

融合表示学习和知识图谱推理的糖尿病及并发症预测方法

王宇翱^① 黄叶琪^① 李青远^① 刘云^② 景慎旗^② 单涛^② 郭永安^{*①}

^①(南京邮电大学智能信息处理与通信技术省高校重点实验室 南京 210003)

^②(江苏省人民医院信息处 南京 210029)

摘要: 糖尿病及其并发症的联合预测对于降低慢性病危害、改善患者预后具有重要意义。然而, 现有预测方法面临数据异构性和稀疏性、实体关系复杂以及疾病与医学概念间高阶关联难以精确捕捉等挑战, 限制了预测准确性和多病症识别能力。针对上述问题, 该文提出一种基于表示学习与知识图谱推理的糖尿病及其并发症预测模型(REKG-MDP)。通过整合电子健康记录与医学补充知识构建医疗知识图谱, 在患者侧完善个人基本信息、检查指标及现病史, 在疾病侧补充疾病共病信息、多发人群、常见病因及诊断依据, 从而缓解数据稀疏性与异构性问题。综合考虑对称、反对称、反转和组合4种关系连接模式, 并设计层次化注意力机制与图卷积网络相结合的推理模块, 在全局和局部动态调整邻居节点权重, 有效聚合多阶邻居信息并捕捉高阶语义关系。基于MIMIC-IV数据集的实验结果表明, 所提模型在糖尿病及并发症联合预测任务中明显优于现有方法, 预测准确率和多病症识别能力均有显著提升。

关键词: 多病症联合预测; 表示学习; 医疗知识图谱; 图神经网络; 注意力机制

中图分类号: TN912.34

文献标识码: A

文章编号: 1009-5896(2026)03-0971-11

DOI: 10.11999/JEIT250798

CSTR: 32379.14.JEIT250798

1 引言

糖尿病是一种常见的代谢性疾病, 与全身多个器官系统紧密关联^[1]。研究表明, 糖尿病引起的长期代谢紊乱会导致眼、肾和心脏等器官发生慢性进行性病变, 最终导致功能减退及衰竭, 是患者多系统并发症和死亡的重要原因^[2]。因此, 研究重点逐渐从单一疾病预测转向糖尿病及其并发症联合预测。与此同时, 人工智能模型在处理多源异构数据、捕捉复杂关联方面具备显著优势, 已成为疾病预测的重要工具。

尽管糖尿病预测模型的开发取得了显著进展, 但精确预测糖尿病及其并发症仍然面临多方面挑战。首先, 数据稀疏性使模型难以全面捕捉糖尿病病程发展规律。电子健康记录 (Electronic Health Record, EHR) 通常包括涵盖长文本 (如病程描述)、数值数据 (如检查指标)、短文本 (如性别、诊断结果) 等异构数据, 传统方法难以进行高效整合

和利用。其次, EHR数据结构复杂, 如表1所示, 其中症状、检查指标、疾病等医疗实体存在对称、反对称、反转和组合等关系连接模式。然而, 现有疾病预测模型大多忽视了针对关系模式的建模, 导致嵌入向量的表达精度不足。最后, 患者的生活方式、生理指标及多种疾病之间存在着复杂关联。例如, 身体质量指数 (Body Mass Index, BMI) 和血糖水平与糖尿病密切相关, 糖尿病与高脂血症关系紧密, 但与胰腺炎的关联较弱。这表明疾病间的相关性差异显著。在此背景下, 如何捕捉多病症场景下实体间相关性程度, 以充分挖掘关键信息, 成为当前研究亟需解决的难题。

针对上述挑战, 本文提出一种基于表示学习与知识图谱推理的糖尿病及其并发症预测模型 (Representation learning-Enhanced Knowledge Graph reasoning for Multiple Disease Prediction, REKG-MDP)。其核心思想是利用表示学习与知识图谱推理的优势, 缓解EHR数据稀疏性和异构性, 同时捕捉多疾病间复杂的语义关系, 提升糖尿病及其并发症的预测性能。具体而言, 该文贡献总结为:

(1) 提出基于表示学习的医疗知识图谱 (Medical Knowledge Graph, MKG) 构建方法, 在患者侧引入基本信息、检查指标和病史, 在疾病侧补充共病信息、多发人群、病因和诊断依据, 实现异构信息的融合表征, 缓解医疗数据稀疏性与异构性带来的建模难题;

(2) 提出关系模式建模机制, 综合考虑对称、反对称、反转和组合4种关系模式, 在知识图谱推

收稿日期: 2025-08-26; 改回日期: 2025-10-27; 网络出版: 2025-11-04

*通信作者: 郭永安 guo@njupt.edu.cn

基金项目: 国家重点研发计划(2023YFC3605800), 江苏省前沿引领技术基础研究专项(BK20202001), 江苏省研究生科研与实践创新计划项目(SJCX24_0285)

Foundation Items: The National Key Research Program of China (2023YFC3605800), The Frontier Leading Technology Basic Research Program of Jiangsu Province (BK20202001), The Post-graduate Research & Practice Innovation Program of Jiangsu Province (SJCX24_0285)

表1 医疗知识图谱中的关系连接模式示例

关系连接模式	解释	医疗案例
对称模式	两个实体之间的关系是相互的, 即如果A与B有这种关系, 那么B也应该与A有这种关系	(糖尿病, 共病, 高脂血症)
反对称模式	如果A与B有这种关系, 那么B与A没有这种关系	(患者, BMI, 肥胖)
反转模式	在某些条件下, 这导致原始关系的反转, 即如果存在 $r_1(A, B)$, 那么存在 $r_2(B, A)$	(高血糖, 导致, 糖尿病) →(糖尿病, 风险因素, 高血糖)
组合模式	一个实体可以通过一系列关系与另一个实体连接, 即如果存在 $r_1(A, B)$ 和 $r_2(B, C)$, 那么可以推断出 $r_3(A, C)$	(患者, 有, 异常检查指标)+ (疾病, 诊断依据, 异常检查指标) →(患者, 患有, 疾病)

理过程中更准确地刻画患者、疾病及诊断依据之间的复杂语义联系, 提升了嵌入表示的表达能力;

(3)设计层次化注意力机制与图卷积网络相结合的推理模块, 在全局和局部同时度量不同阶邻居节点对目标节点的重要性, 并在三元组级别分配权重, 结合图卷积网络聚合节点及其高阶邻居特征, 实现对患者与疾病高阶拓扑信息的有效建模, 提升预测准确率;

(4)在MIMIC-IV数据集上开展实验验证, 并与多种经典算法对比, 结果表明所提模型在糖尿病及其并发症预测任务上显著优于现有方法。

2 相关工作

2.1 单疾病预测模型

机器学习(Machine Learning, ML)和深度学习(Deep Learning, DL)等人工智能算法能够从医疗数据中学习潜在规律, 广泛应用于慢性病的预测、筛查和管理^[3]。传统ML方法多依赖特征工程和采样策略来应对数据不平衡与缺失问题。例如, Zhang等人^[4]结合混合采样和模糊K最近邻(Fuzzy K-nearest Neighbor, FKNN)分类器构建深静脉血栓形成(Deep Vein Thrombosis, DVT)预测模型, 通过全局分析与局部分析对不同风险因素进行评估, 以判断患者是否存在DVT。Rahman等人^[5]利用链式方程多重插补法和边界线合成少数类过采样技术解决数据缺失和类别不平衡的问题, 采用递归特征消除和Boruta算法进行特征选择, 通过患者的各种临床特征判定患者是否患有慢性肾脏病。Althobaiti等人^[6]将梯度提升机与数据降维单元相结合, 提高数据存储效率和糖尿病预测模型性能, 同时采用了包括Bagging和Boosting在内的集成学习方法, 以提升模型的性能与鲁棒性。同时, Ssulami等人^[7]采用ML和数据增强技术进行冠心病的预测, 利用被错误分类的实体进行数据增强, 让模型在训练过程更多地关注难以分类或者错误分类的实例以提高预测准确性。利用ML进行疾病预测可以实现单个疾病高准确度的诊断率。然而, ML中基于特征选择

的阈值设置会导致某些属性特征的丢失而降低预测的准确度。

随着DL的发展, 研究者尝试利用多层神经网络自动提取表征, 以减轻特征工程依赖。金怀平等^[8]对不同分辨率下的切片进行深度特征提取和融合, 预测胃癌是否发生远处转移。季薇等人^[9]利用对抗迁移学习和特征解耦, 提取跨语种语音数据中的病理信息, 实现帕金斯疾病检测。Ghorbani等人^[10]利用图卷积网络对不平衡数据进行加权建模, 避免分类器偏向多数类样本。然而, 此类方法多针对单一疾病, 不能分析利用多源医疗信息与疾病间的复杂联系, 难以应对多并发症预测需求。

2.2 多病症预测模型

相比单疾病预测, 多疾病联合预测更符合临床实际需求, 引起国内外广泛研究。Zhao等人^[11]采用图神经网络进行慢性阻塞性肺疾病及相关疾病预测, 并建立了包含患者各类病理信息以及医学概念的知识图谱。该知识图谱采用通过节点邻域的采样和特征聚合来生成节点的嵌入向量, 不直接对关系进行建模, 限制了链接预测的能力。Pham等人^[12]采用正负空间对比判断不同属性与不同疾病的相关性以及疾病之间的相关性, 实现多病症联合预测。但其语义建模层次较浅, 无法捕捉不同疾病之间的高阶和非线性关系。Qu等人^[13]基于异构图注意力网络模型实现多病症预测。Lu等人^[14]将患者EHR构建为异构图, 设计上下文感知的动态图学习方法, 利用全局疾病图和就诊患者子图探索患者疾病发展模式, 基于前一次诊断数据预测患者后续可能发展的疾病。然而只采用了患者EHR的部分信息, 存在数据稀疏性的问题。

部分工作引入学习机制提升预测模型性能。例如, 熊立鹏等人^[15]使用极限梯度提升树结构对数据进行编码提取其特征重要性, 并使用长短期记忆神经网络, 捕捉不同特征之间的依赖关系, 预测肺部手术后并发症的发生。Sun等人^[16]利用强化学习策略和价值评估机制, 在知识图谱上进行随机游走的方式进行推理, 构建了可预测疾病进展路径并提供

可解释性的模型。但这些方法在关系表示上大多采用独热编码或浅层特征聚合，无法捕捉到关系之间的语义相似性或层次结构，表示能力有一定的局限性。综上所述，现有多病症预测模型为联合建模疾病关系提供了可能，但现有方法对关系模式的建模不足、语义表达有限，难以捕捉糖尿病及其并发症间复杂的语义模式与发展路径，从而限制了预测的精度与临床价值。

2.3 知识图谱嵌入模型

知识图谱通过刻画实体及其关系为疾病预测提供了良好的结构化语义支撑。知识图谱嵌入方法通过将实体和关系映射到低维向量空间，在方便计算的同时保留知识图谱中的结构信息^[17]，是实现知识图谱推理的重要方法之一。典型方法包括：文献[18]提出的TransE将关系视为从头实体到尾实体的向量平移，结构简单但无法区分对称模式；文献[19]将实体和关系分别嵌入到不同空间，提升表达能力但难以捕捉反转和组合模式；文献[20]在复数与旋转空间中实现更灵活的关系建模，能够部分解决上述问题。

尽管已有方法在一般知识图谱任务中表现良好，但医疗场景中的关系类型复杂多样，涉及对称、反对称、反转以及组合关系。现有嵌入模型往往无法同时兼顾这些模式，限制了在医疗知识图谱推理中的应用效果。因此，需要设计能够兼容多种关系模式的知识图谱嵌入方法，以更好地捕捉患者与疾病间的复杂语义联系，为多病症预测提供更精确的语义表示。

3 模型设计与实现

本文所提REKG-MDP模型的整体框架如图1

所示，其整体工作流程包括4个阶段：首先，基于HER和补充医疗信息构建MKG，其中节点表示患者、疾病和检验指标等实体，边表示其语义关系；其次，利用复数空间的表示学习方法将实体和关系映射为潜在向量，以捕捉语义特征和结构模式；然后，通过层次化注意力机制在三元组级别建模局部重要性并在全局建模不同邻居层次的重要性，结合图卷积网络聚合节点及多阶邻居信息，得到更具语义表达力的嵌入表示；最后，基于更新的嵌入向量计算患者与疾病的关联强度，输出糖尿病及潜在并发症的预测概率，实现多病症联合预测。

3.1 知识图谱构建

选取公开数据集MIMIC-IV^[1]中与糖尿病相关的疾病数据，并在此基础上构建MKG。由于数据清理、解密与去标识化等过程超出本文范围，此处省略。知识图谱的构建过程如图2所示：首先整合糖尿病患者的最后一次医疗记录，涵盖个人信息、生活习惯、现病史、症状和检查指标；其次注入相关医学背景知识，增强疾病、症状和检查指标等实体的语义丰富性；随后对年龄、检查指标、BMI等数值特征进行医学标准分级并赋予解释；接着将实体及其关系表示为三元组，如(患者ID, 患有, 二型糖尿病)或(糖尿病, 可能导致, 心脑血管疾病)；最后对多义或同义实体进行消歧^[21]，确保知识图谱的一致性与可扩展性。

如图3所示，构建的知识图谱包括患者-疾病二部图、局部知识图谱和全局知识图谱3种类型。

(1)患者-疾病二部图：在多病症疾病预测场景下，使用 $p \in \mathcal{P}$ 表示患者、 $d \in \mathcal{D}$ 表示疾病。患者-疾病二部图 $\mathcal{G}_{pd} = \{(p, t_{pd}, d)\}$ ，当 $t_{pd} = 1$ 时，表明患者 p 具有疾病 d ，反之， $t_{pd} = 0$ 。

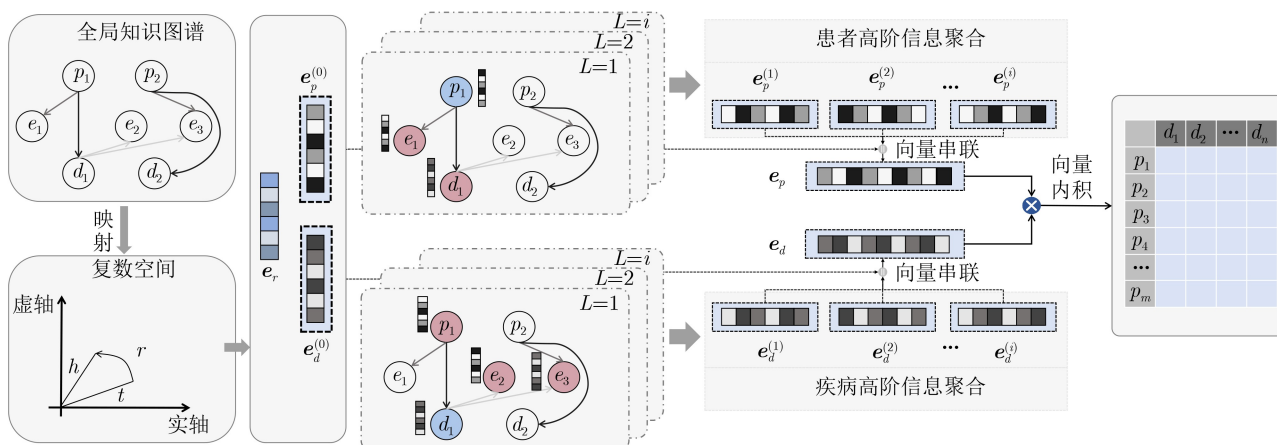


图1 REKG-MDP模型架构图

¹⁾ <https://physionet.org/>，该数据集由美国国立卫生研究院提供

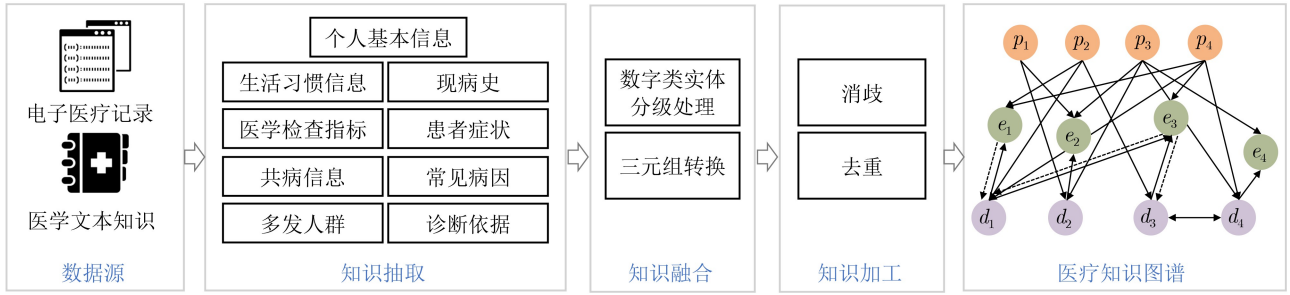


图2 知识图谱构建流程图

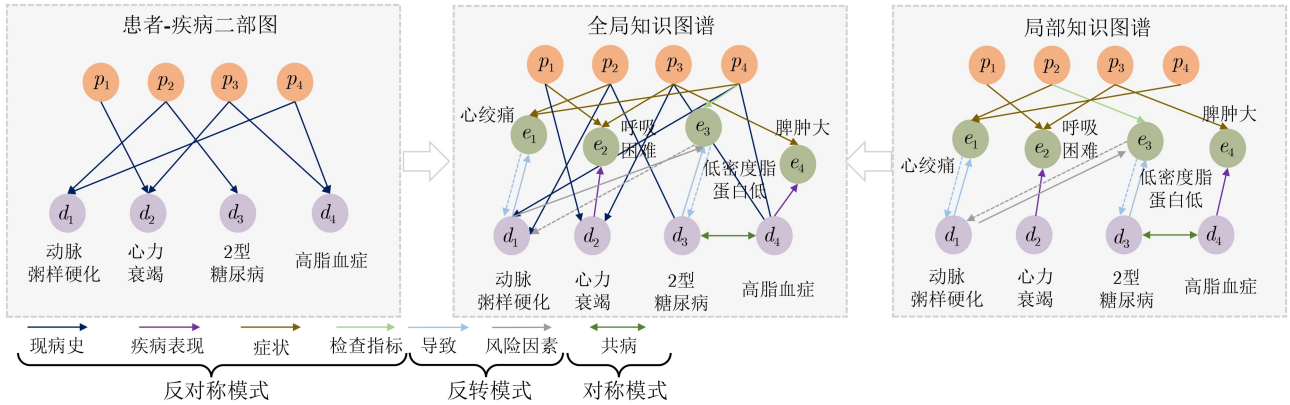


图3 医疗领域知识图谱示例图

(2)局部知识图谱：是由EHR和医学补充知识构成的。其中， $\mathcal{G}_{loc} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$ ， $\mathcal{T} = \{(h, r, t)\}$ 表示 \mathcal{G}_{loc} 中包含的所有三元组。 $h, t \in \mathcal{E}$ ， $r \in \mathcal{R}$ ，其中， \mathcal{E} 是所有节点的集合， \mathcal{R} 代表除患者与疾病关系之外的所有其他关系的集合。显然， $p \in \mathcal{P} \in \mathcal{E}$ 并且 $d \in \mathcal{D} \in \mathcal{E}$ 。

(3)全局知识图谱： $\mathcal{G}_{glo} = (\mathcal{E}_S, \mathcal{R}_S, \mathcal{T}_S)$ 拥有完整的已知关系信息，其中 $\mathcal{E}_S = \mathcal{E}$ ， $\mathcal{R}_S = \mathcal{R} \cup t_{pd}$ ， $\mathcal{T}_S = \{(p, t_{pd}, d)\} \cup \{(h, r, t)\}$ 。

3.2 嵌入

知识图谱嵌入是将实体和关系转化为连续的向量空间，简化操作的同时保留知识图谱原有结构^[22]。对于给定的实体 h, t 和关系 r ，将其映射到复数向量空间中，其关系表示为

$$e_t = e_h \circ e_r, |e_{ri}| = 1 \quad (1)$$

其中， $e_h, e_r, e_t \in \mathbb{C}^k$ 分别是头实体、关系和尾实体的复数嵌入向量， \mathbb{C}^k 表示一个 k 维的复数集合。 \circ 表示逐元素相乘。关系 r 在复数空间中任何维度 i 下的嵌入向量的模为1。

在MKG中不同实体之间的关系也分别属于不同的关系连接模式，因此，在嵌入关系时有以下限制：

(1)当关系为对称模式时，即 $e_t = e_h \circ e_r$ 且 $e_h = e_t \circ e_r$ 时，要求 $e_{ri} = e^{0/i\pi} = \pm 1$ 。类比地，当关系为反对称模式时， $e_{ri} \neq \pm 1$ 。

(2)当关系为反转模式时，即 $e_t = e_h \circ e_{r_1}$ 且 $e_h = e_t \circ e_{r_2}$ ，并且因为 $r_1 \neq r_2$ ，则要求 $e_{r_1} = \overline{e_{r_2}}$ 。

(3)当关系为组合模式时，即 $e_{t_1} = e_{h_1} \circ e_{r_1}$ ， $e_{t_2} = e_{t_1} \circ e_{r_2}$ 且 $e_{t_2} = e_{h_1} \circ e_{r_3}$ ，则要求 $e_{r_3} = e_{r_1} \circ e_{r_2}$ 。

通过对4类连接模式的有效建模，不仅能够嵌入空间中保持关系的语义一致性，还能捕捉疾病之间的潜在逻辑关联，从而提升预测的准确性与医学可解释性。对于每个三元组 (h, r, t) ，评分函数 $f_r(h, t) = \|e_h \circ e_r - e_t\|_1$ ，其中 $\|\cdot\|_1$ 表示1范数，即向量元素绝对值之和。评分函数 $f_r(h, t)$ 的分数越低，说明头实体 h 和尾实体 t 之间存在关系 r 的可能性越大，反之，则可能性越小。

在知识图谱中，由关系连接的节点组成的三元组被视为真实三元组，其中的头尾实体构成正样本，为了训练模型，随机替换真实三元组中的头尾实体以生成负样本构造了损失函数 \mathcal{L}_{KG} ，旨在优化实体和关系在向量空间中的表示。首先引入softplus激活函数来平滑地处理得分差异，则损失函数 \mathcal{L}_{KG} 为

$$\mathcal{L}_{KG} = \text{softplus}(\gamma - f_r(h, t)) + \frac{1}{m} \sum_{i=1}^m \text{softplus}(f_r(h'_i, t'_i) - \gamma) \quad (2)$$

其中， γ 是一个固定的间隔， m 是每个正样本对应的负样本采样个数， β 调节了softplus函数对输入的

敏感性，较大的 β 值使得函数对输入的变化更加敏感，从而影响正负样本的区分度以及模型的性能。

3.3 基于层次化注意力机制的信息传播

3.3.1 注意力机制

考虑到不同实体对于目标实体的影响存在差异，在信息传播过程中采用注意力机制以计算不同相邻节点的权重分数，从而识别对目标节点更具贡献的邻居

$$\text{att}(h, r, t) = \exp\left(-\frac{f_r(\mathbf{h}, \mathbf{t})}{\tau}\right) \quad (3)$$

式中， $\exp(x)$ 为指数函数 e^x ， τ 是温度参数，用于控制注意力分数的分布范围。为确保权重分数的可比性，尤其是在处理同一头实体对应的多个邻居时，对其进行规范化处理

$$\text{att}(h, r, t) = \frac{\exp(\text{att}(h, r, t))}{\sum_{(h, r', t') \in \mathcal{N}_h} \exp(\text{att}(h, r', t'))} \quad (4)$$

式中， \mathcal{N}_h 表示以头实体 h 为中心节点的所有1阶邻居节点^[23]。基于注意力机制的权重分配策略不仅在局部增强了对与目标实体距离更近节点的关注，还可以抑制可能存在的噪声与冗余信息。为更有效地融合多阶邻居信息，进一步设计了层次化注意力机制，计算层次化注意力分数为 $\text{att}_{\text{lev}}(h, r, t, l) = \text{att}(h, r, t) \cdot S_l$ ，对于每1阶邻居节点，构建层次化权重系数 S_l 以捕捉全局重要性信息

$$S_l = \frac{\exp(\theta_l)}{\sum_{k=l}^l \exp(\theta_k)} \quad (5)$$

通过设置层次化权重系数 S_l 和可训练参数 θ_l ，促进模型自动学习不同阶邻居(如1阶、2阶等)对目标节点的贡献程度，动态调整不同层邻居的全局重要性。

3.3.2 1阶信息传播

为有效建模MKG中复杂的多跳关系(例如疾

病—症状—并发症的层级关联)，提出基于层次化注意力机制的图卷积网络聚合策略。具体而言，遍历每个头实体 h 及其1阶邻居构成的三元组 (h, r, t) ，并利用层次化注意力分数 $\text{att}_{\text{lev}}(h, r, t, l)$ 来衡量尾实体的重要性，从而对1阶邻居信息进行加权求和。为充分融合头实体 h 与其1阶邻居节点的嵌入信息，采用双向交互聚合策略。具体而言，将头实体与其1阶邻居节点信息进行线性组合与元素级乘积^[24]，并结合层次化注意力机制捕捉更复杂的交互模式。

$$\begin{aligned} \mathbf{e}_h^{(1)} = f_{\text{mix}}(\mathbf{e}_h^{(0)}, \mathbf{e}_t^{(0)}) = & \text{LeakyReLU}\left(\mathbf{W}_1\left(\mathbf{e}_h^{(0)}\right.\right. \\ & + \sum_{(h, r, t) \in \mathcal{N}_h} \text{att}_{\text{lev}}(h, r, t, l) \mathbf{e}_t^{(0)}\left.\right) + \mathbf{W}_2\left(\mathbf{e}_h^{(0)}\right. \\ & \left.\left. \circ \sum_{(h, r, t) \in \mathcal{N}_h} \text{att}_{\text{lev}}(h, r, t, l) \mathbf{e}_t^{(0)}\right)\right) | l = 1 \quad (6) \end{aligned}$$

式中， $\mathbf{W}_1, \mathbf{W}_2 \in R^{d \times k}$ 是两个可训练的权重矩阵， k 是输入实体嵌入向量的维度， d 是输出向量的维度。LeakyReLU函数作为一种非线性激活函数，用于引入非线性特性，避免梯度消失。融入层次化注意力机制的双向交互聚合方式能够更有效地融合头实体与邻居节点信息，提升模型在多病症预测任务中的性能。

3.3.3 高阶信息传播

为深入挖掘并利用知识图谱中的高阶邻居信息，采用逐步迭代的方法，如图4所示，随着聚合过程的推进，逐步纳入更高阶的邻居信息，从而不断完善实体 h 的嵌入向量，捕捉知识图谱的全局结构信息和语义信息，提升实体嵌入的精确度。具体而言，中心节点 h 的 l 阶嵌入向量 $\mathbf{e}_h^{(l)}$ 表示为

$$\mathbf{e}_h^{(l)} = \sum_{t \in \mathcal{N}_h} f_{\text{mix}}(\mathbf{e}_h^{(l-1)}, \mathbf{e}_t^{(l-1)}), l = 2, 3, 4, \dots \quad (7)$$

式中， $\mathbf{e}_h^{(l-1)}$ 是中心节点 h 的 $l-1$ 阶的嵌入向量，

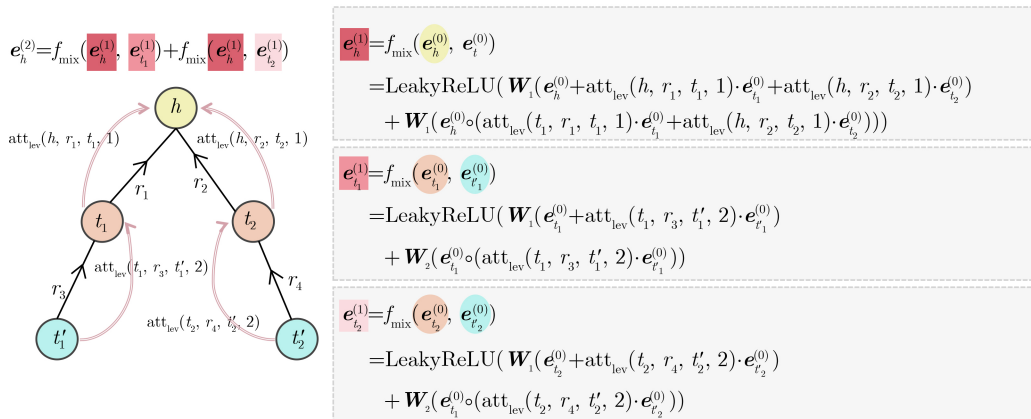


图4 节点嵌入向量聚合过程

$e_t^{(l-1)}$ 是中心节点 h 的 $l-1$ 阶的 1 阶邻居节点嵌入向量, $e_h^{(0)}, e_t^{(0)}$ 表示实体 h, t 的初始嵌入向量。

3.4 多病症疾病预测

经过 l 层的图卷积网络传播后, 每个节点都会获得从第 0 层到第 l 层的嵌入向量, 这些嵌入向量捕捉了节点在不同网络深度下的特征。嵌入过程为

$$\mathbf{h}_p = \mathbf{e}_p^{(0)} \parallel \dots \parallel \mathbf{e}_p^{(l)}, \mathbf{h}_d = \mathbf{e}_d^{(0)} \parallel \dots \parallel \mathbf{e}_d^{(l)} \quad (8)$$

式中, $\mathbf{e}_p^{(l)}, \mathbf{e}_d^{(l)}$ 分别表示患者节点 p 和疾病节点 d 在第 l 层的嵌入向量。 \parallel 表示向量串联操作, 用于将各层的嵌入向量组合成一个完整的嵌入向量。利用全局知识图谱 \mathcal{G}_{glo} , 可以获取到患者 p 和疾病 d 的完整嵌入向量 \mathbf{h}_p 和 \mathbf{h}_d 。通过向量的点积运算 $\hat{y}(p, d) = \mathbf{h}_p^T \mathbf{h}_d$ 来计算两者之间的预测分数。为了提升模型的准确度, 采用 BPR 损失进行损失函数的计算^[25]

$$\mathcal{L}_{\text{CF}} = \sum_{(p, d, d') \in \mathcal{G}_{pd}} -\ln \sigma(\hat{y}(p, d) - \hat{y}(p, d')) \quad (9)$$

式中, (p, d) 是正样本, (p, d') 是负样本。当患者 p 患有疾病 d 的概率越高, 疾病预测损失函数 \mathcal{L}_{CF} 越小。结合知识图谱损失函数 \mathcal{L}_{KG} 、疾病预测损失函数 \mathcal{L}_{CF} 以及正则化项, 定义综合损失函数 \mathcal{L} 为

$$\mathcal{L} = \mathcal{L}_{\text{KG}} + \mathcal{L}_{\text{CF}} + \frac{\lambda}{|\mathcal{E}_S|} \sum_{e \in \mathcal{E}_S} \|e\|_2^2 \quad (10)$$

式中, $|\mathcal{E}_S|$ 表示节点集合 \mathcal{E}_S 的大小。 λ 是正则化系数, 用于控制正则化项的强度。 $\|e\|_2^2$ 表示所有节点嵌入向量的 L2 范数平方。通过最小化综合损失函数 \mathcal{L} 对模型进行训练和优化。

4 实验

4.1 数据集设置

采用 MIMIC-IV 数据集和补充医学知识构建全局知识图谱, 其中从 ADMISSIONS 和 PATIENT 表中提取个人信息, 从 LABEVENTS 表提取检查指标,

表 2 知识图谱统计信息

数据类型	数据集大小
训练集	2910
测试集	1942
疾病数量	18
患者数量	4852
检查指标数量	92
基本个人信息类型数量	18
共病/常见病因/多发人群数量	485
关系类型数量	18
知识图中的三元组数量	163118

从 DIAGNOSES 表提取诊断信息。并结合公开医学资料与权威医学知识库, 获取疾病共病信息、多发人群、常见病因和诊断依据, 其统计信息如表 2 所示。

在患者筛选过程中, 先收集每位患者的最后一次就诊记录, 根据诊断信息筛选出患糖尿病的 20000 名患者, 再结合血糖及相关检查指标进行二次判别, 最终保留 4852 名患者。考虑到代谢性疾病、心脑血管疾病、肾脏疾病及非酒精性脂肪性肝病之间存在复杂关联^[26], 选择其中 18 种疾病作为预测标签, 详见表 3。数据集划分为训练集(约 60%)和测试集(约 40%), 在预测时, REKG-MDP 以患者基本信息、症状和检查指标为输入, 输出每位患者 n 种疾病的预测结果, 其中 n 可以根据需求确定。

4.2 评估指标设置

采用精确率(Precision, P)、F1 分数(F1 Score, F1)、归一化折损累计增益(Normalized Discounted Cumulative Gain, NDCG)3 种指标来衡量模型的预测性能, 以下对各指标进行简要说明。

(1) $P@n$: 对于每个患者, 计算前 n 个预测项中实际为正样本的数量与所有预测个数的比例^[11]。

(2) $F1@n$: F1 分数用于综合衡量精确率和召回率的性能^[12], 其中, 召回率为对于每个患者, 计算前 n 个预测项中实际为正样本的数量与患者真实患病数量的比例。

(3) $NDCG@n$: 用来衡量模型将患者实际患有的疾病放在推荐列表中较高位置的能力^[24]。

表 3 该文中使用的疾病信息和疾病分类

疾病类别	疾病 ICD-10 代码	疾病名称
代谢性疾病	E11	2 型糖尿病
	E78.5	高脂血症
	E11.4	糖尿病性神经病变
	E10.2&E11.2	糖尿病性慢性肾病
	E10.65&E11.65	高血糖症
	E10	1 型糖尿病
	E78.0	高胆固醇血症
心脑血管疾病	E10.1&E11.1	糖尿病酮症酸中毒
	I10	高血压
	150	心力衰竭
	125.1	冠状动脉粥样硬化性心脏病
	121	心肌梗死
肾脏疾病	163	缺血性中风
	G45	短暂性脑缺血发作
	170	动脉粥样硬化
非酒精性脂肪肝病	N18	慢性肾病
	K75.S1	非酒精性脂肪性肝炎
	K76.0	脂肪肝

以上3种评估指标中的 n 分别设置为1, 3, 5。 $n = 1$ 表示衡量模型在单疾病预测时的准确度, $n = 3$ 和 $n = 5$ 来衡量模型的多疾病联合预测性能。

4.3 对比方法

为全面评估REKG-MDP的性能,选取了两种ML方法和两种图模型作为对比。其中前两者经过微调以支持多病症预测^[27]。具体包括(1)bSES-AC-RUN-FKNN^[4]:利用SES-AC-RUN进行特征子集搜索并基于FKNN完成DVT预测。(2)DCKD-RF^[5]:结合随机森林与特征选择方法对临床特征进行建模,用于CKD预测。(3)PyRec^[28]:基于图神经网络建模实体交互以缓解数据稀疏性,实现更精准的推荐/预测。(4)KGRec^[29]:基于图神经网络的知识图谱推荐模型,采用加权知识掩码自编码器自适应屏蔽高分三元组,并结合对比学习方法增强模型对关键连接的重建能力,从而提升预测性能。

4.4 参数设置

在对比实验中,bSES-AC-RUN-FKNN因未处理缺失值,仅使用所有患者均具备的个人信息、症状及检查指标作为特征,共计136项,并基于欧氏距离的FKNN进行训练(训练集和测试集为90%和10%,模糊参数为2,邻居数 $k = 5$)。DCKD-RF采用插补方法补充BMI数据(4852名患者中3919人有BMI信息),以符合原始设定;PyRec和KGRec则使用相同数据集,其中KGRec因数据划分方式对源代码的评估模块做了微调,不影响核心算法。REKG-MDP的主要超参数为负采样数 $m = 3$, $\beta \in \{1, 2, 3\}$, $\gamma = 3$,嵌入向量维度 $\dim \in \{32, 64, 128, 256\}$,学习率为0.001。模型采用Adam优化器^[30],在Ubuntu 20.04和NVIDIA RTX 3090 GPU上运行,所有实验均沿用相同评估指标与 n 值设定,以保证可比性。

4.5 实验结果

表4展示了REKG-MDP模型与4种基线方法的性能对比。其中,最优的性能通过加粗表示,次优的性能通过下划线表示,次优和最优之间的性能差距由括号里面的百分数表示。在4种基线方法中,REKG-MDP模型的预测准确度均优于所有基线方

法。具体来说,当 $n = 1$ 时, P , F1, NDCG分别提高了19.39%, 19.67%, 19.39%;当 $n = 3$ 时,3种指标分别提高了16.71%, 21.83%, 23.53%;当 $n = 5$ 时,3种指标分别提高了22.01%, 20.34%, 20.88%,验证了REKG-MDP在疾病预测中的有效性。此外,基于图神经网络的KGRec和PyRec模型整体优于DCKD-RF和bSES-AC-RUN-FKNN两类传统ML模型,这归因于图神经网络能够高效聚合邻居节点信息,捕捉复杂数据关系,从而在预测任务中展现更高准确性。REKG-MDP在此基础上进一步优化关系建模和信息传播策略,因此表现出更高的准确性。

4.6 消融实验

4.6.1 不同模块的影响

为了评估不同模块对REKG-MDP性能的影响,设计了3项消融实验,具体为(1)REKG-MDP -V1:仅考虑对称和反对称两种关系模式,用于检验复杂关系建模的作用。(2)REKG-MDP -V2:去除层次化注意力机制,利用GraphSAGE模型^[31]实现信息聚合,以评估双向交互信息聚合方式的贡献。(3)REKG-MDP -V3:不引入补充医学知识,仅使用EHR构建知识图谱,以探究图谱稀疏性对模型性能的影响。

如图5所示,完整的REKG-MDP在 P , F1和NDCG上均取得最佳表现,验证了关系模式兼容性、层次化注意力聚合及疾病侧信息3个关键模块的有效性与必要性。具体而言,REKG-MDP -V3因未利用疾病侧信息而表现最差,其在 P , F1, NDCG 3个关键指标上分别平均下降了19.43%, 20.29%, 16.84%,说明该信息对预测性能贡献最大。此外,去除复杂关系模式后的REKG-MDP -V1变体性能分别下降12.48%, 12.81%, 9.90%,表明关系建模对于捕捉节点间复杂关联至关重要。REKG-MDP -V2的性能与其他变体相比较佳,其在3个关键指标上下降幅度分别为5.54%, 5.71%, 4.86%。这一结果表明,层次化注意力机制虽影响相对较小,但仍能增强全局依赖与局部交互建模能力。

4.6.2 不同参数的影响

如图6所示,嵌入向量维度 \dim 直接影响REKG-

表4 REKG-MDP模型与5种基线方法的性能对比

模型	$P@1$	$P@3$	$P@5$	F1@1	F1@3	F1@5	NDCG@1	NDCG@3	NDCG@5
REKG-MDP	0.9655 (↑ 19.39%)	0.8879 (↑ 16.71%)	0.8280 (↑ 22.01%)	0.4200 (↑ 19.67%)	0.7332 (↑ 21.83%)	0.8121 (↑ 20.34%)	0.9655 (↑ 19.39%)	0.9151 (↑ 23.53%)	0.8946 (↑ 20.88%)
DCKD-RF	0.7199	0.4192	0.3455	0.3058	0.3569	0.3375	0.7199	0.4651	0.4329
bSES-AC-RUN-FKNN	0.7106	0.4670	0.3995	0.3086	0.3972	0.3902	0.7106	0.4855	0.4384
KGRec	<u>0.8087</u>	<u>0.7608</u>	<u>0.6786</u>	0.2910	0.4804	0.5057	<u>0.8087</u>	0.6544	0.6017
PyRec	0.7948	0.7018	0.6537	<u>0.3510</u>	<u>0.6018</u>	<u>0.6748</u>	0.7948	<u>0.7408</u>	<u>0.7401</u>

MDP的表示能力。当dim从32增加到128时, P , F1和NDCG平均提升11.33%, 11.54%, 9.39%, 表明更高维度能增强模型从知识图谱中提取信息和建模复杂关系的能力, 从而提升预测精度。然而, 当维度进一步增至256时, 3项指标分别下降5.22%, 5.38%, 4.27%, 说明模型在128维已充分学习节点和关系特征, 过高维度不仅难以带来额外收益, 还可能因特征向量过度平滑而导致性能下降。如图7所示, 当 β 从1增至3时, P , F1和NDCG平均提升9.28%, 27.9%, 8.08%。这一趋势源于softplus函数与 β 的交互作用: 在输入为负时, β 越大, 函数值越逼近0, 使负样本得分在减去 $\gamma = 3$ 后被更敏锐

地识别。因而, 正负样本区分度提高, 模型性能显著增强。

5 结束语

本文提出的REKG-MDP模型充分利用HER和非结构化医学补充知识构建MKG, 有效解决了医疗数据的稀疏性和异构性问题。模型同时引入4种关系连接模式, 以生成更精确的关系与实体嵌入, 并通过层次化注意力机制在全局与局部范围内建模多阶邻居关系, 充分考虑尾实体对头实体的重要性。在MIMIC-IV数据集上的实验表明, REKG-MDP在F1, Precision和NDCG等指标上显著优于

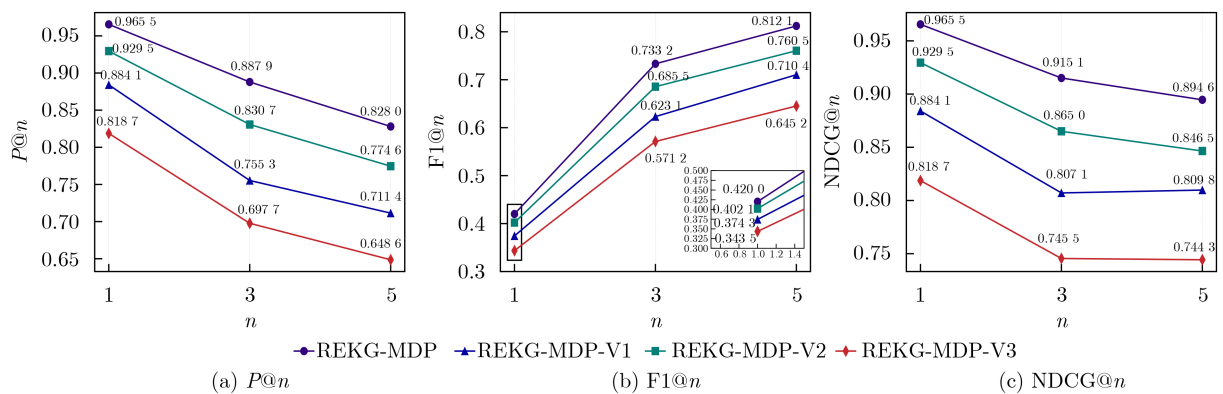


图5 REKG-MDP模型及其3个变体的性能对比图

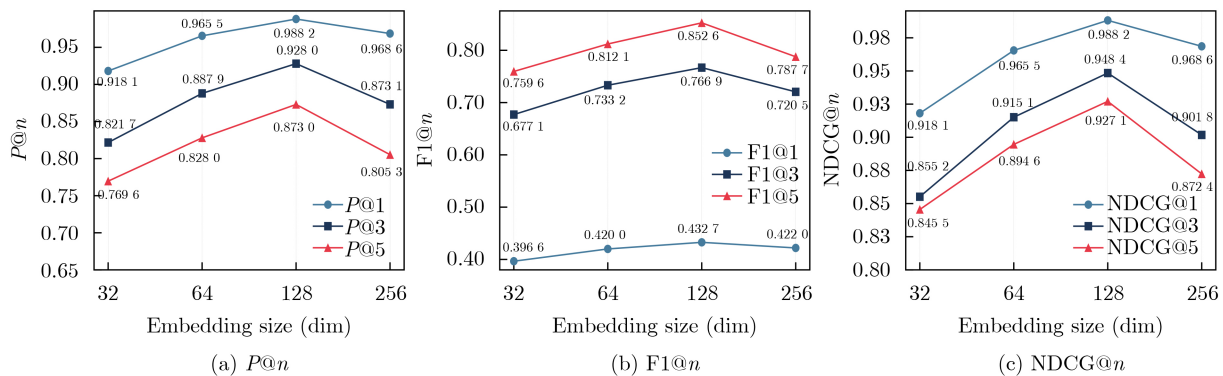


图6 嵌入向量维度对REKG-MDP模型的性能影响图

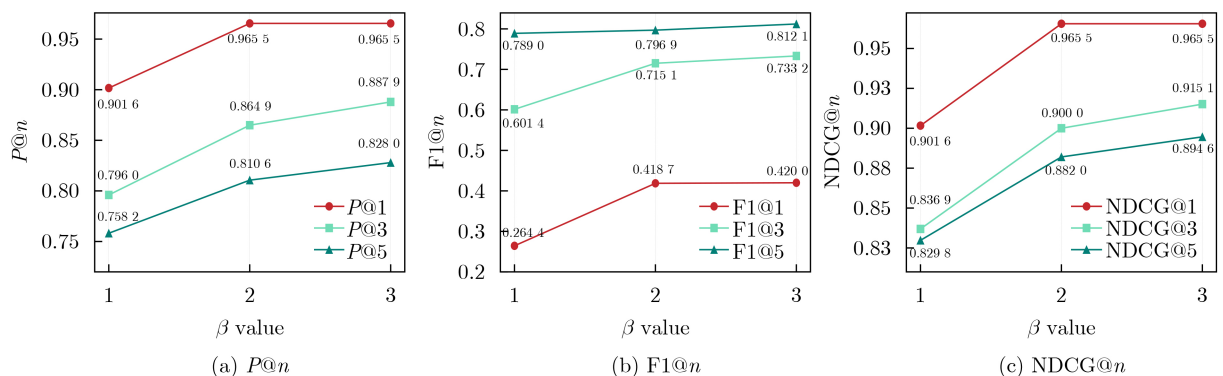


图7 β 对REKG-MDP模型的性能影响图

4种基线方法，展示了在多病症预测任务中的优势。消融实验进一步验证了关系建模与信息传播模块的关键作用。未来的研究将结合时间序列数据与动态知识图谱，实现对潜在疾病风险的精准预测。同时，将进一步推进整合研究，融合多源医学数据与多样化建模方法，以提升框架的鲁棒性与适用性。此外，REKG-MDP可通过灵活调整输入数据推广至其他推荐、预测及知识挖掘场景，具备广泛的应用潜力与实践价值。

参考文献

- [1] American Diabetes Association. Diagnosis and classification of diabetes mellitus[J]. *Diabetes Care*, 2014, 37(S1): S81–S90. doi: [10.2337/dc14-S081](https://doi.org/10.2337/dc14-S081).
- [2] 姚欣卉, 肖洪彬, 卞敬琦, 等. 丹参有效成分在治疗糖尿病及其并发症中的作用机制研究进展[J]. *中国实验方剂学杂志*, 2021, 27(7): 209–218. doi: [10.13422/j.cnki.syfjx.20210401](https://doi.org/10.13422/j.cnki.syfjx.20210401).
YAO Xinhui, XIAO Hongbin, BIAN Jingqi, et al. New progress in mechanism of *Salviae Miltiorrhizae Radix et Rhizoma* in treatment of diabetes and its complications[J]. *Chinese Journal of Experimental Traditional Medical Formulae*, 2021, 27(7): 209–218. doi: [10.13422/j.cnki.syfjx.20210401](https://doi.org/10.13422/j.cnki.syfjx.20210401).
- [3] GUAN Zhouyu, LI Huating, LIU Ruhan, et al. Artificial intelligence in diabetes management: Advancements, opportunities, and challenges[J]. *Cell Reports Medicine*, 2023, 4(10): 101213. doi: [10.1016/j.xcrm.2023.101213](https://doi.org/10.1016/j.xcrm.2023.101213).
- [4] ZHANG Lufang, YU Renyue, CHEN Keya, et al. Enhancing deep vein thrombosis prediction in patients with coronavirus disease 2019 using improved machine learning model[J]. *Computers in Biology and Medicine*, 2024, 173: 108294. doi: [10.1016/j.compbimed.2024.108294](https://doi.org/10.1016/j.compbimed.2024.108294).
- [5] RAHMAN M M, AL-AMIN M, and HOSSAIN J. Machine learning models for chronic kidney disease diagnosis and prediction[J]. *Biomedical Signal Processing and Control*, 2024, 87: 105368. doi: [10.1016/j.bspc.2023.105368](https://doi.org/10.1016/j.bspc.2023.105368).
- [6] ALTHOBAITI T, ALTHOBAITI S, and SELIM M M. An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making[J]. *Alexandria Engineering Journal*, 2024, 94: 311–324. doi: [10.1016/j.aej.2024.03.044](https://doi.org/10.1016/j.aej.2024.03.044).
- [7] AL-SSULAMI A M, ALSORORI R S, AZMI A M, et al. Improving coronary heart disease prediction through machine learning and an innovative data augmentation technique[J]. *Cognitive Computation*, 2023, 15(5): 1687–1702. doi: [10.1007/s12559-023-10151-6](https://doi.org/10.1007/s12559-023-10151-6).
- [8] 金怀平, 薛飞跃, 李振辉, 等. 基于病理图像集成深度学习的胃癌预后预测方法[J]. *电子与信息学报*, 2023, 45(7): 2623–2633. doi: [10.11999/JEIT220655](https://doi.org/10.11999/JEIT220655).
- [9] JIN Huaiping, XUE Feiyue, LI Zhenhui, et al. Prognostic prediction of gastric cancer based on ensemble deep learning of pathological images[J]. *Journal of Electronics & Information Technology*, 2023, 45(7): 2623–2633. doi: [10.11999/JEIT220655](https://doi.org/10.11999/JEIT220655).
- [9] 季薇, 王传瑜, 吴迪, 等. 基于跨语种声学分析的帕金森病检测方法[J]. *电子与信息学报*, 2024, 46(2): 546–554. doi: [10.11999/JEIT230981](https://doi.org/10.11999/JEIT230981).
- [9] JI Wei, WANG Chuanyu, WU Di, et al. Parkinson's disease detection method based on cross-language acoustic analysis[J]. *Journal of Electronics & Information Technology*, 2024, 46(2): 546–554. doi: [10.11999/JEIT230981](https://doi.org/10.11999/JEIT230981).
- [10] GHORBANI M, KAZI A, BAGHSHAH M S, et al. RA-GCN: Graph convolutional network for disease prediction problems with imbalanced data[J]. *Medical Image Analysis*, 2023, 75: 102272. doi: [10.1016/j.media.2021.102272](https://doi.org/10.1016/j.media.2021.102272).
- [11] ZHAO Qing, LI Jianqiang, ZHAO Linna, et al. Knowledge guided feature aggregation for the prediction of chronic obstructive pulmonary disease with Chinese EMRs[J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022, 20(6): 3343–3352. doi: [10.1109/TCBB.2022.3198798](https://doi.org/10.1109/TCBB.2022.3198798).
- [12] PHAM T, TAO Xiaohui, ZHANG Ji, et al. Graph-based multi-label disease prediction model learning from medical data and domain knowledge[J]. *Knowledge-Based Systems*, 2022, 235: 107662. doi: [10.1016/j.knsys.2021.107662](https://doi.org/10.1016/j.knsys.2021.107662).
- [13] QU Zhe, CUI Lizhen, and XU Yonghui. Disease risk prediction via heterogeneous graph attention networks[C]. 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, USA, IEEE, 2022: 3385–3390. doi: [10.1109/BIBM55620.2022.9995491](https://doi.org/10.1109/BIBM55620.2022.9995491).
- [14] LU Chang, HAN Tian, and NING Yue. Context-aware health event prediction via transition functions on dynamic disease graphs[C]. The 36th AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2022: 4567–4574. doi: [10.1609/aaai.v36i4.20380](https://doi.org/10.1609/aaai.v36i4.20380).
- [15] 熊立鹏, 徐修远, 牛颢, 等. 融合nmODE的术后肺部并发症预测模型[J]. *智能系统学报*, 2025, 20(1): 198–205. doi: [10.11992/tis.202401007](https://doi.org/10.11992/tis.202401007).
- [15] XIONG Lipeng, XU Xiuyuan, NIU Hao, et al. Predicting postoperative pulmonary complications after lung surgery using nmODE[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(1): 198–205. doi: [10.11992/tis.202401007](https://doi.org/10.11992/tis.202401007).
- [16] SUN Zhoujian, DONG Wei, SHI Jimlong, et al. Interpretable disease progression prediction based on reinforcement reasoning over a knowledge graph[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, 54(3): 1948–1959. doi: [10.1109/TSMC.2023.3331847](https://doi.org/10.1109/TSMC.2023.3331847).
- [17] CHEN Xiaojun, JIA Shengbin, and XIANG Yang. A review:

- Knowledge reasoning over knowledge graph[J]. *Expert Systems with Applications*, 2020, 141: 112948. doi: [10.1016/j.eswa.2019.112948](https://doi.org/10.1016/j.eswa.2019.112948).
- [18] BORDES A, USUNIER N, GARCIA-DURÁN A, *et al.* Translating embeddings for modeling multi-relational data[C]. The 27th International Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2013: 2787–2795.
- [19] LIN Yankai, LIU Zhiyuan, SUN Maosong, *et al.* Learning entity and relation embeddings for knowledge graph completion[C]. The 29th AAAI Conference on Artificial Intelligence, Austin, USA, 2015: 2181–2187. doi: [10.1609/aaai.v29i1.9491](https://doi.org/10.1609/aaai.v29i1.9491).
- [20] TROUILLON T, WELBL J, RIEDEL S, *et al.* Complex embeddings for simple link prediction[C]. The 33rd International Conference on Machine Learning, New York, USA, 2016: 2071–2080.
- [21] HE Zexue, YAN An, GENTILI A, *et al.* “Nothing abnormal”: Disambiguating medical reports via contrastive knowledge infusion[C]. The 37th AAAI Conference on Artificial Intelligence, Washington, D.C., USA, 2023: 14232–14240. doi: [10.1609/aaai.v37i12.26665](https://doi.org/10.1609/aaai.v37i12.26665).
- [22] SUN Zhiqing, DENG Zhihong, NIE Jianyun, *et al.* Rotate: Knowledge graph embedding by relational rotation in complex space[C]. The 7th International Conference on Learning Representations, New Orleans, USA, 2019: 1–18.
- [23] QIU Jiezhong, TANG Jian, MA Hao, *et al.* DeepInf: Social influence prediction with deep learning[C]. The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom, 2018: 2110–2119. doi: [10.1145/3219819.3220077](https://doi.org/10.1145/3219819.3220077).
- [24] WANG Xiang, HE Xiangnan, CAO Yixin, *et al.* KGAT: Knowledge graph attention network for recommendation[C]. The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, USA, 2019: 950–958. doi: [10.1145/3292500.3330989](https://doi.org/10.1145/3292500.3330989).
- [25] RENDLE S, FREUDENTHALER C, GANTNER Z, *et al.* BPR: Bayesian personalized ranking from implicit feedback[C]. The 25th Conference on Uncertainty in Artificial Intelligence, Montreal, Canada, 2009: 452–461.
- [26] STEFAN N and CUSI K. A global view of the interplay between non-alcoholic fatty liver disease and diabetes[J]. *The Lancet Diabetes & Endocrinology*, 2022, 10(4): 284–296. doi: [10.1016/S2213-8587\(22\)00003-1](https://doi.org/10.1016/S2213-8587(22)00003-1).
- [27] CARRASCO-ZANINI J, PIETZNER M, KOPRULU M, *et al.* Proteomic prediction of diverse incident diseases: A machine learning-guided biomarker discovery study using data from a prospective cohort study[J]. *The Lancet Digital Health*, 2024, 6(7): e470–e479. doi: [10.1016/S2589-7500\(24\)00087-6](https://doi.org/10.1016/S2589-7500(24)00087-6).
- [28] LI Bo, QUAN Haowei, WANG Jiawei, *et al.* Neural library recommendation by embedding project-library knowledge graph[J]. *IEEE Transactions on Software Engineering*, 2024, 50(6): 1620–1638. doi: [10.1109/TSE.2024.3393504](https://doi.org/10.1109/TSE.2024.3393504).
- [29] YANG Yuhao, HUANG Chao, XIA Lianghao, *et al.* Knowledge graph self-supervised rationalization for recommendation[C]. The 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, USA, 2023: 3046–3056. doi: [10.1145/3580305.3599400](https://doi.org/10.1145/3580305.3599400).
- [30] KINGMA D P and BA J. Adam: A method for stochastic optimization[C]. The 3rd International Conference on Learning Representations, San Diego, USA, 2015: 1–15.
- [31] HAMILTON W L, YING R, and LESKOVEC J. Inductive representation learning on large graphs[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 1025–1035.
- 王宇翱: 男, 博士生, 研究方向为人工智能和智能信息处理。
 黄叶琪: 女, 硕士生, 研究方向为医疗人工智能。
 李青远: 男, 硕士生, 研究方向为人工智能和医疗信息处理。
 刘云: 女, 教授, 研究方向为智能医学、医学信息学、临床大数据。
 景慎旗: 男, 高级工程师, 研究方向为医疗信息大数据。
 单涛: 男, 高级工程师, 研究方向为医疗信息大数据。
 郭永安: 男, 教授, 研究方向为智能信息处理。

责任编辑: 马秀强

Integrating Representation Learning and Knowledge Graph Reasoning for Diabetes and Complications Prediction

WANG Yuao^① HUANG Yeqi^① LI Qingyuan^① LIU Yun^② JING Shenqi^②
 SHAN Tao^② GUO Yongan^①

^①(Jiangsu Key Laboratory of Intelligent Information Processing and Communication Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

^②(Department of Information, Jiangsu Province Hospital, Nanjing 210029, China)

Abstract:

Objective Diabetes mellitus and its complications are recognized as major global health challenges, causing severe morbidity, high healthcare costs, and reduced quality of life. Accurate joint prediction of these conditions is essential for early intervention but is hindered by data heterogeneity, sparsity, and complex inter-entity relationships. To address these challenges, a Representation Learning Enhanced Knowledge Graph-based Multi-Disease Prediction (REKG-MDP) model is proposed. Electronic Health Records (EHRs) are integrated with supplementary medical knowledge to construct a comprehensive Medical Knowledge Graph (MKG), and higher-order semantic reasoning combined with relation-aware representation learning is applied to capture complex dependencies and improve predictive accuracy across multiple diabetes-related conditions.

Methods The REKG-MDP framework consists of three modules. First, an MKG is constructed by integrating structured EHR data from the MIMIC-IV dataset with external disease knowledge. Patient-side features include demographics, laboratory indices, and medical history, whereas disease-side attributes cover comorbidities, susceptible populations, etiological factors, and diagnostic criteria. This integration mitigates data sparsity and enriches semantic representation. Second, a relation-aware embedding module captures four relational patterns: symmetric, antisymmetric, inverse, and compositional. These patterns are used to optimize entity and relation embeddings for semantic reasoning. Third, a Hierarchical Attention-based Graph Convolutional Network (HAGCN) aggregates multi-hop neighborhood information. Dynamic attention weights capture both local and global dependencies, and a bidirectional mechanism enhances the modeling of patient–disease interactions.

Results and Discussions Experiments demonstrate that REKG-MDP consistently outperforms four baselines: two machine learning models (DCKD-RF and bSES-AC-RUN-FKNN) and two graph-based models (KGRec and PyRec). Compared with the strongest baseline, REKG-MDP achieves average improvements in P , F1, and NDCG of 19.39%, 19.67%, and 19.39% for single-disease prediction ($n = 1$); 16.71%, 21.83%, and 23.53% for $n = 3$; and 22.01%, 20.34%, and 20.88% for $n = 5$ (Table 4). Ablation studies confirm the contribution of each module. Removing relation-pattern modeling reduces performance metrics by approximately 12%, removing hierarchical attention decreases them by 5~6%, and excluding disease-side knowledge produces the largest decline of up to 20% (Fig. 5). Sensitivity analysis indicates that increasing the embedding dimension from 32 to 128 enhances performance by more than 11%, whereas excessive dimensionality (256) leads to over-smoothing (Fig. 6). Adjusting the β parameter strengthens sample discrimination, improving P , F1, and NDCG by 9.28%, 27.9%, and 8.08%, respectively (Fig. 7).

Conclusions REKG-MDP integrates representation learning with knowledge graph reasoning to enable multi-disease prediction. The main contributions are as follows: (1) integrating heterogeneous EHR data with disease knowledge mitigates data sparsity and enhances semantic representation; (2) modeling diverse relational patterns and applying hierarchical attention improves the capture of higher-order dependencies; and (3) extensive experiments confirm the model's superiority over state-of-the-art baselines, with ablation and sensitivity analyses validating the contribution of each module. Remaining challenges include managing extremely sparse data and ensuring generalization across broader populations. Future research will extend REKG-MDP to model temporal disease progression and additional chronic conditions.

Key words: Joint prediction of multiple diseases; Representation learning; Medical Knowledge Graph (MKG); Graph neural network; Attention mechanism