助于证明标注结果的一致性和可靠性

3.1.3 数据集构建与划分

为避免模型过拟合或评估偏差，建议数据集被严格划分为训练集、验证集和测试集。训练集用于模型参数的训练，验证集用于超参数调优和模型选择，测试集则用于最终的无偏性能评估。测试集应与训练集和验证集保持完全独立，即不能参与模型训练的任何环节。数据集构建如表 1 所示，同时应充分考虑数据集的代表性，确保其在年龄、性别、地域、严重程度、共病情况等关键维度上覆盖目标用户群体，以降低算法偏见的风险。

表 1 数据集构建

数据集类型	主要用途	独立性要求
训练集	用于模型训练与参数优化，样本量需足够大以保证模型学习的充分性和泛化能力基础。	-
验证集	超参数调优与模型选择，用于训练过程中的“质检”，防止对训练集过度拟合，是模型优化迭代的关键参考	与训练集来源一致，但样本独立
测试集	最终性能评估，用于模拟真实应用场景，提供对模型最终性能的公正评估，此评估结果应最接近实际应用效果。	建议与训练集、验证集完全独立

3.1.4 数据预处理标准化

原始数据通常包含噪声和缺失值，需经过标准化的预处理流程。预处理方法应在研究方案中预先明确，并在整个数据集中一致应用。常见的预处理步骤包括缺失值处理、数据增强和特征标准化。具体方法的选择需根据数据类型和算法需求而定，其标准化方法可归纳如表2。

表 2 数据预处理标准化方法

预处理步骤	常见方法	注意事项
缺失值处理	直接删除、均值/中位数/众数填充、基于模型的插补	需记录缺失比例，并在报告中说明所采用的具体方法及理由。
数据增强	SMOTE（合成少数类过采样技术）、加性噪声、时间序列变换	仅限用于训练集，旨在平衡类别或增加数据多样性，严禁在测试集中使用。
特征标准化	Z-score 标准化、最大－最小值归一化	消除不同特征间的量纲影响，标准化参数应从训练集中计算并应用于验证集和测试集。

3.2 测试过程

测试过程建议在受控且可重现的环境下，依据预先定义的指标和流程执行，以支持评估结果的科学性和可比性。

3.2.1 测试环境

测试环境应最大程度地模拟产品真实的使用场景。这包括硬件环境（覆盖主流品牌和型号的移动设备）、软件环境（搭建与开发环境隔离的独立测试服务器，并涵盖不同的操作系统版本）以及网络环境（模拟 4G、5G、Wi-Fi 及弱网条件）。考虑到检验算法鲁棒性和稳定性的需要，建议测试环境配置具备一定的多样性。

3.2.2 测试指标

算法性能建议通过一系列定量指标进行衡量。这些指标宜全面反映算法在分类或回归任务上的表现。核心指标包括准确度、敏感度、特异度、精确度、F1-Score 以及受试者工作特征曲线下面积（AUC）。这些指标的定义和计算公式如下表所示。其中，敏感度和特异度在医疗场景中尤为重要，因为高敏感度有助于减少漏诊（假阴性），高特异度有助于减少误诊（假阳性）。AUC 值则提供了模型在不同分类阈值下整体性能的综合度量，AUC 值越高，表明模型的区分能力越强。如表 3 所示。

3.3 测试结果

测试结果的评估不应仅局限于单一的准确度指标，而需建立一个多维度的评估体系，以确保算法满足临床应用的实际需求。该体系应涵盖以下六个核心维度，如表 4 所示。

最终，算法的价值宜体现在其临床有用性上。这意味着算法输出应能有效辅助临床决策，例如，准确识别出需要紧急干预的高风险患者，或精准推荐个性化的治疗模块。临床有用性的黄金验证标准是通

过前瞻性的随机对照试验，证明集成了该算法的数字疗法产品能够显著改善患者的临床结局（如症状减轻、复发率下降、生活质量提升）。

表 3 测试指标

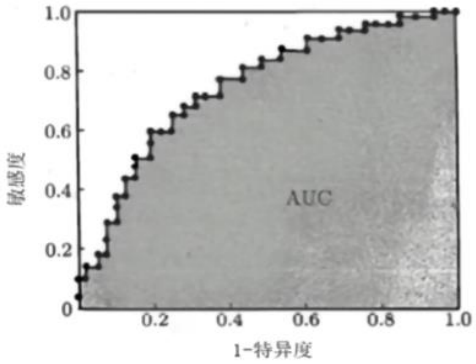
性能指标	计算公式/说明
准确度	$\frac{TP+TN}{TP+TN+FP+FN}\times 100\%$
敏感度	$\frac{TP}{TP+FN}\times 100\%$
特异度	$\frac{TN}{TN+FP}\times 100\%$
精确度	$\frac{TP}{TP+FP}\times 100\%$
F-Score	$(1+\beta^2)\times\frac{精确度\times敏感度}{(\beta^2\times精确度)+敏感度}$ 其中β为平衡系数，一般情况下取值为1。
ROC 曲线	绘图过程：以每一个检测结果作为可能的诊断界值，计算得到相应的真阳率（敏感度）和假阳率（1-特异度），以假阳率为横坐标，以真阳率为纵坐标绘制而成的曲线。  图1 ROC曲线绘制示意图
AUC 值	计算公式：假定ROC曲线是由坐标为$\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$的点按序连接而形成，如图1所示，则AUC可估算为： $AUC=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\times(y_i+y_{i+1})$

表 4 算法性能多维度评估标准

评估维度	核心要求	具体评估方法
准确性	预测结果与金标准高度一致。	关键指标（如 AUC）应达到预设阈值，敏感度与特异度需均衡，避免以牺牲一方为代价提升另一方。
鲁棒性	在数据存在噪声、缺失或分布变化时性能保持稳定。	在测试集中引入合成噪声或使用来自不同采集源的数据，性能下降幅度不影响产品使用。
泛化性	在未见过的数据上表现良好。	在独立的外部验证集（如来自不同医院或地区的数据）上性能与内部测试集相比无显著差异。
公平性	在不同人口统计学亚组间性能差异最小。	分析算法在年龄、性别、种族等亚组上的性能指标（如敏感度、特异度）
效率	计算速度快，资源消耗低。	在目标硬件上，算法的推理时间应满足实时性要求（如秒级响应），且内存和电量消耗在可接受范围内。
可解释性	决策过程透明，可供人类理解。	提供特征重要性分析或使用可解释 AI 技术（如 SHAP、LIME）来阐明模型决策的依据。

总结

抑郁症数字疗法产品代表了精神健康领域的技术前沿，其发展有望弥补传统诊断和治疗的不足。本报告系统梳理了抑郁症的常规诊疗现状、数字疗法产品的演进及核心技术，并重点提出一套算法性能测试的建议性流程框架。测试结果表明，稳健的算法是数字疗法安全有效的基石，但需在准确性、鲁棒性、泛化性、公平性、效率和可解释性之间取得平衡。未来，随着法规完善和技术进步，数字疗法可能成

为集成式护理的一部分，但需持续优化算法以应对真实世界的复杂性。建议行业加强合作，建立共享数据集和基准测试平台，推动创新成果的临床转化。