(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
2022 年 11 月 3 日 (03.11.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/227294 A1

(51) 国际专利分类号:
G16H 10/60 (2018.01) *G16H 50/20* (2018.01)

(21) 国际申请号: PCT/CN2021/106860

(22) 国际申请日: 2021 年 7 月 16 日 (16.07.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110486200.2 2021年4月30日 (30.04.2021) CN

(71) 申请人: 山 东 大 学 (**SHANDONG UNIVERSITY**) [CN/CN]; 中国山东省济南市历下区经十路17923号, Shandong 250061 (CN)。

(72) 发明人: 刘治 (**LIU, Zhi**); 中国山东省济南市历下区经十路17923号, Shandong 250061 (CN)。 李玉军 (**LI, Yujun**); 中国山东省济南市历下区经十路17923号, Shandong 250061 (CN)。 胡喜风 (**HU, Xifeng**); 中国山东省济南市历下区经十路17923号, Shandong 250061 (CN)。 胡伟风 (**HU, Weifeng**); 中国山东省济南市历下区经十路17923号, Shandong 250061 (CN)。

(74) 代理人: 济南圣达知识产权代理有限公司 (**JINAN SHENGDA INTELLECTUAL PROPERTY AGENCY CO., LTD.**); 中国山东省济南市历下区经十路17703号华特广场B510室, Shandong 250061 (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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,

(54) **Title:** DISEASE RISK PREDICTION METHOD AND SYSTEM BASED ON MULTI-MODAL FUSION

(54) 发明名称: 一种基于多模态融合的疾病风险预测方法和系统

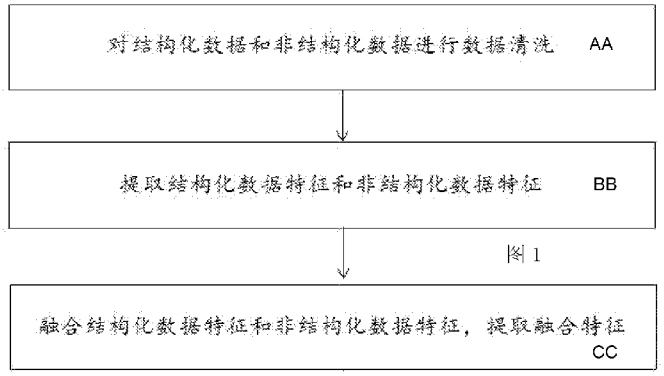


图 1

AA Perform data cleaning on structured data and unstructured data
BB Extract a structured data feature and an unstructured data feature
CC Fuse the structured data feature and the unstructured data feature, and extract a fused feature

(57) **Abstract:** A disease risk prediction method based on multi-modal fusion. The method comprises: acquiring EHR data of a patient, wherein the EHR data comprises structured data and unstructured data; inputting the EHR data into a disease risk prediction model, so as to obtain a disease risk prediction result; and outputting the disease risk prediction result, wherein execution steps of the disease risk prediction model comprise: extracting a structured data feature and an unstructured data feature; fusing the structured data feature and the unstructured data feature, and extracting a fused feature; and making a decision on the fused feature to obtain the disease risk prediction result. In addition, further disclosed are a disease risk prediction system based on multi-modal fusion and used for implementing the method, and a method for processing EHR data, a method for constructing a disease risk prediction model, and a computer device and a computer-readable storage medium, which can implement a method.

WO 2022/227294 A1

[见续页]

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。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, 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：一种基于多模态融合的疾病风险预测方法，获取患者的EHR数据，包括结构化数据和非结构化数据，将EHR数据输入疾病风险预测模型，得到疾病风险预测结果，输出疾病风险预测结果；其中，疾病风险预测模型执行步骤包括：提取结构化数据特征和非结构化数据特征，融合结构化数据特征和非结构化数据特征，提取融合特征，对融合特征进行决策，得到疾病风险预测结果。此外，还公开了实现方法的基于多模态融合的疾病风险预测系统、处理EHR数据的方法、疾病风险预测模型的构建方法，以及可实现方法的计算机设备、计算机可读存储介质。

一种基于多模态融合的疾病风险预测方法和系统

技术领域

本申请涉及医学大数据信息处理领域，尤其涉及一种基于多模态融合的疾病风险预测方法和系统。

背景技术

公开该背景技术部分的信息仅仅旨在增加对本申请的总体背景的理解，而不必然被视为承认或以任何形式暗示该信息构成已经成为本领域一般技术人员所公知的现有技术。

电子健康记录（EHR）为健康研究创建了大量廉价的数据，其涵盖电子病历、既往病史信息、患者病历的文本记录等数据。数字化和对病历的后续分析构成了一种数字化转换领域，旨在以 EHR 的形式收集有关患者的多种医学信息，包括数字化测量（实验室结果）、口头描述（症状和便笺、生命体征等）、图像（X 射线、CT 和 MR 扫描等）并记录患者的治疗过程。这种数字化为挖掘健康记录创造了机会，以提高护理质量和临床结果。

然而临床医生只有有限的时间来处理所有可用数据并检测类似病历中的模式。电子健康记录包含具有重要研究和临床价值的结构化和非结构化数据，随着大量 EHR 数据的标准化和数字化，通过对大量多源异构数据进行挖掘进而建立风险预测模型来实现个性化医疗是亟需的。以往的大多数尝试都是建立在结构化的 EHR 字段上，非结构化文本数据中的大量信息被丢失。

发明内容

本发明发明人在了解现有技术存在的缺陷的基础上，发现通过对医学文本进行有效挖掘，并通过有效的数据融合手段将多源异构数据进行深层次的融合研究，能够避免单一数

据所导致的局限性和片面性。因此，发明人在进一步地将深度学习与疾病预测相结合进行研究。然而，两者的结合伴随着以下问题，包括：

数据集数量和分布不均衡的问题：不带目的的数据收集往往会造成记录数据的完整度、准确度和颗粒度无法形成系统的体系，造成数据的缺失和不规范。因此，需要耗费一定的人力和物力进行数据采集。限于时间和财力，能够得到的良好的样本数量有限，比如，在本发明的一些实施方式中，得到良好的样本数量仅有 1300 例，且正负样本分布不均衡，这会极大地影响深度神经网络的学习和训练。

医学文本数据不能直接用于计算的问题：在现有的处理方式中，医学文本往往首先需要进行数字化表示。但是，这些文本数据通常是长文本且带有医学实体，采用 CNN(Convolutional Neural Network，卷积神经网络)、word2vec(词向量产生模型)、LSTM(Long-Short Term Memory，长短期记忆网络)、Bi-LSTM(Bi-directional Long-Short Term Memory)等进行医学文本数据的向量表示时不尽人意。

以及，目前临床的真实数据大多是以多模态的形式存在，然而现在对于多模态方面的研究较少，单点突破已经做了很多事情，只考虑单模态因素不能对潜在风险进行综合评估，临床数据也未被充分挖掘利用。

为了解决现有研究中的不足以及上述问题，本发明通过堆叠的 Transformer 编码器模块对文本病历进行有效的向量化表示，其能有效捕获长文本前后语序包含的丰富的语义关系，并对医疗实体进行正确表示。接着将多源异构数据进行特征级融合，充分考虑到不同模态数据的特点，进而对患者结局进行预测。本发明提供了一种处理 EHR 数据（包含结构化数据和非结构化数据）的方法、并构建了一种基于多模态融合的疾病风险预测模型、使用该模型进行预测的方法和系统，及实现这些功能的软件设备等。本发明通过对患者的人口统计学信息、治疗信息、诊断信息化验信息和相关文本治疗病历进行信息的充分的融合

挖掘提高对患者结局的预判性，能有效的帮助医生提供有效的参考信息，预判患者病情的发展情况，辅助医生制定相应的治疗方案，及时帮助救治，防止病情往恶化的方向发展。同时可以在临床每次访问时向患者展示个性化治疗后的疾病发展方向以提高其治疗的积极性。

多模态数据是指在多种不同设备或场景下采集到的数据。现实世界中的数据集往往是多模态的，例如：一个的故事可以由文本叙述也能用图像或者音频来描述；一个文档可以由多种不同的语言表示也能用用户评价来表示等等。多模态数据库的建立旨在通过分析和处理多模态数据得到其重要特征和代表性检索标签，并以此为基础建立便于后续数据检索的数据库。

非结构化数据是指没有固定结构的数据，例如，所有格式的办公文档、文本、图片、各类报表、图像和音频、视频信息。医学中的非结构化数据包含有医疗影像、心电图、文本病历等。

结构化数据：传统的关系数据模型、行数据，存储于数据库，可用二维表结构表示的数据，例如，存储于 csv，excel 的数据、二维表。

具体地，本发明提供了下述的技术特征，以下技术特征的一个或多个的结合构成本发明的技术方案。

在本发明的第一方面，本发明提供了一种基于多模态融合的疾病风险预测方法，所述方法包括：

获取待预测患者的 EHR 数据，所述数据包括结构化数据和非结构化数据；在本发明的实施方式中，所述非结构化数据尤其指文本；

将 EHR 数据输入疾病风险预测模型，得到疾病风险预测结果；

输出疾病风险预测结果。

其中，疾病风险预测模型执行步骤：

提取结构化数据特征和非结构化数据特征；

融合结构化数据特征和非结构化数据特征，提取融合特征；

对融合特征进行决策，得到疾病风险预测结果。

在本发明的一些实施方式中，所述疾病风险预测模型在提取结构化数据特征和非结构化数据特征前还包括执行数据清洗的步骤；

其中，所述数据清洗包括对异常值的替换、采用均值对缺失值进行补全，以及删除脏数据。

在本发明的一些实施方式中，采用全卷积网络(Fully Convolutional Networks，FCN)提取结构化数据特征。

在本发明的一些实施方式中，采用 BERT（Bidirectional Encoder Representations from Transformers）提取非结构化特征。

在本发明的一些实施方式中，所述提取融合特征的操作包括：将非结构化数据特征和结构化数据特征沿指定维度进行并联，采用合成少数类过采样技术（Synthetic Minority Oversampling Technique，SMOTE）通过对少数类样本数据进行分析并新生成该类样本的方法来降低不平衡率，然后采用分段池化操作，提取得到融合特征。

在本发明的一些实施方式中，进行预测时，将融合特征作为 input 输入到全连接层（Fully connected dence layers），然后通过 Softmax 分类器进行疾病风险预测。

以及，在本发明的实施方式中，本发明采用交叉熵损失和合页损失的加权来共同约束模型。交叉熵损失能够衡量同一随机变量中的两个不同概率分布的差异程度，交叉熵损失值越小，两个概率分布越接近。然而单独使用交叉熵损失易导致边界变量分类混淆，合页损失专用于二分类问题，它不仅要求分类正确，而且确信度足够高时损失才会尽可能的小。

由于合页损失不仅度量了模型对训练数据的拟合程度，而且通过加入正则化项度量了模型自身的复杂度，因此可以大大降低拟合风险。

在本发明的第二方面，本发明提供了一种处理 EHR 数据的方法，其包括：

获取 EHR 数据，所述数据包括结构化数据和非结构化数据；

对结构化数据和非结构化数据分别进行数据处理，包括对结构化数据和非结构化数据分别进行数据清洗，对清洗后的结构化数据和非结构化数据分别进行特征提取，将分别提取得到的非结构化数据特征和结构化数据特征进行融合处理后提取融合特征；

以提取的融合特征数据作为待识别数据用于医疗用途。

在本发明的一些实施方式中，所述数据清洗包括对异常值的替换、采用均值对缺失值进行补全，以及删除脏数据；优选地，所述非结构化数据为文本。

在本发明的一些实施方式中，提取结构化数据特征采用 FCN；提取非结构化特征采用 BERT。

在本发明的一些实施方式中，所述提取融合特征的操作包括：将非结构化数据特征和结构化数据特征沿指定维度进行并联，采用 SMOTE 通过对少数类样本数据进行分析并新生成该类样本的方法来降低不平衡率，然后采用分段池化操作，提取得到融合特征。

在本发明的第三方面，本发明一种疾病风险预测模型的构建方法，其包括：

获取已知疾病风险结局的患者的 EHR 数据，所述数据包括结构化数据和非结构化数据；以获取的 EHR 数据构建数据集，包括结构化数据集和非结构化数据集，以已知的最终结局构建标签集；

构建疾病风险预测网络，包括：构建对于结构化数据进行提取的特征提取模块、对于非结构化数据进行提取的特征提取模块、和特征融合模块，结构化数据特征提取模块和非结构化数据特征提取模块并联连接后在特征融合模块决策层进行串联连接；所述疾病风险

预测网络基于 Pytorch 框架实现；

以标签集为标签，利用数据集（结构化数据集和非结构化数据集）训练构建的疾病风险预测网络，构建得到疾病风险预测模型；

以及，采用交叉熵损失和合页损失的加权来共同约束模型。

在本发明的一些实施方式中，构建数据集前还包括对获取的 EHR 数据进行数据清洗的步骤，数据清洗包括对异常值的替换、采用均值对缺失值进行补全，以及删除脏数据。

在本发明的一些实施方式中，对于结构化数据进行提取的特征提取模块为 FCN 模块；对于非结构化数据进行提取的特征提取模块为 BERT 模块（transformer 模块）。

在本发明的一些实施方式中，特征融合模块执行：将非结构化数据特征和结构化数据特征沿指定维度进行并联，采用 SMOTE 通过对少数类样本数据进行分析并新生成该类样本的方法来降低不平衡率，然后采用分段池化操作，提取得到融合特征；

在本发明的一些实施方式中，利用数据集训练时，以融合特征作为 input 输入到全连接层，训练 Softmax 分类器。

以及，本发明还包括由上述第三方面构建得到的基于多模态融合的疾病风险预测模型。

在本发明的第四方面，本发明提供了一种基于多模态融合的风险预测系统，所述系统包括：

特征提取模块，其用于对 EHR 数据进行特征提取，得到非结构化数据特征和结构化数据特征；

特征融合模块，其用于对非结构化数据特征和结构化数据特征进行融合处理并提取得到融合特征；

分类模块，其以提取的融合特征作为 input，得到疾病风险预测结果。

在本发明的一些实施方式中，所述特征提取模块包括结构化数据特征提取模块和非结构化数据特征提取模块；

其中，所述结构化数据特征提取模块以预处理后的结构化数据作为FCN的input，将数据映射到各个隐语义节点，得到结构化数据特征。

其中，所述非结构化数据特征提取模块采用BERT对预处理后的非结构化数据进行特征提取；优选地，BERT由BERT Encoder组成，BERT Encoder由多层BERT Layer组成，每一层的BERT Layer均为Transformer中的Encoder Block；每一个encoder层包含两层，分别为自注意力机制层和前馈神经网络层。

在本发明的一些实施方式中，所述特征融合模块将非结构化数据特征和结构化数据特征沿指定维度进行并联，采用SMOTE通过对少数类样本数据进行分析并新生成该类样本的方法来降低不平衡率，然后采用分段池化操作，提取得到融合特征。

在本发明的一些实施方式中，分类模块将融合特征作为input输入到全连接层，然后通过Softmax分类器分类，得到疾病风险预测结果。

在本发明的一些实施方式中，所述系统还包括数据获取模块，其用于获取EHR数据。

在本发明的一些实施方式中，所述系统还包括数据清洗模块，其用于在获取EHR数据后、在对EHR数据进行特征提取前对EHR数据进行预处理，所述预处理包括对所述EHR数据清洗模块执行对异常值替换和采用均值对缺失值进行补全并删除脏数据的操作。

在本发明的一些实施方式中，所述系统还包括结果输出模块，其用于输出疾病风险预测结果。

在本发明的第五方面，本发明提供了一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现如本发明上述第一方面和/或第二方面和/或第三方面中任一项所述方法的步骤。

在本发明的第六方面，本发明提供了一种计算机可读存储介质，其上存储有计算机程序指令，该程序指令被处理器执行时实现如本发明上述第一方面和/或第二方面和/或第三方面中任一项所述方法的步骤。

通过上述一个或多个技术手段，可实现以下有益效果：

本发明提供了一种端到端的患者结局预测模型，通过自动读取患者的 EHR 数据，将读取到的数据作为模型的输入，结合深度学习方法对相应数据进行挖掘分析后，输出即为预测的患者的事件结局。其能有效的帮助医生提供有效的参考信息，预判患者病情的发展情况，及时帮助救治。同时增加患者配合治疗的积极性。

本发明采用了双向语言模型 BERT 对医学文本进行特征提取，可以对多组输入做并行计算，捕获不同的子空间信息。引入注意力机制帮助模型更有效的获取上下文信息，学习到句子内部的词依赖关系，捕获句子的内部结构。对于模型的预训练采用中文医疗问答、中文医疗百科和中文电子病历等数据，类似于“腹痛”等医疗实体可以得到更有效的向量化表示。

本发明采用多模态融合技术，对患者的电子病历、既往病史信息、患者病历的文本记录等数据进行预处理、分析和挖掘，构建预测患者结局的风险预测模型，为临床真实数据的利用、疾病结局评估提供辅助手段，帮助医生为每位患者提供个性化的治疗方案。

附图说明

构成本申请的一部分的说明书附图用来提供对本申请的进一步理解，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。以下，结合附图来详细说明本申请的实施方案，其中：

图 1 为本发明第一实施例中处理 EHR 数据的方法的流程图。

图 2 为本发明第一实施例中处理 EHR 数据的系统结构图。

图 3 为本发明的一个或多个实施例中特征融合模块的功能流程图。

图 4 为本发明第三实施例中基于多模态融合的疾病风险预测方法的流程图。

图 5 为本发明一个或多个实施例中疾病风险预测模型的功能流程图。

图 6 为本发明第四实施例中一种基于多模态融合的风险预测系统的结构图。

图 7 为本发明第四实施例中一种基于多模态融合的风险预测系统的结构图。

图 8 为本发明第四实施例中一种基于多模态融合的风险预测系统的结构图。

具体实施方式

下面结合具体实施例，进一步阐述本申请。应理解，这些实施例仅用于说明本申请而不用于限制本申请的范围。下列实施例中未注明具体条件的实验方法，通常按照常规条件或按照制造厂商所建议的条件。

文中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，单独存在 B，同时存在 A 和 B 三种情况，本文中术语“/和”是描述另一种关联对象关系，表示可以存在两种关系，例如，A/和 B，可以表示：单独存在 A，单独存在 A 和 B 两种情况，另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

本文使用的术语仅用于描述特定实施例，并且不意在限制本申请的示例实施例。如本文所使用的，单数形式“一”、“一个”以及“该”意在包括复数形式，除非上下文明确指示相反意思。还应当理解术语“包括”、“包括了”、”包含”、和/或“包含了”当在本文中使用时，指定所声明的特征、整数、步骤、操作、单元和/或组件的存在性，并且不排除一个或多个其他特征、数量、步骤、操作、单元、组件和/或他们的组合的存在或增加。

应当理解，尽管在本申请可能采用术语第一、第二、第三等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离

本申请范围的情况下，第一信息也可以被称为第二信息，类似地，第二信息也可以被称为第一信息。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

此外，特定特征、结构、功能或特性可以以任何适合的方式组合到一个或多个实施例中。例如，第一实施例可以结合第二实施例，只要与这些实施例或具体实施方式相关联的特定特征、结构、功能或特性不互相排斥。

在本发明的第一实施例中，本发明提供了一种处理 EHR 数据的方法，包括：获取 EHR 数据，所述数据包括结构化数据和非结构化数据；

对 EHR 数据进行处理，处理流程如图 1 中所示，包括：对结构化数据和非结构化数据分别进行数据处理，包括对结构化数据和非结构化数据分别进行数据清洗，对清洗后的结构化数据和非结构化数据分别进行特征提取，将分别提取得到的非结构化数据特征和结构化数据特征进行融合处理后提取融合特征；

以提取的融合特征数据作为待识别数据用于医疗用途。

以及，基于第一实施例中的方法，本发明还提供了一种处理 EHR 数据的系统，其核心模块包括：特征提取模块和特征融合模块；

可选地，获取待处理 EHR 数据后可对数据进行数据清洗，因此，所述系统中还可包括数据清洗模块，如图 2 所示。

其中，数据清洗模块执行对异常值替换和采用均值对缺失值进行补全并删除脏数据的操作。比如可采用首先对于数据进行异常值筛选，使用空值替换该异常值，接着对数据进行加权平均，使用平均值替换异常值和缺失值，可采用 spss 对数据进行清洗操作。

特征提取模块对 EHR 数据中包含的结构化数据和非结构化数据（比如文本）进行特征提取；可选地，所述特征提取模块包括结构化数据特征提取模块和非结构化数据特征提取

模块。

其中，结构化数据特征提取模块以清洗后的结构化数据作为FCN的input，将数据映射到各个隐语义节点，得到结构化数据特征；在该实施方式中，结构化数据特征提取模块通过Dence层来学习权重W，进而得到结构化数据的重置特征，由于数据的离散的，特征之间的位置信息对于决策影响很小，因此在这个过程中可选择舍弃位置信息。

非结构化数据特征提取模块采用BERT对清洗后的非结构化文本数据进行特征提取。所述BERT由BERT Encoder组成，BERT Encoder由多层BERT Layer组成，每一层的BERT Layer均为Transformer中的Encoder Block；每一个encoder层包含两层，分别为自注意力机制层和前馈神经网络层。在该实施方式中，对于非结构化文本数据挖掘的模块，采用的是堆叠的Transformer编码器模块，分别得到词嵌入张量、语句分块张量和位置编码张量来提取到医学文本数据语义信息、句子信息和位置信息，计算得到文本病历的向量化表示。

对于特征融合模块，如图3所示，连接层将结构化数据特征与非结构化数据特征沿指定维度进行并联，采用SMOTE通过对少数类样本数据进行分析并新生成该类样本的方法来降低不平衡率，并通过添加分段池化操作按照数据类型的不同分别提取到不同结构数据的重要信息。由于医学数据通常样本量较少，批归一化会受到batch_size大小的影响，因此在本发明的实施方式中，每个子层的输出采用层归一化。

在本发明的第二实施例中，本发明提供了一种构建疾病风险预测模型的方法，包括：

获取已知疾病风险结局的患者的EHR数据(所述数据包括结构化数据和非结构化数据，非结构化数据主要指文本)；以其EHR数据构建数据集（结构化数据集和文本数据集），以其最终结局构建标签集；

可选地，对获取的EHR数据进行数据清洗，数据清洗包括对异常值的替换、采用均值

对缺失值进行补全，以及删除脏数据；

构建疾病风险预测网络，包括：构建对于结构化数据进行提取的特征提取模块（FCN）、对于非结构化数据进行提取的特征提取模块（transformer 模块）、特征融合模块，结构化数据特征提取模块和非结构化数据特征提取模块并联连接后在特征融合模块决策层进行串联连接，所述模型架构基于 Pytorch 框架实现；

以标签集为标签，利用数据集训练上述构建的疾病风险预测网络，构建疾病风险预测模型；在该实施例中，以疾病风险结局为标签，以融合特征作为 input 输入到全连接层，训练 Softmax 分类器，构建疾病风险预测模型。

将待预测患者的 EHR 数据输入至训练完成的疾病风险预测模型中，即可输出患者的结局属性情况。

进一步地，采用交叉熵损失和合页损失的加权来共同约束疾病风险预测模型。交叉熵损失能够衡量同一随机变量中的两个不同概率分布的差异程度，交叉熵损失值越小，两个概率分布越接近。然而单独使用交叉熵损失易导致边界变量分类混淆，合页损失专用于二分类问题，它不仅要求分类正确，而且确信度足够高时损失才会尽可能的小。由于合页损失不仅度量了模型对训练数据的拟合程度，而且通过加入正则化项度量了模型自身的复杂度，因此可以大大降低拟合风险。

在本发明的第三实施例中，基于第二实施例构建的疾病风险预测模型，本发明提供了一种基于多模态融合的疾病风险预测方法，如图 4 所示，其包括：

获取待预测患者的 EHR 数据，EHR 数据可以包括结构化数据和非结构化数据（文本）；

将获取的 EHR 数据输入疾病风险预测模型，得到疾病风险预测结果；

输出疾病风险预测结果。

其中，所述疾病风险预测模型执行步骤，如图 5 所示，包括：

提取结构化数据特征和非结构化数据特征；

提取融合特征，所述融合特征为非结构化数据特征和结构化数据特征的融合特征；

对融合特征进行决策，得到疾病风险预测结果。

在该实施例中，采用交叉熵损失和合页损失的加权来共同约束模型。交叉熵损失能够衡量同一随机变量中的两个不同概率分布的差异程度，交叉熵损失值越小，两个概率分布越接近。然而单独使用交叉熵损失易导致边界变量分类混淆，合页损失专用于二分类问题，它不仅要求分类正确，而且确信度足够高时损失才会尽可能的小。由于合页损失不仅度量了模型对训练数据的拟合程度，而且通过加入正则化项度量了模型自身的复杂度，因此可以大大降低拟合风险。

在本发明的第四实施例中，本发明提供了一种基于多模态融合的风险预测系统，如图6所示，包括：特征提取模块、特征融合模块和分类模块。

其中，特征提取模块包括：结构化数据提取模块和非结构化数据提取模块，如图7所示。

在该实施例的基础上，所述基于多模态融合的风险预测系统中还可以包括数据获取模块和/或数据清洗模块和/或结果输出模块。

比如，在该实施例中，所述系统可如图8中所示。

如图8所示，所述系统获取待预测患者的EHR数据（包括结构化数据和非结构化数据，非结构化数据比如文本）后，数据清洗模块对EHR数据进行预处理包括对异常值的替换和采用均值对缺失值进行补全并删除脏数据。

清洗处理后的非结构化数据比如文本数据在文本特征提取模块内进行特征提取，该模块内应用双向语言模型BERT，对医疗文本数据进行特征提取。模型的核心由BERT Encoder组成，BERT Encoder由多层BERT Layer组成，每一层的BERT Layer其实都是Transformer

中的 Encoder Block。每一个 encoder 层包含两层，一个自注意力机制层一个前馈神经网络层。

清洗处理后的结构化数据在结构化数据特征提取模块内进行特征提取，其中，清洗处理后的结构化数据作为 FCN 的输入，将原始数据映射到各个隐语义节点，得到结构化数据特征。

如图 3 所示，融合模块将结构化数据的特征与文本数据的特征沿指定维度进行拼接并联，采用 SMOTE 通过对少数类样本数据进行分析，新生成该类样本的方法来降低不平衡率。接着采用分段池化操作，提取不同结构数据的重要信息，得到融合特征。

分类模块将融合后提取的融合特征作为 input 输入到全连接层，然后通过 Softmax 分类器进行患者的结局预测。

进一步，分类模块得到的预测解决可通过结果输出模块输出。

医疗患者可根据输出的结果结合自己的判断得出最终的结论。

该实施例中所述的系统可实现第三实施例中所述的基于多模态融合的疾病风险预测方法。

在本发明的第五实施例中，本发明提供了一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现第一实施例中所述方法的步骤；

和/或，所述处理器执行所述计算机程序时实现第二实施例中所述方法的步骤；

和/或，所述处理器执行所述计算机程序时实现第三实施例中所述方法的步骤。

在本发明的第六实施例中，本发明提供了一种计算机可读存储介质，其上存储有计算机程序指令，该程序指令被处理器执行时实现第一实施例中所述方法的步骤；

和/或，该程序指令被处理器执行时实现第二实施例中所述方法的步骤；

和/或，该程序指令被处理器执行时实现第三实施例中所述方法的步骤。

以上所描述的装置实施例仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助加必需的通用硬件平台的方式来实现，当然也可以通过硬件和软件结合的方式来实现。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以计算机产品的形式体现出来，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质，包括但不限于磁盘存储器、CD-ROM、光学存储器等上实施的计算机程序产品的形式。

以上所述仅为本申请的优选实施例而已，并不用于限制本申请，尽管参照前述实施例对本申请进行了详细的说明，对于本领域的技术人员来说，其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权利要求书

1、一种基于多模态融合的疾病风险预测方法，其特征在于，所述方法包括：

获取患者的 EHR 数据，包括结构化数据和非结构化数据；

将 EHR 数据输入疾病风险预测模型，得到疾病风险预测结果；

输出疾病风险预测结果；

其中，疾病风险预测模型执行步骤，包括：

提取结构化数据特征和非结构化数据特征；

融合结构化数据特征和非结构化数据特征，提取融合特征；

对融合特征进行决策，得到疾病风险预测结果。

2、根据权利要求 1 所述的方法，其特征在于，采用全卷积网络提取结构化数据特征；

优选地，采用 BERT 提取非结构化特征。

3、根据权利要求 1 或 2 所述的方法，其特征在于，所述提取融合特征的操作包括：将非结构化数据特征和结构化数据特征沿指定维度进行并联，采用 SMOTE 通过对少数类样本数据进行分析并新生成该类样本的方法来降低不平衡率，然后采用分段池化操作，提取得到融合特征；

优选地，进行预测时，将融合特征作为 input 输入到全连接层，然后通过 Softmax 分类器进行疾病风险预测；

优选地，采用交叉熵损失和合页损失的加权来共同约束疾病风险预测模型。

4、根据权利要求 1 所述的方法，其特征在于，所述疾病风险预测模型在提取结构化数据特征和非结构化数据特征前还包括执行数据清洗的步骤；

优选地，所述数据清洗包括对异常值的替换、采用均值对缺失值进行补全，以及删除脏数据；

优选地，所述非结构化数据为文本。

5、一种基于多模态融合的风险预测系统，其特征在于，所述系统包括：

特征提取模块，其用于对 EHR 数据进行特征提取，得到非结构化数据特征和结构化数据特征；

特征融合模块，其用于对非结构化数据特征和结构化数据特征进行融合处理并提取得到融合特征；

分类模块，其以提取的融合特征作为 input，得到疾病风险预测结果。

6、根据权利要求 5 所述的系统，其特征在于，所述特征提取模块包括结构化数据特征提取模块和非结构化数据特征提取模块；

其中，所述结构化数据特征提取模块以结构化数据作为 FCN 的 input，将数据映射到各个隐语义节点，得到结构化数据特征；

其中，所述非结构化数据特征提取模块采用 BERT 对非结构化数据进行特征提取；优选地，BERT 由 BERT Encoder 组成，BERT Encoder 由多层 BERT Layer 组成，每一层的 BERT Layer 均为 Transformer 中的 Encoder Block；每一个 encoder 层包含两层，分别为自注意力机制层和前馈神经网络层；

优选地，所述特征融合模块将非结构化数据特征和结构化数据特征沿指定维度进行并联，采用 SMOTE 通过对少数类样本数据进行分析并新生成该类样本的方法来降低不平衡率，然后采用分段池化操作，提取得到融合特征；

优选地，分类模块将融合特征或结构化数据作为 input 输入到全连接层，然后通过 Softmax 分类器对患者的结局进行预测；

优选地，所述系统还包括数据获取模块，其用于获取 EHR 数据；

优选地，所述系统还包括数据清洗模块，其用于在获取 EHR 数据后、在对 EHR 数据进行特征提取前预处理 EHR 数据，所述预处理包括对所述 EHR 数据清洗模块执行对异常值替

换和采用均值对缺失值进行补全并删除脏数据的操作；

优选地，所述系统还包括结果输出模块，其用于输出疾病风险预测结果。

7、一种处理 EHR 数据的方法，其特征在于，包括：

获取 EHR 数据，所述数据包括结构化数据和非结构化数据；

对结构化数据和非结构化数据分别进行数据处理，包括对结构化数据和非结构化数据分别进行数据清洗，对清洗后的结构化数据和非结构化数据分别进行特征提取，将分别提取得到的非结构化数据特征和结构化数据特征进行融合处理后提取融合特征；

以提取的融合特征数据作为待识别数据用于医疗用途；

优选地，所述数据清洗包括对异常值的替换、采用均值对缺失值进行补全，以及删除脏数据；优选地，所述非结构化数据为文本；

优选地，提取结构化数据特征采用 FCN；

优选地，提取非结构化特征采用 BERT；

优选地，所述提取融合特征的操作包括：将非结构化数据特征和结构化数据特征沿指定维度进行并联，采用 SMOTE 通过对少数类样本数据进行分析并新生成该类样本的方法来降低不平衡率，然后采用分段池化操作，提取得到融合特征。

8、一种疾病风险预测模型的构建方法，其特征在于，包括：

获取已知疾病风险结局的患者的 EHR 数据，所述数据包括结构化数据和非结构化数据；以获取的 EHR 数据构建数据集，包括结构化数据集和非结构化数据集，以已知的最终结局构建标签集；

构建疾病风险预测网络，包括：构建对于结构化数据进行提取的特征提取模块、对于非结构化数据进行提取的特征提取模块、和特征融合模块，结构化数据特征提取模块和非结构化数据特征提取模块并联连接后在特征融合模块决策层进行串联连接；所述疾病风险

预测网络基于 Pytorch 框架实现；

以标签集为标签，利用数据集训练构建的疾病风险预测网络，构建疾病风险预测模型；

优选地，构建数据集前包括对获取的 EHR 数据进行数据清洗的步骤，数据清洗包括对异常值的替换、采用均值对缺失值进行补全，以及删除脏数据；

优选地，对于结构化数据进行提取的特征提取模块为 FCN 模块；

优选地，对于非结构化数据进行提取的特征提取模块为 BERT 模块；

优选地，特征融合模块执行：将非结构化数据特征和结构化数据特征沿指定维度进行并联，采用 SMOTE 通过对少数类样本数据进行分析并新生成该类样本的方法来降低不平衡率，然后采用分段池化操作，提取得到融合特征；

优选地，利用数据集训练时，以融合特征作为 input 输入到全连接层，训练 Softmax 分类器。

9、一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求 1 至 4 中任一项所述方法的步骤；

和/或，所述处理器执行所述计算机程序时实现权利要求 7 中所述方法的步骤；

和/或，所述处理器执行所述计算机程序时实现权利要求 8 中所述方法的步骤。

10、一种计算机可读存储介质，其上存储有计算机程序指令，其特征在于，该程序指令被处理器执行时实现权利要求 1 至 4 中任一项所述方法的步骤；

和/或，该程序指令被处理器执行时实现权利要求 7 中所述方法的步骤；

和/或，该程序指令被处理器执行时实现权利要求 8 中所述方法的步骤。

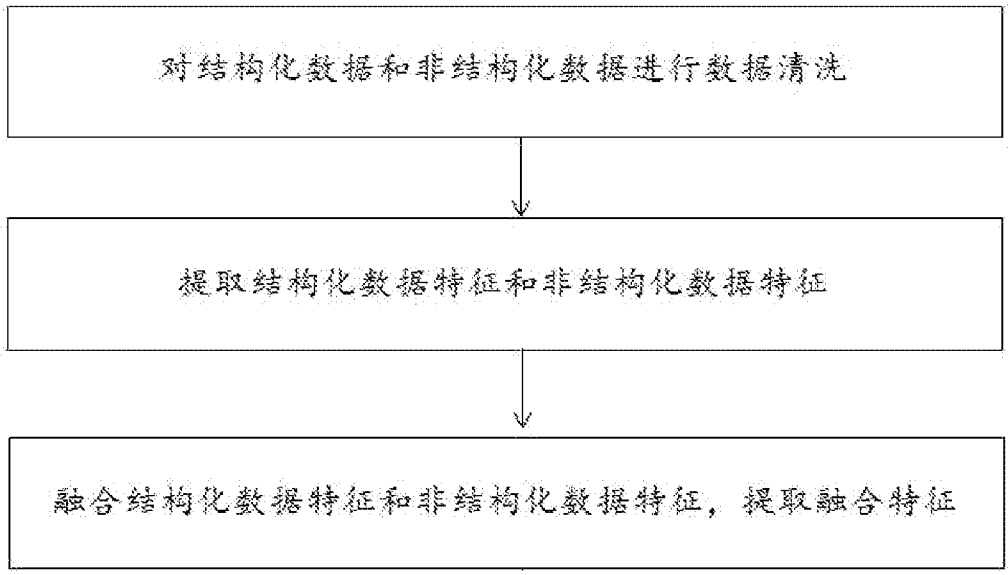

对结构化数据和非结构化数据进行数据清洗
提取结构化数据特征和非结构化数据特征
融合结构化数据特征和非结构化数据特征，提取融合特征

图 1

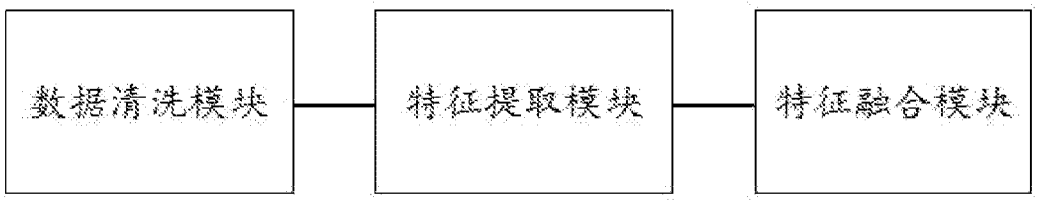

数据清洗模块
特征提取模块
特征融合模块

图 2

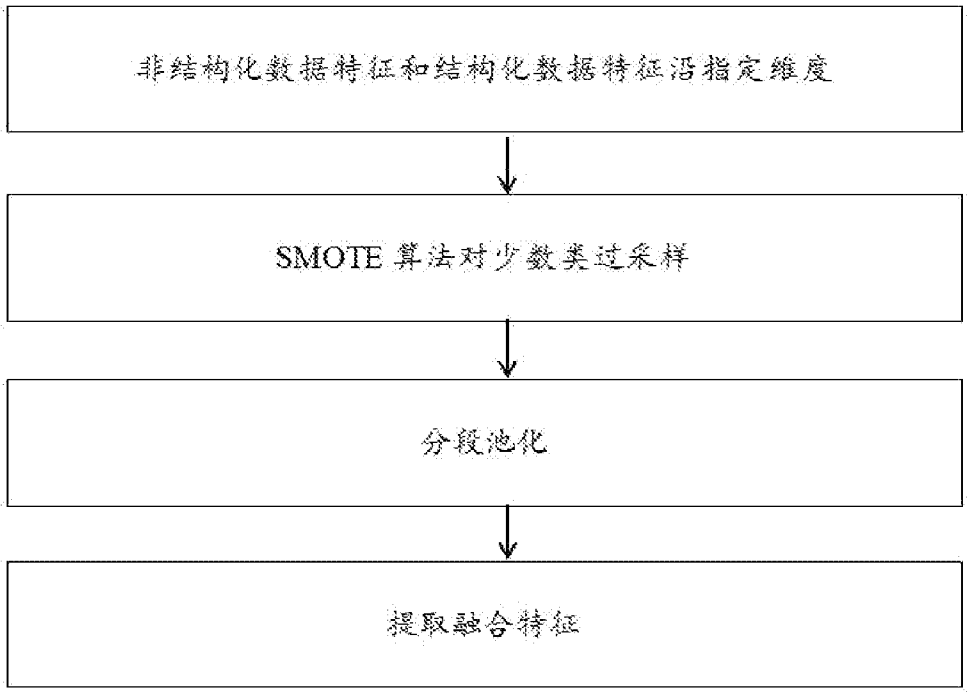

非结构化数据特征和结构化数据特征沿指定维度
SMOTE 算法对少数类过采样
分段池化
提取融合特征

图 3

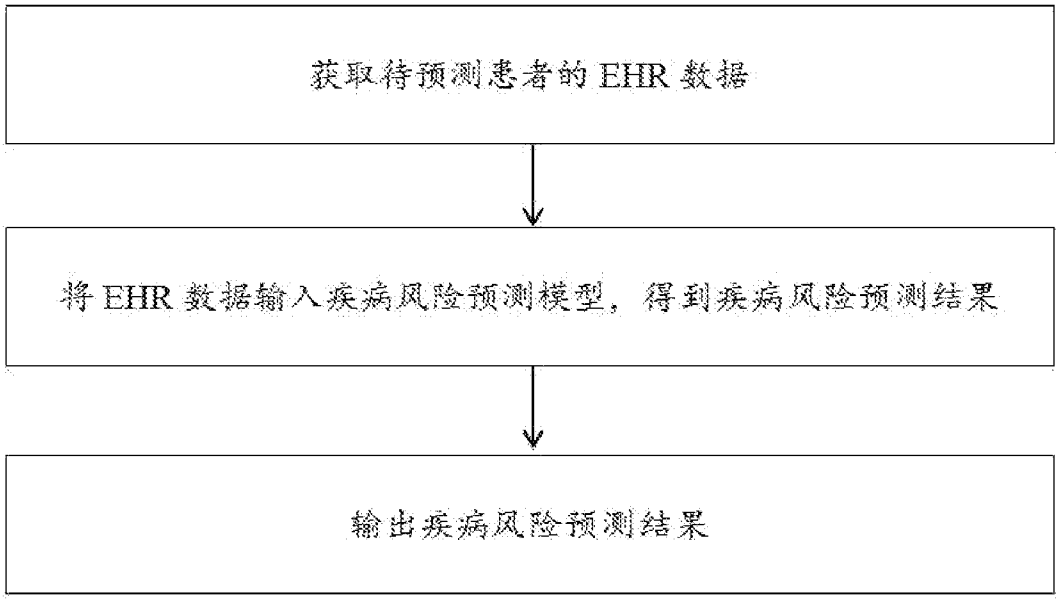

获取待预测患者的 EHR 数据
将 EHR 数据输入疾病风险预测模型，得到疾病风险预测结果
输出疾病风险预测结果

图 4

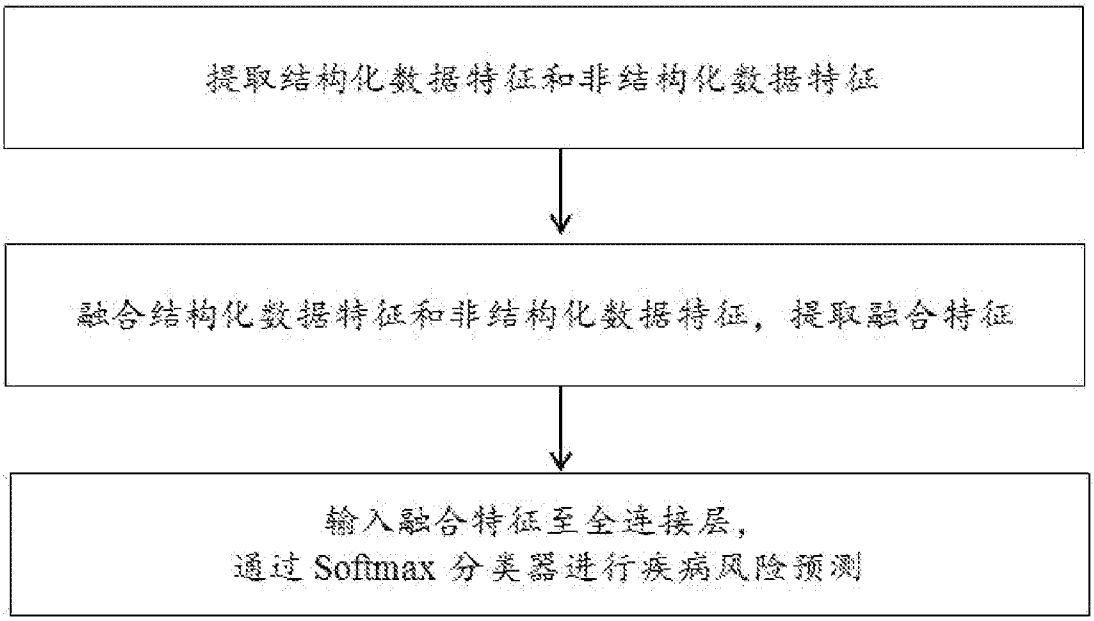

提取结构化数据特征和非结构化数据特征
融合结构化数据特征和非结构化数据特征，提取融合特征
输入融合特征至全连接层，
通过 Softmax 分类器进行疾病风险预测

图 5

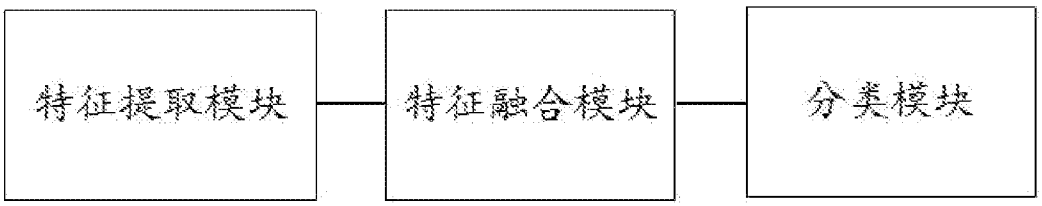

特征提取模块
特征融合模块
分类模块

图 6

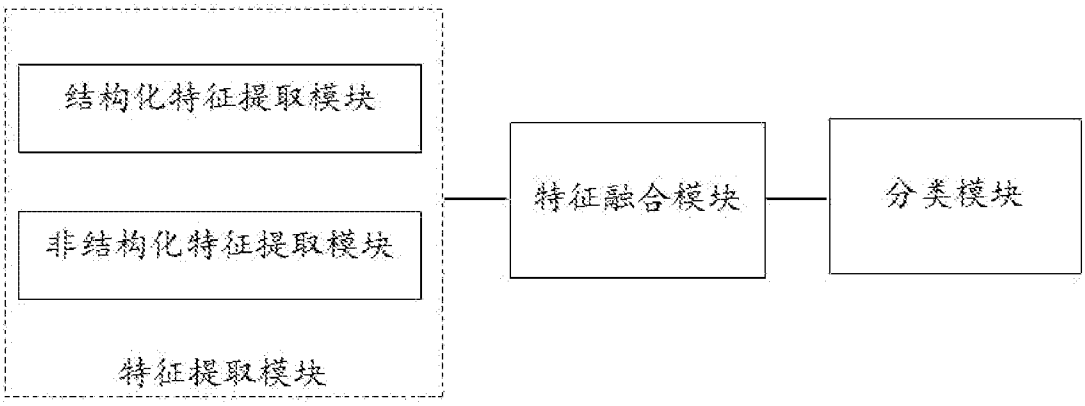

结构化特征提取模块
非结构化特征提取模块
特征提取模块
特征融合模块
分类模块

图 7

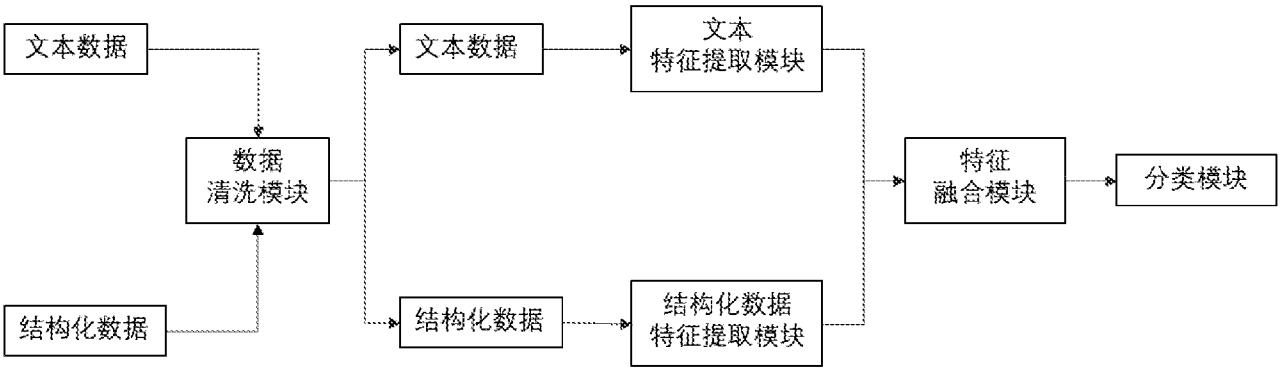

文本数据
结构化数据
数据
清洗模块
文本数据
结构化数据
文本
特征提取模块
结构化数据
特征提取模块
特征
融合模块
分类模块

图 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/106860

A. CLASSIFICATION OF SUBJECT MATTER

G16H 10/60(2018.01)i; G16H 50/20(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)

G16H10/-;G16H50/20

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)

CNABS, CNTXT, CNKI: 数据, 融合, 特征, BERT, 疾病, 预测, 病例, 电子, EHR, 结构化, 非结构化, 文本, 图片, 影像, SMOTE, FCN, Softmax, 数据清洗, 山东大学, 李玉军, 胡喜风, 刘治, 胡伟风; VEN, USTXT, WOTXT, EPTXT, 互联网检索, Internet Search: data, fusion, synthesis, confluence, feature?, characteristic?, disease, predict+, case note, electronic, structur+, unstructur+, text, picture?, image, clean+, Liu Zhi, EHR , BERT, SMOTE, FCN, Softmax;

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113241135 A (SHANDONG UNIVERSITY) 10 August 2021 (2021-08-10) claims 1-10	1-10
X	CN 111916207 A (HANGZHOU DEEPWISE BOLIAN TECHNOLOGY CO., LTD.; BEIJING DEEPWISE BOLIAN TECHNOLOGY CO., LTD.) 10 November 2020 (2020-11-10) description, paragraphs [0007]-[0016]	7
Y	CN 109117864 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 01 January 2019 (2019-01-01) description, paragraphs [0019]-[0085]	1-6, 9-10
Y	CN 111916207 A (HANGZHOU DEEPWISE BOLIAN TECHNOLOGY CO., LTD.; BEIJING DEEPWISE BOLIAN TECHNOLOGY CO., LTD.) 10 November 2020 (2020-11-10) description, paragraphs [0007]-[0016]	1-6, 9-10

☑ 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
"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	Date of mailing of the international search report
29 September 2021	**28 December 2021**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/106860

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Laila Rasmy等 (Rasmy, L. et al.). "“Med-BERT:pre-trained contextualized embeddings on larges-sacle structured electronic health records for disease prediction”" *arxiv.org/abs/2005.12833v1,* 22 May 2020 (2020-05-22), pp. 1-23	1-10
A	US 2016342764 A1 (UNIVERSIDAD DE VIGO) 24 November 2016 (2016-11-24) entire document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2021/106860

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 113241135 A	10 August 2021	None	
CN 111916207 A	10 November 2020	None	
CN 109117864 A	01 January 2019	CN 109117864 B	28 February 2020
US 2016342764 A1	24 November 2016	EP 3096253 A1	23 November 2016

国际检索报告

国际申请号
PCT/CN2021/106860

A. 主题的分类

G16H 10/60(2018.01)i; G16H 50/20(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)

G16H10/-;G16H50/20

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））

CNABS、CNTXT、CNKI:数据，融合，特征，BERT，疾病，预测，病例，电子，EHR，结构化，非结构化，文本，图片，影像，SMOTE，FCN，Softmax，数据清洗，山东大学，李玉军，胡喜风，刘治，胡伟风，VEN、USTXT、WOTXT、EPTXT、互联网检索：data，fusion，synthesis，confluence，feature?，characteristic?，disease，predict+，case note，electronic，structur+，unstructur+，text，picture?，image，clean+，Liu Zhi，EHR，BERT，SMOTE，FCN，Softmax;

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 113241135 A (山东大学) 2021年8月10日 (2021-08-10) 权利要求1-10	1-10
X	CN 111916207 A (杭州深睿博联科技有限公司 北京深睿博联科技有限责任公司) 2020年11月10日 (2020-11-10) 说明书第[0007]-[0016]段	7
Y	CN 109117864 A (华南理工大学) 2019年1月1日 (2019-01-01) 说明书第[0019]-[0085]段	1-6，9-10
Y	CN 111916207 A (杭州深睿博联科技有限公司 北京深睿博联科技有限责任公司) 2020年11月10日 (2020-11-10) 说明书第[0007]-[0016]段	1-6，9-10
A	Laila Rasmy等. "“Med-BERT:pre-trained contextualized embeddings on larges-sacle structured electronic health records for disease prediction”"《arxiv.org/abs/2005.12833v1》, 2020年5月22日 (2020-05-22), 第1-23页	1-10

☑ 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

“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	国际检索报告邮寄日期
2021年9月29日	2021年12月28日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 王倩 电话号码 (86-28)62967630

国际检索报告

国际申请号
PCT/CN2021/106860

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
A	US 2016342764 A1 (UNIVERSIDAD DE VIGO) 2016年11月24日 (2016 - 11 - 24) 全文	1-10

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/106860

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 113241135 A	2021年8月10日	无	
CN 111916207 A	2020年11月10日	无	
CN 109117864 A	2019年1月1日	CN 109117864 B	2020年2月28日
US 2016342764 A1	2016年11月24日	EP 3096253 A1	2016年11月23日