

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2024 年 3 月 28 日 (28.03.2024)



(10) 国际公布号
WO 2024/060508 A1

(51) 国际专利分类号:
G16H 50/20 (2018.01)

(21) 国际申请号: PCT/CN2023/077386

(22) 国际申请日: 2023 年 2 月 21 日 (21.02.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202211146285.0 2022年9月20日 (20.09.2022) CN

(71) 申请人: 浙江大学 (ZHEJIANG UNIVERSITY) [CN/CN]; 中国浙江省杭州市西湖区余杭塘路 866 号, Zhejiang 310058 (CN)。

(72) 发明人: 李昊旻 (LI, Haomin); 中国浙江省杭州市滨江区滨盛路 3333 号, Zhejiang 310052 (CN)。杨建 (YANG, Jian); 中国浙江省杭州市西湖区浙大路 38 号, Zhejiang 310027 (CN)。段会龙 (DUAN, Huilong); 中国浙江省杭州市西湖区浙大路 38 号, Zhejiang 310027 (CN)。舒强 (SHU,

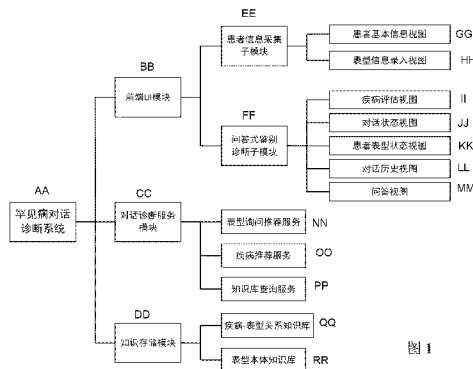
Qiang); 中国浙江省杭州市滨江区滨盛路 3333 号, Zhejiang 310052 (CN)。

(74) 代理人: 杭州天勤知识产权代理有限公司 (HANGZHOU TIANQIN INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国浙江省杭州市西湖区竞舟路 1 号筑品金座 501 室, Zhejiang 310013 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(54) Title: KNOWLEDGE-DRIVEN SYSTEM AND METHOD FOR VISUALIZED QUESTION-AND-ANSWER ASSISTED DIFFERENTIAL DIAGNOSIS OF RARE DISEASE

(54) 发明名称: 知识驱动的罕见病可视化问答式辅助鉴别诊断系统及方法



AA Dialogue diagnosis system for rare diseases
BB Front-end UI module
CC Dialogue diagnosis service module
DD Knowledge storage module
EE Patient information collection sub-module
FF Question-and-answer differential diagnosis sub-module
GG Patient basic information view
HH Phenotypic information entry view
II Disease assessment view
JJ Dialogue state view
KK Patient phenotypic state view
LL Dialogue history view
MM Question-and-answer view
NN Phenotype inquiry recommendation service
OO Disease recommendation service
PP Knowledge base query service
QQ Disease-phenotype relation knowledge base
RR Phenotype ontology knowledge base

图 1

(57) Abstract: A system and method for visualized question-and-answer assisted differential diagnosis of a rare disease. The system comprises: a front-end UI module, which collects basic information and phenotypic information entry of a patient to acquire an initial self-report of the patient; a dialogue diagnosis service module, which provides a phenotype inquiry recommendation service, a disease diagnosis recommendation service and a knowledge base query service by means of an API, wherein the phenotype inquiry recommendation service is to recommend an appropriate phenotype on the basis of the current dialogue state so as to achieve relatively high dialogue benefits, the disease diagnosis recommendation service is to recommend an appropriate disease on the basis of the initial self-report of the patient and new phenotypic information generated during a dialogue, and the knowledge base query service is to query and call, according to requirements, knowledge stored in a knowledge storage module; and the knowledge storage module, which is used for storing a human phenotype ontology knowledge base and a disease-phenotype relation knowledge base. The system is capable of collecting additional symptoms beyond a self-report, and displaying auxiliary information for differential diagnosis using a visualization technique.

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种罕见病可视化问答式辅助鉴别诊断系统及方法, 系统包括: 前端UI模块, 采集患者基本信息和表型信息录入以获取患者的初始自我报告; 对话诊断服务模块, 通过API提供表型询问推荐服务、疾病诊断推荐服务以及知识库查询服务; 表型询问推荐服务是根据当前的对话状态推荐合适的表型以达到较高的对话收益; 疾病诊断推荐服务是根据患者初始自我报告和对话过程中产生的新的表型信息推荐合适的疾病; 知识库查询服务是根据需要, 查询并调用知识存储模块中存储的知识; 知识存储模块, 用于存储人类表型本体知识库以及疾病-表型关系知识库。该系统能够收集超出其自我报告的附加症状并利用可视化技术显示鉴别诊断的辅助信息。

知识驱动的罕见病可视化问答式辅助鉴别诊断及方法

技术领域

本发明涉及智能辅助诊断技术领域，尤其涉及一种知识驱动的罕见病可视化问答式辅助鉴别诊断及方法。

背景技术

罕见病，又称孤儿病，是指那些发病率极低的疾病。世界卫生组织（WHO）将罕见病定义为每10000人中患病率为6.5到10人的疾病。《中国罕见病定义研究报告2021》将“新生儿发病率小于1/万、患病率小于1/万、患病人数小于14万的疾病”列入罕见病。国际确认的罕见病有近7000种，大约占人类疾病的10%。

罕见病诊疗困难，具体表现为确诊困难、治疗成本高和药物研发困难。罕见病种类繁多且通常具有复杂的表型特征和分子特征，这使得罕见病的诊断难度要高于其他种类的疾病。根据欧洲罕见病组织（EURORDIS）的一项调查，四分之一的罕见遗传病患者需要5到30年的时间才能被诊断出来，且初诊误诊率高达40%。这些误诊会导致医疗资源浪费和救治时间的延误，甚至直接危害患者生命。因此，如何实施快速有效的诊断成为罕见病领域的一项重要课题。

目前许多罕见病知识库被建立起来。Orphanet于1997年在法国成立，旨在收集稀缺的罕见病知识，提高对罕见病患者的诊断、护理和治疗水平。迄今为止，Orphanet已成为罕见病诊断的重要参考来源。人类表型本体（Human Phenotype Ontology, HPO）在2008年由Robinson教授等人提出，是人类疾病中已发现的表型异常的标准化词汇表，广泛应用罕见疾病表型基因型诊断分析。临床医生可以利用罕见病知识库，根据患者的表型特征去筛选患者潜在疾病，然而这一方法在临床实践中有颇多的局限性：一是在临床采集患者表型过程中经常会采集到不准确表型甚至引入噪声表型；二是因遗漏表型导致患者表型信息不完整；三是罕见病之间表型交叠较多导致鉴别诊断难度较大。这些问题在各种罕见病诊断场景中都会对诊断效果

造成不利影响。

当前，对话式聊天机器人或面向任务的对话系统引起越来越多的关注，尤其在医疗领域，以疾病诊断为目标的智能对话系统能够进行主动的患者额外症状信息收集以达到更好的疾病诊断效果。现有的医疗诊断对话系统大多依赖于数据驱动的机器学习方法。然而，在罕见病领域，由于疾病种类繁多且病例资源匮乏，使得真实数据驱动的学习难以开展。在这一背景下，整合知识库开展知识驱动的医疗诊断对话成为一种潜在的有效的办法。

发明内容

本发明提供了一种知识驱动的罕见病可视化问答式辅助鉴别诊断系统，该系统能够通过与患者/医生对话的方式收集超出其自我报告的附加症状并利用可视化技术显示鉴别诊断的辅助信息，进一步缓解罕见病诊断难的问题。

本发明技术方案如下：

一种罕见病可视化问答式辅助鉴别诊断系统，包括：

前端 UI 模块，用于进行人机交互，采集患者基本信息和表型信息录入以获取患者的初始自我报告，并利用多个维度可视化视图展现相关信息供临床人员进行鉴别诊断；

对话诊断服务模块，通过 API（Application Programming Interface）提供表型询问推荐服务、疾病诊断推荐服务以及知识库查询服务；表型询问推荐服务是根据当前的对话状态推荐合适的表型以达到较高的对话收益，用户对系统询问的表型可以选择回答“存在”、“不存在”或者“不确定”；疾病诊断推荐服务是根据前端 UI 模块采集的患者初始自我报告和对话过程中产生的新的表型信息推荐合适的疾病；知识库查询服务是根据表型询问推荐服务和疾病诊断推荐服务的需要，查询调用知识存储模块中存储的知识；

知识存储模块，用于存储人类表型本体知识库以及疾病-表型关系知识库。

所述的人类表型本体知识库为人类表型本体 (Human Phenotype Ontology, HPO)，HPO 是基于本体论思想构建的人类异常表型的标准化结构化本体。HPO 被构造为具有层级结构的有向无环图，HPO 术语之间

通过“is_a”的关系进行连接，子项比父项拥有更为精确的定义。截至 2022 年 2 月份，HPO 总共有了 16226 条现行的表型异常表型（HP:0000118-Phenotypic abnormality 及其后代表型），每条异常表型都有一个唯一的 HPO 编码。

所述的疾病-表型关系知识库为罕见病权威知识库 Orphanet，罕见病权威知识库 Orphanet 中包含疾病流行性、平均发病死亡年龄、疾病表型频率特征以及致病基因信息。

知识存储模块为对话诊断服务模块提供知识驱动。

前端 UI 模块包括：

患者信息采集子模块，包括患者基本信息视图和表型信息录入视图；

可视化的问答式鉴别诊断子模块，疾病评估视图、对话状态视图、患者表型状态视图、对话历史视图以及问答视图，利用多个维度可视化视图展现相关信息供临床人员进行鉴别诊断。

患者信息采集子模块包括患者基本信息视图和表型信息录入视图，用户可以在该模块输入患者年龄、出生时性别、阳性表型以及阴性表型等信息。

问答式鉴别诊断子模块包括：

疾病评估视图，用于展示疾病患病率、后验概率以及表型匹配情况；

对话状态视图，以雷达图的形式展示当前优势最大的疾病的后验概率（标签为 P）、表型匹配情况（标签为 M）以及疾病推荐结果的不确定性度量（信息熵标签为 E，基尼指数标签为 G）；

患者表型状态视图，为基于人类表型本体并利用注释传播规则绘制的反应患者表型特征（阳性、阴性、不确定）的旭日图；

对话历史视图以及问答视图，用于展示系统询问的表型信息并为用户提供反馈按钮（存在、不存在、不确定）。

问答式鉴别诊断子模块以多个维度可视化视图展现相关信息供临床人员进行鉴别诊断。

对话诊断服务模块，通过 API（Application Programming Interface）提供表型询问推荐服务、疾病诊断推荐服务以及知识库查询服务。

疾病推荐服务包括疾病先验概率估计和疾病后验概率计算，对话诊断

服务模块在对话结束时推荐后验概率最大的疾病；

疾病先验概率估计的方法包括：

(a-1) 将 Orphanet 知识库中的疾病流行性类型统一转换为点患病率；

(a-2) 在罕见病范围内，根据疾病的点患病率计算疾病先验概率，公式为：

$$Prob_{pretest}(D) = \frac{PointPrev_D}{\sum PointPrev} \#(1)$$

其中， $Prob_{pretest}(D)$ 代表疾病 D 的先验概率（Pretest Probability）， $PointPrev_D$ 代表疾病 D 的点患病率（Point Prevalence）， $\sum PointPrev$ 表示所有罕见病点患病率之和；

疾病后验概率计算的方法包括：

(b-1) 胜算（Odds）与概率（Prob）之间的换算关系为：

$$Odds = \frac{Prob}{1 - Prob}, Prob = \frac{Odds}{1 + Odds} \#(2)$$

根据公式（2）将疾病先验概率转换为疾病先验胜算；

(b-2) 根据疾病表型频率注释和注释传播规则依照公式（3）计算表型信息的复合似然比：

$$CompositeLR(P, D) = \frac{Pr(p_1, p_2, \dots, p_n | D)}{Pr(p_1, p_2, \dots, p_n | \bar{D})} = \prod_{i=1}^n \frac{Pr(p_i | D)}{Pr(p_i | \bar{D})} \#(3)$$

其中， P 代表已采集的表型集合， $(p_1, p_2, \dots, p_n) \in P$ ， $\bar{D}$ 代表除罕见病 D 以外的所有罕见病， $Pr(p_i | D)$ 表示表型 p_i 在疾病 D 中的出现频率，表型集合 P 的复合似然比被计算为单个表型 p_i 似然比的乘积；

(b-3) 根据公式（4）计算疾病的后验胜算 $Odds_{posterior}(P, D)$ ：

$$Odds_{posterior}(P, D) = Odds_{pretest}(D) * CompositeLR(P, D) \#(4)$$

其中， $Odds_{pretest}(D)$ 代表罕见病 D 的先验胜算；

(b-4) 根据公式（5）计算疾病后验概率 $Prob_{posterior}(D)$ ：

$$Prob_{posterior}(D) = \frac{Odds_{posterior}(D)}{1 + Odds_{posterior}(D)} \#(5)$$

。

步骤（a-1）中，将疾病流行性类型统一转换为点患病率的方法为：

①如果疾病的流行性类型为点患病率，优先选择患病率地理区域为“世界范围内”的条目，使用该条目的数值型点患病率或者类别型点患病率

的中值；如果患病率地理区域没有“世界范围内”的条目但有以国家或地区统计的患病率条目，则使用以国家或地区统计的患病率条目的点患病率平均值作为该疾病的点患病率；

②如果疾病的流行性类型为年发病率，则从Orphanet库获取该疾病的平均发病年龄以及死亡年龄信息，将发病年龄到死亡年龄作为疾病的持续时长，用年发病率乘以疾病持续年份估算该疾病的点患病率；

③如果疾病的流行性类型为病例数，则直接使用病例数除以世界总人口作为该疾病的点患病率；

④如果疾病的流行性类型为出生时患病率，且疾病发病年龄在新生儿期，则使用出生时患病率乘以该疾病持续年龄段的人口在世界总人口中的比例值作为疾病的点患病率；

⑤如果疾病的流行性类型为终生患病率，则使用默认的百万分之一作为该疾病的点患病率。

针对临床采集的患者表型不准确有噪声的问题，本发明在计算表型似然比时会在人类表型本体上应用注释传播规则，使疾病-表型关系的分析更加高效。

优选的，所述的注释传播规则为：患者的阳性表型可以推断出其祖先表型在患者处也为阳性，患者的阴性表型可以推断出其后代表型在患者处均为阴性。

表型推荐服务根据当前的疾病后验概率分布情况和疾病-表型注释关系，选择期望收益最大的表型由系统进行询问。

表型推荐服务包括：

表型的信息增益（Info Gain）计算，包括：

（A-1）使用信息熵（Information Entropy）来表示疾病推荐结果的不确定性，以公式（6）定义有分类权重的信息熵 $Entropy(W)$ ：

$$Entropy(W) = - \sum_{w_i \in W} \frac{w_i}{W_{sum}} * \log \frac{w_i}{W_{sum}} \quad \#(6)$$

其中， W 是一个权重集合， w_i 表示权重集合 W 中单一元素的权值， W_{sum} 表示 W 中所有元素权值之和；

当前对话状态下的疾病后验概率的信息熵 $Entropy_{cur}(D)$ 为：

$$Entropy_{cur}(D) = Entropy(\{Prob_{posterior}(d) | d \in D\}) \quad \#(7)$$

其中, D 表示所有罕见病的集合, $Prob_{posterior}(d)$ 表示疾病 d 的后验概率;

(A-2) 单一表型条件下的信息增益为:

$$InfoGain(p, D) = Entropy_{cur}(D) - \gamma * Entropy(\{Prob_{posterior}(d) * Pr(p|d) | d \in D\}) - (1 - \gamma) * Entropy(\{Prob_{posterior}(d) * Pr(\bar{p}|d) | d \in D\}) \quad \#(8)$$

其中, 第一项为当前对话状态下的疾病后验概率的信息熵, 第二项为表型 p 为阳性时的条件熵, 第三项为表型 p 为阴性时的条件熵, D 表示所有罕见病的集合, $Prob_{posterior}(d)$ 表示疾病 d 的后验概率, γ 为阳性表型分支权重, 即 $\{Prob_{posterior}(d) * Pr(p|d) | d \in D\}$ 集合元素之和, $Pr(p|d)$ 表示表型 p 在疾病 d 中的出现频率, $Pr(\bar{p}|d)$ 表示表型 p 不在疾病 d 中出现的频率; 表型的基尼指数计算, 包括:

(B-1) 使用基尼指数 (Gini Index) 来表示疾病推荐结果的纯度, 以公式 (9) 定义有分类权重的基尼指数 $Gini(W)$:

$$Gini(W) = 1 - \sum_{w_i \in W} \left(\frac{w_i}{W_{sum}} \right)^2 \quad \#(9)$$

其中, W 是一个权重集合, w_i 表示权重集合 W 中单一元素的权值, W_{sum} 表示 W 中所有元素权值之和;

当前对话状态下的疾病后验概率的基尼指数为:

$$Gini_{cur}(D) = Gini(\{Prob_{posterior}(d) | d \in D\}) \quad \#(10)$$

其中, D 表示所有罕见病的集合, $Prob_{posterior}(d)$ 表示疾病 d 的后验概率;

(B-2) 单一表型条件下的基尼指数增益为:

$$GiniGain(p, D) = Gini_{cur}(D) - \gamma * Gini(\{Prob_{posterior}(d) * Pr(p|d) | d \in D\}) - (1 - \gamma) * Gini(\{Prob_{posterior}(d) * Pr(\bar{p}|d) | d \in D\}) \quad \#(11)$$

其中, 第一项为当前对话状态下的疾病后验概率基尼指数, 第二项为表型 p 为阳性时的条件基尼指数, 第三项为表型 p 为阴性时的条件基尼指数, 其余参数定义等同公式 (8);

将信息增益和基尼指数进行自适应融合, 获得融合度量AIGGI (Adaptive Information Gain and Gini Index):

$$AIGGI(p, D) = \alpha * InfoGain(p, D) + (1 - \alpha) * GiniGain(p, D) \quad (12)$$

其中， D 表示所有罕见病的集合， p 表示单一表型， α 为反应当前对话状态下疾病后验概率分布的不确定程度或杂乱度， α 使用 $Gini_{cur}(D)$ 来表示；

系统根据当前对话状态选择融合度量 AIGGI 最大的表型进行询问，用户针对该表型选择“存在”、“不存在”或者“不确定”的回答。

优选的，为了提高对话的效率，在表型询问推荐时应用注释转播规则来过滤掉一些表型，包括：

已汇报的表型不会在接下来的对话轮次中被询问；

对于已汇报的阳性表型，默认其祖先表型均为阳性，将其祖先表型过滤掉，不再询问；

对于已汇报的阴性表型，默认其后代表型均阴性，将其后代表型过滤掉，不再询问。

本发明还提供了采用罕见病可视化问答式辅助鉴别诊断系统进行辅助鉴别诊断的方法，包括：

(1) 用户在前端 UI 模块的患者信息采集子模块中录入患者信息，包括患者年龄、出生时性别、阳性表型以及阴性表型；

(2) 对话诊断服务模块根据当前对话状态选择融合度量 AIGGI 最大的表型进行询问，用户针对该表型选择“存在”、“不存在”或者“不确定”的回答后，之后进入下一轮对话；

问答过程中，用户可以在前端 UI 模块的可视化的问答式鉴别诊断子模块中获得包括疾病优先级评估、当前对话状态、患者表型状态以及对话历史等多种可视化信息；

(3) 重复步骤 (2)，直到疾病推荐结果的不确定性低于阈值或者达到最大对话轮次时，系统根据疾病后验概率给出疾病推荐并结束对话。

本发明的核心是疾病推荐服务，其算法核心是如何在数万种表型中获得最有效的表型问答，以最优路径逼近真实诊断。其中重要贡献之一是将一种基于贝叶斯医学检验计算疾病后验概率的方法，应用于对话诊断服务模块中的疾病推荐服务，包括疾病先验概率估计和疾病后验概率计算两个步骤。首先本发明设计了不同流行性类型的转换方法，将异构的疾病流行性知识同质化为点患病率，并根据其计算疾病先验概率。使用似然比

(Likelihood Ratio) 指标作为单个表型条件对疾病优先级贡献程度的度量, 累乘得到全部表型条件的复合似然比, 进而得到疾病在表型基础上的后验概率。

本发明的另一个核心是自适应融合信息增益和基尼指数的表型推荐方法, 应用于对话诊断服务模块中的表型询问推荐服务。该方法的目的是根据当前的对话状态, 即疾病后验概率分布情况和疾病-表型注释关系, 选择期望收益最大的表型由系统进行询问。该方法包括表型的信息增益 (Info Gain) 计算、基尼指数增益 (Gini index Gain) 计算以及二者的自适应融合度量 AIGGI (Adaptive Information Gain and Gini Index) 计算三个方面。本发明使用疾病后验概率作为分类权重计算单个表型的信息增益和基尼指数增益, 根据疾病推荐结果的不确定程度将二者进行自适应融合, 由系统推荐 AIGGI 最大的表型作为询问对象。

与现有技术相比, 本发明的有益效果为:

(1) 相较于基于知识库的表型关键词检索推荐疾病以及常见的基于患者相似性的疾病推荐方法, 本发明采用表型问答的形式, 通过人机交互进行主动的患者额外症状信息收集以丰富患者表型信息, 对话过程中可以启发医生注意到之前漏采的患者表型, 最终实现更有效的疾病诊断。

(2) 针对临床采集的患者表型不准确有噪声的问题, 本发明设计了一种基于贝叶斯医学检验计算疾病后验概率的方法, 该方法在人类表型本体上应用注释传播规则来提高疾病推荐的鲁棒性。

(3) 患病率是影响疾病诊断的一个非常重要的因素, 而大多数疾病推荐方法没有考虑患病率。本发明针对疾病流行性知识的异构性问题, 设计了一套流行性类型转换计算方法, 将疾病流行性类型统一为点患病率, 以实现同质性, 并根据点患病率计算出疾病先验概率, 将其应用到疾病推荐方法中。

(4) 本发明设计了一种自适应融合信息增益和基尼指数的表型推荐方法。根据疾病推荐结果的不确定程度, 将表型条件的信息增益和基尼指数增益两种度量进行自适应融合, 得到更高效的表型问答。

附图说明

图 1 为罕见病可视化问答式辅助鉴别诊断系统的模块结构示意图;

图 2 为罕见病可视化问答式辅助鉴别诊断状态雷达图；

图 3 为罕见病可视化问答式辅助鉴别诊断系统的问答流程图示意。

具体实施方式

本发明设计了一种知识驱动的罕见病对话诊断系统，能够通过和用户（患者/家属/医生）对话的方式收集超出其自我报告的附加表型，并基于此自动开展评估并给出疾病诊断。系统以 Orphanet 罕见病知识库和人类表型本体作为先验知识，在初始患者信息录入后由系统进行表型询问，系统根据实时对话状态选择期望收益最大的表型，并由用户给出判断，迭代以上过程实现多轮问答，从而丰富患者的结构化表型数据采集并提高疾病诊断效果。

(一)罕见病对话诊断系统模块介绍

本发明设计的罕见病问答对话诊断系统包括前端 UI（User Interface）模块、对话诊断服务模块和知识存储模块，如图 1 所示，其中：

前端 UI 模块，用以进行人机交互，具体包括患者信息采集子模块和可视化的问答式鉴别诊断子模块。患者信息采集子模块包括患者基本信息视图和表型信息录入视图，用户可以在该模块输入患者年龄、出生时性别、阳性表型以及阴性表型等信息。问答式鉴别诊断子模块包括疾病评估视图、对话状态视图、患者表型状态视图、对话历史视图以及问答视图。疾病评估视图展示疾病患病率、后验概率以及表型匹配情况；对话状态视图以雷达图的形式展示当前优势最大的疾病的后验概率（标签为 P）、表型匹配情况（标签为 M）以及疾病推荐结果的不确定性度量（信息熵标签为 E，基尼指数标签为 G），如图 2 所示，雷达图阴影面积越大代表疾病推荐结果的不确定性越低；患者表型状态视图为一个基于人类表型本体绘制的反应患者表型特征（阳性、阴性、不确定）的旭日图，并利用了注释传播规则；问答视图为对话诊断子模块的核心交互模块，展示系统询问的表型信息并为用户提供反馈按钮（存在、不存在、不确定）。问答式鉴别诊断子模块以多个维度可视化视图展现相关信息供临床人员进行鉴别诊断。

对话诊断服务模块，主要是通过 API（Application Programming Interface）提供表型询问推荐服务、疾病诊断推荐服务以及知识库查询服务。表型询问推荐服务是根据当前的对话状态推荐合适的表型以达到较高

的对话收益，用户对系统询问的表型可以选择回答“存在”、“不存在”或者“不确定”；疾病诊断推荐服务是根据患者初始的自我报告和对话过程中产生的新的表型信息推荐合适的疾病。我们为表型询问和疾病诊断设计了相应的算法，作为本发明的两个重要附加方面。

知识存储模块，存储从罕见病权威知识库 Orphanet 中获取的疾病流行性、平均发病死亡年龄、疾病表型频率特征以及致病基因等信息。截至 2022 年 2 月份，Orphanet 知识库中总共统计了 4257 个带有表型特征的罕见病，这 4257 个罕见病为本系统的目标推荐疾病集合。Orphanet 提供 XML 文件格式的知识图谱，将其解析后存储到关系型数据库 MySQL 中。

除 Orphanet 知识库之外，本系统还整合了人类表型本体 (Human Phenotype Ontology, HPO)，HPO 是基于本体论思想构建的人类异常表型的标准化结构化本体。HPO 被构造为具有层级结构的有向无环图，HPO 术语之间通过“is_a”的关系进行连接，子项比父项拥有更为精确的定义。截至 2022 年 2 月份，HPO 总共有了 16226 条现行的表型异常表型 (HP:0000118-Phenotypic abnormality 及其后代表型)，每条异常表型都有一个唯一的 HPO 编码。

人类表型本体适用注释传播规则 (Annotation Propagation Rule)，即患者的阳性表型可以推断出其祖先表型在患者处也为阳性，患者的阴性表型可以推断出其后代表型在患者处均为阴性，该规则在疾病后验概率计算和表型问答推荐中发挥了巨大作用。

(二)基于贝叶斯医学检验计算疾病后验概率的方法

本发明的重要附加方面之一是一种基于贝叶斯医学检验计算疾病后验概率的方法，该方法应用于对话诊断服务模块中的疾病推荐服务，包括疾病先验概率估计和疾病后验概率计算两个步骤。我们通过设计流行性类型转换方法，将异构的疾病流行性知识同质化为点患病率，并根据其计算疾病先验概率。然后，我们使用似然比 (Likelihood Ratio, LR) 指标作为单个表型条件对疾病优先级贡献程度的度量，将其累乘得到全部表型条件的复合似然比，进而得到疾病的后验概率。详细实施方式如下：

(1) 疾病先验概率估计。

Orphanet 知识库中的疾病流行性知识有明显的异构性，一个罕见病可

能被注释为五种流行性类型（点患病率、出生时患病率、终生患病率、年发病率以及病例数）中的一个或几个。通过设计流行性类型转换计算方法，将疾病流行性类型统一为点患病率，以实现同质性。然后，在罕见病范围内根据疾病点患病率计算先验概率：

$$Prob_{pretest}(D) = \frac{PointPrev_D}{\sum PointPrev} \#(1)$$

其中， $Prob_{pretest}(D)$ 代表疾病 D 的先验概率（Pretest Probability）， $PointPrev_D$ 代表疾病 D 的点患病率（Point Prevalence）， $\sum PointPrev$ 表示所有罕见病点患病率之和。

具体的罕见病患病率信息同质化方法如下，从五种流行性类型的知识中计算出一个代表点患病率的具体值：

如果疾病的流行性类型为点患病率，优先选择患病率地理区域为“世界范围内”的条目，使用该条目的数值型点患病率或者类别型点患病率的中值。如果疾病没有“世界范围内”的条目但有以国家或地区统计的患病率条目，则使用这些条目的点患病率平均值作为该疾病的点患病率。

如果疾病的流行性类型为年发病率，我们需要额外从Orphanet获取该疾病的平均发病年龄以及死亡年龄信息。由于罕见病难以治愈，因此我们将发病年龄到死亡年龄作为疾病的持续时长，用年发病率乘以疾病持续年份估算疾病的点患病率。

如果疾病的流行性类型为病例数，我们直接使用病例数除以世界总人口作为疾病的点患病率。

如果疾病的流行性类型为出生时患病率，且疾病发病年龄在新生儿期，我们使用出生时患病率乘以疾病持续年龄段的人口在世界总人口中的比例值作为疾病的点患病率。

最后，对于疾病的流行性类型为终生患病率的情况，由于终生患病率是在某个时间段里患病的人数在目标人群中的比例，包括测量时间点目前患病人数和以前患病但是现在痊愈的人数，因此该类型很难对应到点患病率上，我们使用默认的百万分之一作为该疾病的点患病率。

（2）疾病后验概率计算。

为了更好地提供单个表型观察结果对疾病优先级的贡献程度的度量，我们使用基于贝叶斯定理的似然比计算方法来估计单一表型的LR以及疾

病的后验概率。首先根据公式（2）计算先验胜算，根据疾病表型频率注释和注释传播规则依照公式（3）计算表型信息的复合似然比，然后根据公式（4）计算后测胜算，最后根据公式（5）得到疾病的后验概率。

胜算（Odds）与概率（Prob）之间的换算关系：

$$Odds = \frac{Prob}{1 - Prob}, Prob = \frac{Odds}{1 + Odds} \#(2)$$

多表型的复合似然比（Composite Likelihood Ratio）计算公式：

$$CompositeLR(P, D) = \frac{Pr(p_1, p_2, \dots, p_n | D)}{Pr(p_1, p_2, \dots, p_n | \bar{D})} = \prod_{i=1}^n \frac{Pr(p_i | D)}{Pr(p_i | \bar{D})} \#(3)$$

其中， P 代表已采集的表型集合， $(p_1, p_2, \dots, p_n) \in P$ ， $\bar{D}$ 代表除罕见病 D 以外的所有罕见病， $Pr(p_i | D)$ 表示表型 p_i 在疾病 D 中的出现频率，表型集合 P 的复合似然比被计算为单个表型 p_i 似然比的乘积。

后验胜算计算公式：

$$Odds_{posterior}(P, D) = Odds_{pretest}(D) * CompositeLR(P, D) \#(4)$$

其中， $Odds_{pretest}(D)$ 代表罕见病 D 的先验胜算。

最后是后验概率的计算：

$$Prob_{posterior}(D) = \frac{Odds_{posterior}(D)}{1 + Odds_{posterior}(D)} \#(5)$$

值得一提的是，针对临床采集的患者表型不准确有噪声的问题，本方法在计算表型似然比时会在人类表型本体上应用注释传播规则，使疾病-表型关系的分析更加高效。具体来讲，对于疾病注释的相关表型，可以推断疾病也与该表型的祖先表型相关。如果录入的一个阳性的表型信息为疾病相关表型的祖先表型，那么该阳性表型对该疾病的似然比度量应该为积极的，即似然比大于1。如果录入的一个阴性的表型信息为疾病相关表型的祖先表型，那么该阴性表型对该疾病的似然比度量应该为消极的，即似然比小于1。

(三)自适应融合信息增益和基尼指数的表型推荐方法

本发明的另一个重要附加方面为一种自适应融合信息增益和基尼指数的表型推荐方法，该方法应用在对话诊断服务模块中的表型推荐服务。该方法的目的是根据当前的对话状态，即疾病后验概率分布情况和疾病-表型注释关系，选择期望收益最大的表型由系统进行询问。该方法包括表

型的信息增益 (Info Gain) 计算、基尼指数增益 (Gini Index Gain) 计算以及二者的自适应融合度量 AIGGI (Adaptive Information Gain and Gini Index) 计算三个方面。我们使用疾病后验概率所谓分类权重计算单个表型的信息增益和基尼指数增益, 根据疾病推荐结果的不确定程度将二者进行自适应融合, 推荐 AIGGI 最大的表型作为询问对象。在问答诊断前期或优势疾病后验概率不够突出时, 系统更侧重于询问期望信息增益较大的表型来降低疾病诊断的不确定性; 在问答诊断后期或优势疾病后验概率较为突出时, 系统更侧重于询问期望基尼指数较小的表型来进一步提升疾病诊断的纯度。

(1) 表型问答的信息增益计算。

本发明使用信息熵 (Information Entropy) 来表示疾病推荐结果的不确定性。信息熵关联于系统信息含量, 能够表征样本集合的不确定性, 信息熵侧重于信息表示, 对较乱的数据具有更强的表现力, 经典的决策树算法 ID3 就是以信息熵为度量设计的。我们将疾病后验概率作为分类的权重计算当前状态的信息熵, 如公式 (7) 所示。如果疾病后验概率分布均匀, 说明疾病诊断的不确定性较高, 信息熵值较大。

首先定义基本的有分类权重的信息熵 $Entropy(W)$:

$$Entropy(W) = - \sum_{w_i \in W} \frac{w_i}{W_{sum}} * \log \frac{w_i}{W_{sum}} \quad \#(6)$$

其中, W 是一个权重集合, w_i 表示权重集合 W 中单一元素的权值, W_{sum} 表示 W 中所有元素权值之和。

当前对话状态下的疾病后验概率信息熵 $Entropy_{cur}(D)$ 为:

$$Entropy_{cur}(D) = Entropy(\{Prob_{posterior}(d) | d \in D\}) \quad \#(7)$$

其中, D 表示所有罕见病的集合, $Prob_{posterior}(d)$ 表示疾病 d 的后验概率。

单一表型条件下的信息增益为:

$$\begin{aligned} InfoGain(p, D) = & Entropy_{cur}(D) - \\ & \gamma * Entropy(\{Prob_{posterior}(d) * Pr(p|d) | d \in D\}) - \\ & (1 - \gamma) * Entropy(\{Prob_{posterior}(d) * Pr(\bar{p}|d) | d \in D\}) \quad \#(8) \end{aligned}$$

其中, 第一项为当前对话状态下的疾病后验概率信息熵, 第二项为表型 p 为阳性时的条件熵, 第三项为表型 p 为阴性时的条件熵, D 表示所有罕

见病的集合, $Prob_{posterior}(d)$ 表示疾病 d 的后验概率, γ 为阳性表型分支权重, 即 $\{Prob_{posterior}(d) * Pr(p|d) | d \in D\}$ 集合元素之和, $Pr(p|d)$ 表示表型 p 在疾病 d 中的出现频率, $Pr(\bar{p}|d)$ 表示表型 p 不在疾病 d 中出现的频率。

(2) 表型问答的基尼指数计算。

本发明使用基尼指数 (Gini Index) 来表示疾病推荐结果的纯度。基尼指数关联于数据杂乱度或纯度, 基尼指数侧重于代数表示, 更擅长较纯集合分类, 相关的CART算法用于构造分类决策树。我们将疾病后验概率作为分类的权重计算当前状态的基尼值, 如公式 (10) 所示。如果疾病后验概率分布集中在少数疾病上, 说明疾病诊断结果的纯度较高, 基尼值较小。首先定义基本的有分类权重的基尼指数 $Gini(W)$:

$$Gini(W) = 1 - \sum_{w_i \in W} \left(\frac{w_i}{W_{sum}} \right)^2 \quad \#(9)$$

其中, W 是一个权重集合, w_i 表示权重集合 W 中单一元素的权值, W_{sum} 表示 W 中所有元素权值之和。

当前对话状态下的疾病后验概率基尼指数为:

$$Gini_{cur}(D) = Gini(\{Prob_{posterior}(d) | d \in D\}) \quad \#(10)$$

其中, D 表示所有罕见病的集合, $Prob_{posterior}(d)$ 表示疾病 d 的后验概率。

单一表型条件下的基尼指数增益为:

$$\begin{aligned} GiniGain(p, D) = & Gini_{cur}(D) - \\ & \gamma * Gini(\{Prob_{posterior}(d) * Pr(p|d) | d \in D\}) - \\ & (1 - \gamma) * Gini(\{Prob_{posterior}(d) * Pr(\bar{p}|d) | d \in D\}) \quad \#(11) \end{aligned}$$

其中, 第一项为当前对话状态下的疾病后验概率基尼指数, 第二项为表型 p 为阳性时的条件基尼指数, 第三项为表型 p 为阴性时的条件基尼指数, 其余参数定义等同公式 (8)。

(3) 信息增益和基尼指数的自适应融合。

信息熵和基尼指数具有相似的不确定性度量功能, 但具有不同的定义和测量重点。信息熵侧重于信息表示, 对较乱数据具有更强区分性; 基尼指数则侧重于代数表示, 更擅长较纯集合分类。融合度量可以将二者的优势进行整合。在问答诊断前期或优势疾病后验概率不够突出时, 系统更侧重询问期望信息增益较大的表型来降低疾病诊断的不确定性; 在问答诊断

后期或优势疾病后验概率较为突出时，系统更侧重询问期望基尼指数较小的表型来进一步提升疾病诊断的纯度。本发明将信息增益和基尼指数进行自适应加权集成，产生具有更完备不确定性刻画的融合度量AIGGI (Adaptive Information Gain and Gini Index):

$$AIGGI(p, D) = \alpha * InfoGain(p, D) + (1 - \alpha) * GiniGain(p, D) \quad (12)$$

其中， D 表示所有罕见病的集合， p 表示单一表型， α 为反应当前对话状态下疾病后验概率分布的不确定程度或杂乱度，由于基尼指数的取值范围为 $[0,1]$ ，因此 α 可以直接使用 $Gini_{cur}(D)$ 来表示。

值得一提的是，为了提高对话的效率，我们在表型询问推荐时会应用注释转播规则来过滤掉一些表型。首先，已汇报的表型不会在接下来的对话轮次中被询问。除此之外，对于已汇报的阳性表型，默认其祖先表型均为阳性，将祖先表型过滤掉；对于已汇报的阴性表型，默认其后代表型均为阴性，将其后代表型过滤掉，系统不再询问。

(四)罕见病对话诊断系统对话诊断流程介绍

罕见病对话诊断系统对话诊断流程如图3所示，在初始的患者自我报告后，以多轮问答的形式进行人机交互来获得更多的患者表型信息，当达到最大对话轮次或疾病推荐结果不确定性很低的时候，系统给出疾病诊断推荐并结束对话。对话诊断流程详细步骤为：

(1) 初始自我报告。用户在前端UI模块的患者信息采集子模块中录入患者信息，包括患者年龄、出生时性别、阳性表型以及阴性表型。

(2) 表型问答。系统根据当前对话状态选择融合度量AIGGI最大的表型进行询问，用户针对该表型选择“存在”、“不存在”或者“不确定”的回答后，进入下一轮对话。问答过程中，用户可以在前端UI模块的对话诊断子模块中获得包括疾病优先级评估、当前对话状态、患者表型状态以及对话历史等多种可视化信息。

(3) 疾病推荐。重复步骤(2)直到疾病推荐结果的不确定性很低或者达到最大对话轮次时，系统根据疾病后验概率给出疾病推荐并结束对话。

以上所述的实施例对本发明的技术方案和有益效果进行了详细说明，应理解的是以上所述仅为本发明的具体实施例，并不用于限制本发明，凡在本发明的原则范围内所做的任何修改、补充和等同替换等，均应包含在本发明的保护范围之内。

权 利 要 求 书

1、一种罕见病可视化问答式辅助鉴别诊断系统，其特征在于，包括：

前端 UI 模块，用于进行人机交互，采集患者基本信息和表型信息录入以获取患者的初始自我报告，并利用多个维度可视化视图展现相关信息供临床人员进行鉴别诊断；

对话诊断服务模块，通过 API 提供表型询问推荐服务、疾病诊断推荐服务以及知识库查询服务；表型询问推荐服务是根据当前的对话状态推荐合适的表型以达到较高的对话收益，用户对系统询问的表型可以选择回答“存在”、“不存在”或者“不确定”；疾病诊断推荐服务是根据前端 UI 模块采集的患者初始自我报告和对话过程中产生的新的表型信息推荐合适的疾病；知识库查询服务是根据表型询问推荐服务和疾病诊断推荐服务的需要，查询并调用知识存储模块中存储的知识；

知识存储模块，用于存储人类表型本体知识库以及疾病-表型关系知识库。

2、根据权利要求 1 所述的罕见病可视化问答式辅助鉴别诊断系统，其特征在于，前端 UI 模块包括：

患者信息采集子模块，包括患者基本信息视图和表型信息录入视图；

可视化的问答式鉴别诊断子模块，疾病评估视图、对话状态视图、患者表型状态视图、对话历史视图以及问答视图，利用多个维度可视化视图展现相关信息供临床人员进行鉴别诊断。

3、根据权利要求 2 所述的罕见病可视化问答式辅助鉴别诊断系统，其特征在于，患者信息采集子模块包括患者基本信息视图和表型信息录入视图，用户可以在该模块输入信息包括患者年龄、出生时性别、阳性表型以及阴性表型。

问答式鉴别诊断子模块包括：

疾病评估视图，用于展示疾病患病率、后验概率以及表型匹配情况；

对话状态视图，以雷达图的形式展示当前优势最大的疾病的后验概率、表型匹配情况以及疾病推荐结果的不确定性度量；

患者表型状态视图，为基于人类表型本体并利用注释传播规则绘制的

反应患者表型特征的旭日图；

对话历史视图以及问答视图，用于展示系统询问的表型信息并为用户提供反馈按钮。

4、根据权利要求 1 所述的罕见病可视化问答式辅助鉴别诊断系统，其特征在于，疾病推荐服务包括疾病先验概率估计和疾病后验概率计算，对话诊断服务模块在对话结束时推荐后验概率最大的疾病；

疾病先验概率估计的方法包括：

(a-1) 将 Orphanet 知识库中的疾病流行性类型统一转换为点患病率；

(a-2) 在罕见病范围内，根据疾病的点患病率计算疾病先验概率，公式为：

$$Prob_{pretest}(D) = \frac{PointPrev_D}{\sum PointPrev} \#(1)$$

其中， $Prob_{pretest}(D)$ 代表疾病 D 的先验概率， $PointPrev_D$ 代表疾病 D 的点患病率， $\sum PointPrev$ 表示所有罕见病点患病率之和；

疾病后验概率计算的方法包括：

(b-1) 胜算与概率之间的换算关系为：

$$Odds = \frac{Prob}{1 - Prob}, Prob = \frac{Odds}{1 + Odds} \#(2)$$

根据公式 (2) 将疾病先验概率转换为疾病先验胜算；

(b-2) 根据疾病表型频率注释和注释传播规则依照公式 (3) 计算表型信息的复合似然比：

$$CompositeLR(P, D) = \frac{Pr(p_1, p_2, \dots, p_n | D)}{Pr(p_1, p_2, \dots, p_n | \bar{D})} = \prod_{i=1}^n \frac{Pr(p_i | D)}{Pr(p_i | \bar{D})} \#(3)$$

其中， P 代表已采集的表型集合， $(p_1, p_2, \dots, p_n) \in P$ ， $\bar{D}$ 代表除罕见病 D 以外的所有罕见病， $Pr(p_i | D)$ 表示表型 p_i 在疾病 D 中的出现频率，表型集合 P 的复合似然比被计算为单个表型 p_i 似然比的乘积；

(b-3) 根据公式 (4) 计算疾病的后验胜算 $Odds_{posterior}(P, D)$ ：

$$Odds_{posterior}(P, D) = Odds_{pretest}(D) * CompositeLR(P, D) \#(4)$$

其中， $Odds_{pretest}(D)$ 代表罕见病 D 的先验胜算；

(b-4) 根据公式 (5) 计算疾病后验概率 $Prob_{posterior}(D)$ ：

$$Prob_{posterior}(D) = \frac{Odds_{posterior}(D)}{1 + Odds_{posterior}(D)} \#(5)$$

。

5、根据权利要求 4 所述的罕见病可视化问答式辅助鉴别诊断系统，其特征在于，步骤 (a-1) 中，将疾病流行性类型统一转换为点患病率的方法为：

①如果疾病的流行性类型为点患病率，优先选择患病率地理区域为“世界范围内”的条目，使用该条目的数值型点患病率或者类别型点患病率的中值；如果患病率地理区域没有“世界范围内”的条目但有以国家或地区统计的患病率条目，则使用以国家或地区统计的患病率条目的点患病率平均值作为该疾病的点患病率；

②如果疾病的流行性类型为年发病率，则从Orphanet库获取该疾病的平均发病年龄以及死亡年龄信息，将发病年龄到死亡年龄作为疾病的持续时长，用年发病率乘以疾病持续年份估算该疾病的点患病率；

③如果疾病的流行性类型为病例数，则直接使用病例数除以世界总人口作为该疾病的点患病率；

④如果疾病的流行性类型为出生时患病率，且疾病发病年龄在新生儿期，则使用出生时患病率乘以该疾病持续年龄段的人口在世界总人口中的比例值作为疾病的点患病率；

⑤如果疾病的流行性类型为终生患病率，则使用默认的百万分之一作为该疾病的点患病率。

6、根据权利要求 1 所述的罕见病可视化问答式辅助鉴别诊断系统，其特征在于，表型推荐服务包括：

表型的信息增益计算，包括：

(A-1) 使用信息熵来表示疾病推荐结果的不确定性，以公式 (6) 定义有分类权重的信息熵 $Entropy(W)$ ：

$$Entropy(W) = - \sum_{w_i \in W} \frac{w_i}{W_{sum}} * \log \frac{w_i}{W_{sum}} \quad \#(6)$$

其中， W 是一个权重集合， w_i 表示权重集合 W 中单一元素的权值， W_{sum} 表示 W 中所有元素权值之和；

当前对话状态下的疾病后验概率的信息熵 $Entropy_{cur}(D)$ 为：

$$Entropy_{cur}(D) = Entropy(\{Prob_{posterior}(d) | d \in D\}) \quad \#(7)$$

其中， D 表示所有罕见病的集合， $Prob_{posterior}(d)$ 表示疾病 d 的后验

概率；

(A-2) 单一表型条件下的信息增益为：

$$\begin{aligned} \text{InfoGain}(p, D) = & \text{Entropy}_{\text{cur}}(D) - \\ & \gamma * \text{Entropy}(\{ \text{Prob}_{\text{posterior}}(d) * \text{Pr}(p|d) | d \in D \}) - \\ & (1 - \gamma) * \text{Entropy}(\{ \text{Prob}_{\text{posterior}}(d) * \text{Pr}(\bar{p}|d) | d \in D \}) \end{aligned} \quad \#(8)$$

其中，第一项为当前对话状态下的疾病后验概率的信息熵，第二项为表型 p 为阳性时的条件熵，第三项为表型 p 为阴性时的条件熵， D 表示所有罕见病的集合， $\text{Prob}_{\text{posterior}}(d)$ 表示疾病 d 的后验概率， γ 为阳性表型分支权重，即 $\{ \text{Prob}_{\text{posterior}}(d) * \text{Pr}(p|d) | d \in D \}$ 集合元素之和， $\text{Pr}(p|d)$ 表示表型 p 在疾病 d 中的出现频率， $\text{Pr}(\bar{p}|d)$ 表示表型 p 不在疾病 d 中出现的频率；表型的基尼指数计算，包括：

(B-1) 使用基尼指数来表示疾病推荐结果的纯度，以公式(9)定义有分类权重的基尼指数 $\text{Gini}(W)$ ：

$$\text{Gini}(W) = 1 - \sum_{w_i \in W} \left(\frac{w_i}{W_{\text{sum}}} \right)^2 \quad \#(9)$$

其中， W 是一个权重集合， w_i 表示权重集合 W 中单一元素的权值， W_{sum} 表示 W 中所有元素权值之和；

当前对话状态下的疾病后验概率的基尼指数为：

$$\text{Gini}_{\text{cur}}(D) = \text{Gini}(\{ \text{Prob}_{\text{posterior}}(d) | d \in D \}) \quad \#(10)$$

其中， D 表示所有罕见病的集合， $\text{Prob}_{\text{posterior}}(d)$ 表示疾病 d 的后验概率；

(B-2) 单一表型条件下的基尼指数增益为：

$$\begin{aligned} \text{GiniGain}(p, D) = & \text{Gini}_{\text{cur}}(D) - \\ & \gamma * \text{Gini}(\{ \text{Prob}_{\text{posterior}}(d) * \text{Pr}(p|d) | d \in D \}) - \\ & (1 - \gamma) * \text{Gini}(\{ \text{Prob}_{\text{posterior}}(d) * \text{Pr}(\bar{p}|d) | d \in D \}) \end{aligned} \quad \#(11)$$

其中，第一项为当前对话状态下的疾病后验概率基尼指数，第二项为表型 p 为阳性时的条件基尼指数，第三项为表型 p 为阴性时的条件基尼指数，其余参数定义等同公式(8)；

将信息增益和基尼指数进行自适应融合，获得融合度量AIGGI：

$$\text{AIGGI}(p, D) = \alpha * \text{InfoGain}(p, D) + (1 - \alpha) * \text{GiniGain}(p, D) \quad \#(12)$$

其中， D 表示所有罕见病的集合， p 表示单一表型， α 为反应当前对话

状态下疾病后验概率分布的不确定程度或杂乱度， α 使用 $Gini_{cur}(D)$ 来表示；

系统根据当前对话状态选择融合度量 AIGGI 最大的表型进行询问，用户针对该表型选择“存在”、“不存在”或者“不确定”的回答。

7、根据权利要求 6 所述的罕见病可视化问答式辅助鉴别诊断系统，其特征在于，在表型询问推荐时应用注释转译规则来过滤掉一些表型，包括：

已汇报的表型不会在接下来的对话轮次中被询问；

对于已汇报的阳性表型，默认其祖先表型均为阳性，将其祖先表型过滤掉，不再询问；

对于已汇报的阴性表型，默认其后代表型均阴性，将其后代表型过滤掉，不再询问。

8、一种采用如权利要求 1-7 任一项所述的系统进行罕见病辅助鉴别诊断的方法，包括：

(1) 用户在前端 UI 模块的患者信息采集子模块中录入患者信息，包括患者年龄、出生时性别、阳性表型以及阴性表型；

(2) 对话诊断服务模块根据当前对话状态选择融合度量 AIGGI 最大的表型进行询问，用户针对该表型选择“存在”、“不存在”或者“不确定”的回答后，之后进入下一轮对话；

问答过程中，用户可以在前端 UI 模块的可视化的问答式鉴别诊断子模块中获得多种可视化信息，包括疾病优先级评估、当前对话状态、患者表型状态以及对话历史；

(3) 重复步骤 (2)，直到疾病推荐结果的不确定性低于阈值或者达到最大对话轮次时，系统根据疾病后验概率给出疾病推荐并结束对话。

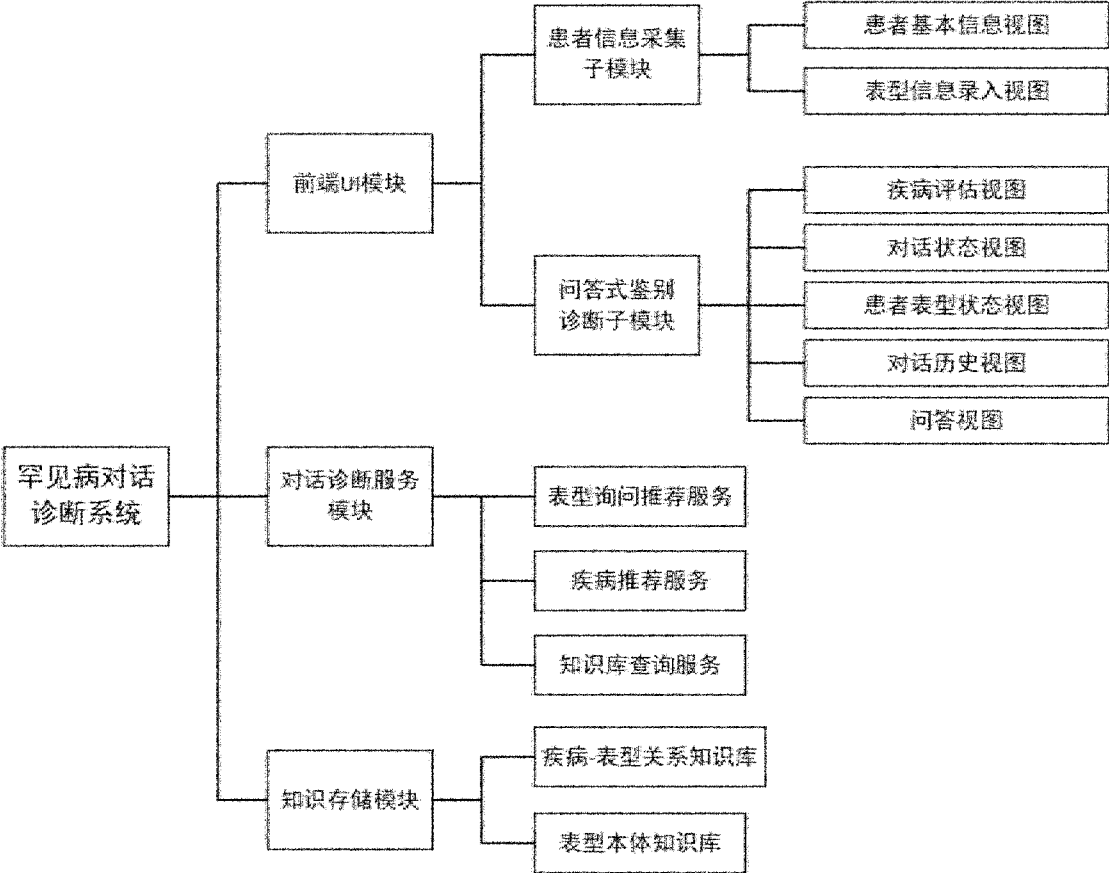


图 1

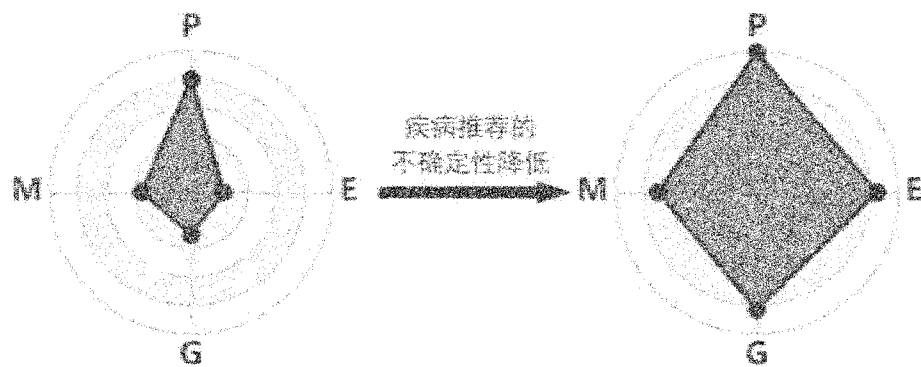


图 2

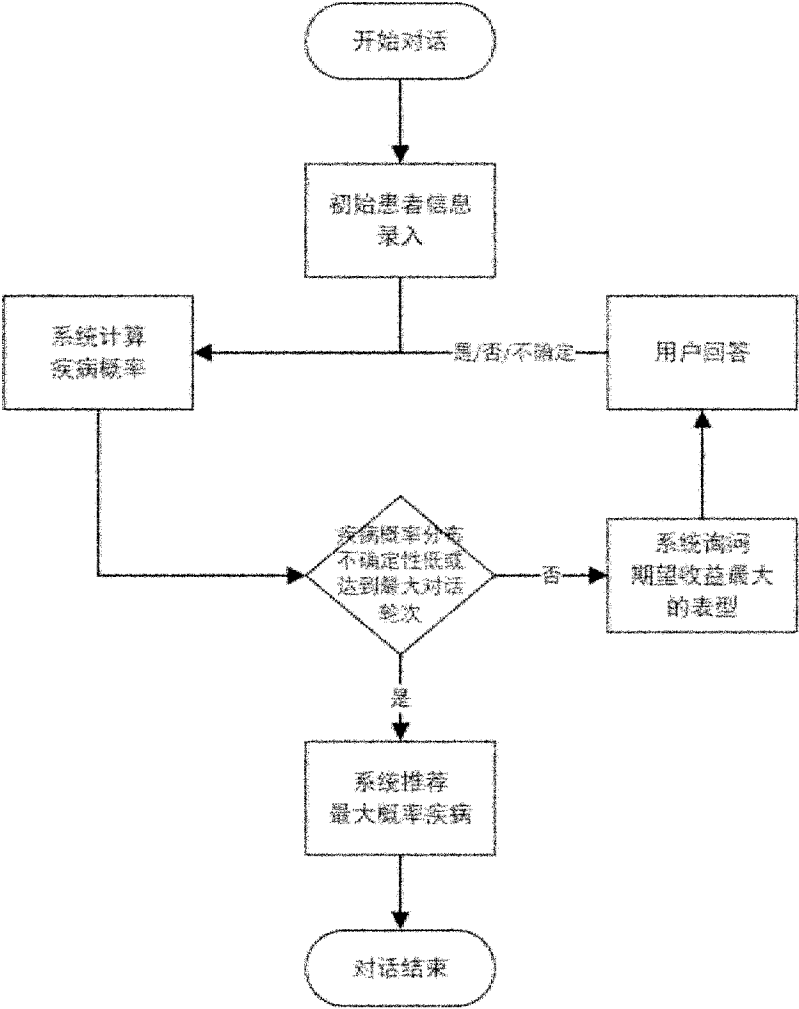


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/077386

A. CLASSIFICATION OF SUBJECT MATTER

G16H50/20(2018.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G16H

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 罕见病, 疾病, 诊断, 辅助, 采集, 表型, 问答, 对话, 推荐, 知识库, rare, disease, diagnosis, assist, AIDS, phenotype, answer, recommendation, knowledge, database

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 115482926 A (ZHEJIANG UNIVERSITY) 16 December 2022 (2022-12-16) claims 1-8	1-8
X	CN 113889259 A (ZHEJIANG UNIVERSITY OF TECHNOLOGY) 04 January 2022 (2022-01-04) claim 1, and description, paragraphs [0058]-[0103]	1-3, 8
A	CN 113889265 A (ZHEJIANG UNIVERSITY) 04 January 2022 (2022-01-04) entire document	1-8
A	CN 112270988 A (XIAMEN GENOKON MEDICAL TECHNOLOGY CO., LTD.) 26 January 2021 (2021-01-26) entire document	1-8
A	CN 107577907 A (GENOMCAN INC.) 12 January 2018 (2018-01-12) entire document	1-8
A	US 2020286622 A1 (GACHON UNIVERSITY OF INDUSTRY-ACADEMIC COOPERATION FOUNDATION et al.) 10 September 2020 (2020-09-10) entire document	1-8



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

17 March 2023

Date of mailing of the international search report

22 March 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/077386

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	115482926	A	16 December 2022	None			
CN	113889259	A	04 January 2022	None			
CN	113889265	A	04 January 2022	None			
CN	112270988	A	26 January 2021	None			
CN	107577907	A	12 January 2018	None			
US	2020286622	A1	10 September 2020	WO	2020111378	A1	04 June 2020
				KR	20200064453	A	08 June 2020
				KR	102147847	B1	25 August 2020

国际检索报告

国际申请号

PCT/CN2023/077386

A. 主题的分类

G16H50/20 (2018.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

IPC: G16H

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 罕见病, 疾病, 诊断, 辅助, 采集, 表型, 问答, 对话, 推荐, 知识库, rare, disease, diagnosis, assist, AIDS, phenotype, answer, recommendation, knowledge, database

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 115482926 A (浙江大学) 2022年12月16日 (2022 - 12 - 16) 权利要求1-8	1-8
X	CN 113889259 A (浙江工业大学) 2022年1月4日 (2022 - 01 - 04) 权利要求1, 说明书第[0058]-[0103]段	1-3, 8
A	CN 113889265 A (浙江大学) 2022年1月4日 (2022 - 01 - 04) 全文	1-8
A	CN 112270988 A (厦门基源医疗科技有限公司) 2021年1月26日 (2021 - 01 - 26) 全文	1-8
A	CN 107577907 A (成都奇恩生物科技有限公司) 2018年1月12日 (2018 - 01 - 12) 全文	1-8
A	US 2020286622 A1 (GACHON UNIVERSITY OF INDUSTRY-ACADEMIC COOPERATION FOUNDATION 等) 2020年9月10日 (2020 - 09 - 10) 全文	1-8

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2023年3月17日

国际检索报告邮寄日期

2023年3月22日

ISA/CN的名称和邮寄地址

中国国家知识产权局

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

王莹

电话号码 (+86) 010-53961411

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2023/077386

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	115482926	A	2022年12月16日	无			
CN	113889259	A	2022年1月4日	无			
CN	113889265	A	2022年1月4日	无			
CN	112270988	A	2021年1月26日	无			
CN	107577907	A	2018年1月12日	无			
US	2020286622	A1	2020年9月10日	WO	2020111378	A1	2020年6月4日
				KR	20200064453	A	2020年6月8日
				KR	102147847	B1	2020年8月25日