Stage-Aware Hierarchical Attentive Relational Network for Diagnosis Prediction

Liping Wang, Qiang Liu, *Member, IEEE,* Mengqi Zhang, Yaxuan Hu, Shu Wu, *Senior Member, IEEE,* Liang Wang, *Fellow, IEEE*

Abstract—Recently, Electronic Health Records (EHR) have become valuable for enhancing medical decision making, as well as online disease detection and monitoring. Meanwhile, deep learning-based methods have achieved great success in health risk prediction and diagnosis prediction based on EHR. Nevertheless, deep learning-based models usually require high volumes of data due to the vast amount of parameters. In addition, a considerable proportion of medical codes appear rarely in the EHR data which poses huge difficulties for clinical applications. Hence, some works propose to adopt medical ontologies to enhance the prediction performance and provide interpretable prediction results. However, these medical ontologies are often small-scale and coarse-grained, most of diagnoses and medical concepts are not included, lacking many diagnoses and medical concepts, let alone various relationships between these concepts. To overcome this limitation, we propose to incorporate existing large-scale medical knowledge graphs (KGs) into diagnosis prediction and devise a Stage-aware **H**ierarchical **A**ttentive **R**elational Network, named **HAR**. Specifically, for each visit, a personalized sub-KG is extracted from the existing medical KG, on which HAR conducts relation-specific message passing and hierarchical message aggregation to refine representations of nodes that correspond to medical codes in visits. HAR takes the specific stage of a patient's disease progression into consideration, which participates in the computation of relation-level and node-level attention. Extensive experiments on two public datasets demonstrate the effectiveness of HAR in improving both the visit-level precision and code-level accuracy of the diagnosis prediction task.

Index Terms—diagnosis prediction, electronic health record, knowledge graph, relational graph neural network.

1 Introduction

Electronic Health Records (EHR) [1] have become a pervasive healthcare information technology. EHR data are represented by a temporal sequence of visits, where each visit includes multiple medical codes that represent clinical diagnoses. Even though EHRs were initially designed to improve healthcare efficiency from an operational standpoint, researchers have found secondary use for health services and clinical research [2], [3]. To be more specific, EHR data are employed for such tasks as medical concept extraction [4], [5] and disease prediction [6], [7], [8].

Meanwhile, deep learning models have achieved great success among various domains, including computer vision [9], [10], natural language processing [11], [12], [13], graph neural networks [14], [15], [16] and data mining [17], [18], [19]. Naturally, a lot of deep learning-based methods [20], [21], [22], [23], [24], [25] have been proposed to model EHR data. These deep-learning-based methods not only require less preprocessing and feature engineering but also achieve better performance.

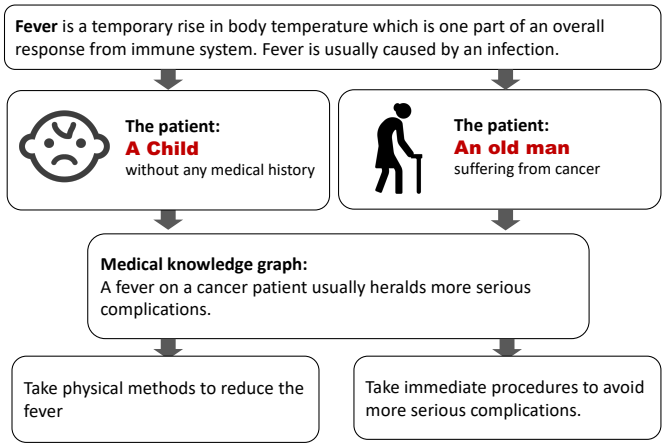

Fig. 1: An example to illustrate the employment of an external medical knowledge graph in treating patients with a fever.

One critical task based on EHR data is predicting future diagnoses using a patient's historical EHR data, known as diagnosis prediction. Recurrent neural networks are adopted to model temporal correlation among EHR sequence data. For instance, Dipole [26] employs an attention-based bidirectional recurrent neural network for learning low-dimensional representations for patient visits, which are then used for future diagnosis prediction.

Nevertheless, deep learning-based EHR models usually require high volumes of data due to the vast number of

- *Liping Wang, Qiang Liu, Mengqi Zhang, Yaxuan Hu, Shu Wu and Liang Wang are with the Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China (E-mail: wangliping2019@ia.ac.cn, qiang.liu@nlpr.ia.ac.cn, mengqi.zhang@cripac.ia.ac.cn, yaxuan.hu@cripac.ia.ac.cn, shu.wu@nlpr.ia.ac.cn, wangliang@nlpr.ia.ac.cn).*

Manuscript received October 31, 2022;
(Corresponding author: Shu Wu)

parameters. Hence, the performance is usually unsatisfactory when the size of the training dataset is limited. In addition, a considerable proportion of medical codes appear infrequently in the EHR data. Then, it is challenging to learn accurate representations for these rare medical codes. In this situation, researchers propose to incorporate external medical knowledge to boost the performance. For example, GRAM [27] proposes to infuse a medical ontology – Clinical Classifications Software (CCS) into deep learning models via neural attention [12]. However, the use of a medical ontology has limited benefits for two main reasons. On the one hand, the scale of most medical ontologies is relatively small. For example, CCS only contains hundreds of concepts. Most diagnoses have no corresponding concepts in the ontology. On the other hand, ontology is essentially a disease classification tree and contains no information about various kinds of relationships between different diseases, let alone, reflecting the progression of diseases. Hence, we propose to incorporate existing large-scale medical knowledge graph, SemMed, into diagnosis prediction.

Even though some existing works [25], [28], [29] also propose to make use of existing large-scale medical knowledge graphs, they fail to take the specific stage of the patient into consideration. As illustrated in Figure 1, taking diagnosis *fever* as an example, it can appear in different stages and reflect different degree of severity. Encountering a patient with a fever, an experienced doctor may take physical methods to reduce the fever if the patient is a child without any medical history. But if the patient is an old man suffering from cancer, the doctor must take immediate procedures to avoid more serious complications that the fever heralds. Existing large-scale medical knowledge graph contains a large amount of information covering different stages of disease. Only a small part of the knowledge applies to the patient at a specific stage. Blind application of the medical KG may result in adverse outcomes. Therefore, it is essential to take the specific stage of the patient into consideration when incorporating existing medical knowledge graph into diagnosis prediction or health risk prediction.

To tackle all the aforementioned problems and challenges, we propose a Stage-Aware **H**ierarchical **A**ttention **R**elational Network, named **HAR**, for diagnosis prediction task in this paper. HAR consists of four components: stage-aware relation-level attention, stage-aware node-level attention, relation-specific message passing and hierarchical message aggregation. Our model is designed as a general-purpose plug-in module, which can be built on all kinds of temporal prediction models. Specifically, for each visit of a patient, we extract a personalized sub knowledge graph from the existing large-scale medical knowledge graph – SemMed[1]. Compared with the raw large-scale knowledge graph, personalized sub-knowledge graph avoids message propagation between nodes not related to the patient and decreases the difficulty of learning significantly. On the extracted personalized sub-knowledge graph, HAR conducts relation specific message passing and hierarchical message aggregation based on both relation-level and node-level attention. Compared with existing works that also make use of medical KGs, HAR takes the specific stage of the patient in disease progression into consideration while calculating the relation-level and node-level attention coefficients. Finally, the obtained refined representations of nodes which have corresponding medical codes in the original EHR data are fed into downstream existing prediction models. Prediction models pass their hidden states back to HAR to represent the current stage of the patient.

Applying GNN to knowledge enhanced diagnosis prediction, while theoretically feasible, faces lots of difficulties in practice and thus requires novel model design. Since there are multiple kinds of nodes and relations in the knowledge graph, we propose specialized network structures that can handle heterogeneous graph input. Even after extracting the subgraph, there are still a large number of nodes and edges in the sub-knowledge graph. Therefore, we design sophisticated attention mechanisms considering current stage of patients to assign different weights in message propagation and aggregation. The main contributions of this work can be summarized as follows:

- We incorporate a large-scale medical knowledge graph (KG) – SemMed into diagnosis prediction and propose a hierarchical stage-aware attention mechanism which extracts informative knowledge from the KG to assist the prediction task.
- We propose HAR, a general-purposed framework for diagnosis prediction which can be built on various temporal prediction models.
- We conduct extensive experiments on two public benchmark datasets to verify the effectiveness of HAR framework.

2 RELATED WORK

In this section, we first review some recent works about health risk prediction and diagnosis prediction based on EHR data. We also provide a brief introduction about recent progress in multi-relation graph neural networks which serve as a useful and powerful tool in enhancing prediction models with external medical knowledge graph.

2.1 Deep Learning based Methods for Diagnosis Prediction

EHR data contain rich historical health information of patients. Building powerful health risk prediction models based on EHR data paves the way for web-enabled personalized health care applications. Recently, deep learning techniques, including convolutional neural networks (CNNs), recurrent neural networks (RNNs) and graph neural networks (GNNs) have achieved great success in various applications among multiple domains, including health risk prediction and diagnosis prediction based on EHR data. In viewing that EHR data exist in temporal sequential form, it is natural to adopt RNNs or LSTMs to model disease progression in time dimension. In comparison, CNNs are adopted to capture local dependence in EHR data.

In Dipole [26], bidirectional recurrent neural networks are employed to remember all the information of both the past visits and the future visits, and three attention mechanisms are introduced to measure the influence of different visits for the prediction. RETAIN [30] develops a

1. https://skr3.nlm.nih.gov/SemMed

reverse time attention model for EHR data which achieves high accuracy while remaining clinically interpretable. Its two-level neural attention detects influential past visits and significant clinical variables within those visits (e.g. key diagnoses).

Another line of works try to model disease progression by taking time intervals into consideration. For example, T-LSTM [31] proposes a novel LSTM [11] unit called Time-Aware LSTM (T-LSTM) to handle irregular time intervals in longitudinal patient records. StageNet [32] integrates inter-visit time information into LSTM cell states to capture the stage variation of patients' health conditions.

TABLE 1: Some Important Notations Used in This Paper

Variable	Description
C	The set of all the unique medical codes
V_i	EHR sequence for the i-th patient
T_i	The length of EHR sequence for the i-th patient
v_t	Medical codes appeared in the t-th visit
N_{v_t}	The set of nodes corresponding to the codes in v_t
$\mathbf{x}_t$	$\|C\|$-length multi-hot vector to represent v_t
$\mathbf{X}_i$	0-1 valued matrix that encodes all the EHR data
G	External medical knowledge graph
g_t	Extracted sub-graph from G according to v_t
f_t	Hidden state of prediction model after the t-th timestep

2.2 Knowledge-Enhanced Methods for Diagnosis Prediction

Deep learning based methods for EHR often require vast sample size to achieve satisfactory performance. In addition, some diagnoses of rare diseases appear in the EHR data much infrequently, making it even harder for accurate prediction for them. To address these problems and learn a robust prediction model, researchers propose to incorporate existing medical knowledge.

As an instance, GRAM [27] infuses information from a medical ontology DAG (Directed acyclic graph) [33] – CCS (Clinical Classifications Software) [2] into deep learning models via neural attention. GRAM can learn accurate and interpretable representations for medical concepts and show significant improvement in the prediction performance, especially on low-frequency diseases and small datasets. HAP [34] adopts the same medical ontology DAG with GRAM [27], but hierarchically propagates attention across the entire ontology structure with two rounds of knowledge propagation. Nevertheless, in both GRAM and HAP, medical ontology information is only used when learning code representations which implicitly affects the final predictions. Hence, Ma et al. [35] propose KAME which directly exploits medical knowledge in the whole prediction process, i.e. learning code representations, generating visit embeddings and making predictions. However, compared with our HAR, the ignorance of considering patient's specific stage limits the performance improvement brought by knowledge graphs.

In viewing that the domain knowledge introduced by GRAM, HAP and KAME are only coarse-grained division of medical concepts where causal relationships are not included, KnowRisk [28] and DG-RNN [29] incorporate a more powerful and larger scaled knowledge graph KnowLife [36][3] to enrich the information extracted from insufficient inputs and guide the prediction. And they propose sophisticated knowledge graph attention to obtain the latent information from embeddings of the input events in the knowledge graph.

2.3 Knowledge Graph and Graph Neural Networks

Knowledge graphs [37] reflect structural relations between entities in the real world which pave the way for cognition and intelligence. A knowledge graph consists of entities and relationships between entities. From the perspective of graph, a knowledge graph could be viewed as a multi-relation graphs in which there are multiple kinds of edges.

To model knowledge graph data, researchers propose various translational models such as TransE [38], TransR [39] and TransH [40]. In order to better capture relationships between entities, some studies [41], [42], [43] propose to represent entities and relationships in complex space instead of real-valued spaces. Each representation are split into real part and imaginary part.

Graph neural networks (GNNs) aims to apply deep neural networks to graph-structured data [44], [45]. Some works generalize convolution operations to graph domain, i.e. Graph Convolution Networks (GCNs) [14], [15], [46], [47]. ChebNet [46] defines graph convolution operator in the Fourier domain and approximates the filter through Chebshev polynomials of the diagonal matrix of eigenvalues. Further, GCN [14] adopts first order Chebshev approximation and bridge the gap between spatial and spectral based graph convolution network approaches. GraphSAGE [47] samples a fixed number of neighbors and employs several aggregation functions. GAT [15] introduces attention mechanisms into GNN and adopts attention coefficients as information aggregation weights.

Later, a lot of works [48], [49], [50], [51] attempt to extend graph neural networks for modeling multi-relational or heterogeneous graphs, especially knowledge graphs. For example, Schlichtkrull et al. [52] propose RGCN (relational graph convolutional networks) to model knowledge graphs which adopts a distinct linear transformation for each kind of relation. Wang et al. [53] propose an extension of graph attention networks (GAT) [15] by maintaining different weights for different pre-defined meta-paths. Messages are passed and aggregated through these meta-paths to convey high-level semantic information. To tackle the problem that weights for relation type with insufficient occurrences cannot be learned accurately, instead of parameterizing each type of edges, HGT [54] defines heterogeneous mutual attention according to the edge types and types of source and target nodes.

3 PRELIMINARY

In this section, we mainly provide some background knowledge about EHR data and external medical knowledge graphs. Basic notations are summarized for easier understanding. Finally, we formulate the diagnosis prediction task.

2. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp
3. http://knowlife.mpi-inf.mpg.de/

3.1 Electronic Health Records

Electronic Health Records (EHR) is a special kind of data which consist of medical history of a patient and provide rich information for the prediction of the future health status of the patient. Even though EHR data were initially designed for improving healthcare efficiency from an operational standpoint, researchers have found secondary use for health services and clinical research such as medical concept extraction and disease prediction. For each visit to the hospital of a specific patient, the diagnoses appeared are recored as medical codes in a pre-defined system such as ICD[4] (International Classification of Diseases) or CUI [5] (Concept Unique Identifiers). In this paper, we mainly focus on diagnosis prediction based on EHR data.

3.2 External Medical Knowledge Graph

External medical knowledge graphs such as KnowLife [36] and SemMed [55] consist of triplets extracted from publications in medical domain. These triplets provide rich information about disease progression and can improve prediction accuracy. Each triplet consists of a head entity, a type of relation and a tail entity. The head and tail entities are concepts in the medical area. Part of those entities appear in EHR data as medical codes which are usually diagnoses given by doctors. The relation type reflects the relationship between the two concepts. For instance, triplet <*Heart Valve Disorder*, *Causes*, *Heart Failure*> reflects that heart valve disorder is the cause of heart failure. With this triplet, model can make prediction of heart failure if heart valve disorder has appeared on the patient.

3.3 Basic Notations

In this paper, all the unique medical codes from EHR data are denoted as $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{|C|} \in C$, where C is the set consisting of all the medical codes. For the i-th patient, the EHR data are denoted as $V_i = \{v_1, v_2, \ldots, v_{T_i}\}$, in which T_i is the total number of visits. Visit v_j is a subset of C, representing medical codes appeared in the j-th visit. For the convenience of calculation, v_j can also be represented as a $|C|$-length multi-hot vector $\mathbf{x}_j$, where each element is zero or one, representing each medical code appears or not respectively. By stacking those multi-hot vectors, we reach a 0-1 valued matrix $\mathbf{X}_i \in \{0,1\}^{T_i \times |C|}$ to represent all the historical EHR data for the i-th patient. Denote the adopted external knowledge graph as G, for visit v_t, the personalized sub-graph extracted from G is g_t. We summarize some important notations used in this paper in Table 1.

3.4 Diagnosis Prediction Task

Diagnosis prediction is one of the most important tasks in health care area which aims to predict potential diagnoses according to historical EHR data. Here, we give the formulation based on notations provided before. For a specific patient, denote his or her EHR data for T consecutive visits as $\mathbf{X} \in \{0,1\}^{T \times |C|}$, the goal is to tell which diagnosis is likely to appear in the next visit, i.e. the value of $\mathbf{x}_{T+1}$. Hence, clinical doctors could intervene in advance for better health service. In addition, interpretability is of great importance in diagnosis prediction scenario. Accurate attribution analysis enable doctors figure out causes in an efficient manner.

4. https://www.cdc.gov/nchs/icd/icd9.htm
5. https://www.nlm.nih.gov

4 METHODOLOGY

In this section, we first provide an overview of our model. After that, the architecture and function of each component are explained in detail. It is worthy note that HAR is a general-purposed plug-in module, it should be combined with a downstream temporal prediction model. Then, joint end-to-end training of HAR and existing prediction models is described. Finally, model interpretation for prediction results is provided.

4.1 Overall Architecture of HAR

As illustrated in Figure 2, HAR is a flexible framework that can be combined with any existing prediction model. For visit v_t, HAR first adopts a sub-graph sampler to extract a personalized knowledge graph g_t from the external knowledge graph G. In sub-graph g_t, medical codes in visit v_t are represented as nodes constituting a node set N_{v_t}. In addition, nodes with less than k-hop distance to those nodes in N_{v_t} are also included in sub-graph g_t. Nodes in N_{v_t} represent symptoms that have already appeared while the neighboring nodes reflect the potential development trend of the disease.

In order to make full use of the valuable information, we decide to adopt message passing and aggregation framework. However, different from general homogeneous graphs, the personalized knowledge graph is a multi-relational graph where edges between nodes have different types. In addition, a medical knowledge graph contains vast volume of information covering different stages of disease, for a patient in a specific stage of a disease, only a small part of the knowledge graph is informative for the diagnosis prediction. Hence, instead of direct adopting multi-relational graph neural networks on heterogeneous (multi-relational) sub-graph g_t, we devise a stage-aware hierarchical attention mechanism to capture relation-level and node-level attention simultaneously. Specifically, downstream temporal prediction models feed their hidden vectors to HAR as patient stage indiactor which enables discriminative adoption of the knowledge graph. After that, hierarchical message aggregation aggregates information from neighborhood and update node representations. Finally, refined embeddings for diagnoses are passed to the downstream existing temporal predictor.

4.2 Personalized Graph Extraction

Even though there are lots of sophisticated choices for sub-graph samplers, we find direct extraction of a k-hop sub-graph of N_{v_t} from G works well, due to the relative small size of N_{v_t}.

$$g_t = k\text{-hop sub-graph}(G, N_{v_t}). \tag{1}$$

As for the specific value of k, we tried different values experimentally, and empirically we find $k = 1$ or 2 works relatively well and larger values may result in computation burden and even negative effect to the performance.

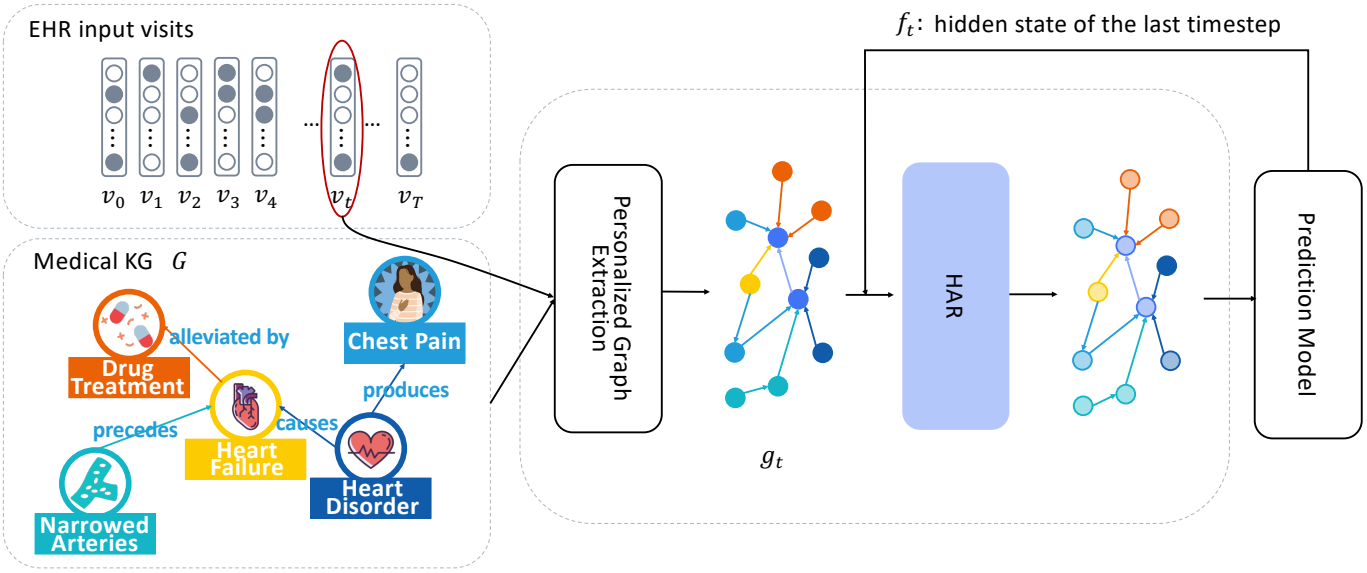

Fig. 2: Data flow for joint-training of HAR and existing prediction models. For visit v_t, a personalized sub-graph g_t is extracted from existing medical knowledge graph G. Then HAR conducts message passing and aggregation on g_t to update embeddings of nodes that correspond to codes in v_t. Finally, refined embeddings are fed to existing prediction model to reach final prediction results. Meanwhile, vector f_t reflecting the current stage of the patient are returned back to HAR.

4.3 Medical Code Embedding

An important step is to convert discrete medical codes to reasonable and learnable representations. There are some advanced strategies for medical code embedding, for example, in KnowRisk [28], GRAM [27] is adopted. However, in HAR framework, we simply employ a parameter embedding matrix $\mathbf{E} \in \mathbb{R}^{|C| \times d}$, where each row encodes a medical code. Parameter matrix $\mathbf{E}$ is learned automatically in an end-to-end manner.

4.4 Stage-Aware Hierarchical Attentive Relational Network

As Figure 3 shows, each HAR layer consists of four components: (1) Stage-Aware Multi-Relational Attention, (2) Stage-Aware Node-Level Attention, (3) Relation-Specific Message Passing and (4) Hierarchical Message Aggregation. Multi-head technique [15] is adopted in the first two components. For the sake of brevity, only one head is assumed in the following text.

4.4.1 Stage-Aware Relation-Level Attention

During the development of a disease, different relation types in the medical knowledge graph may contribute unequally during different stages. For example, in the early stage, causes of a specific diagnosis may reveal more specific and detailed information, which determines the trend of development. However, in the late stage of a disease, *cause* relation contributes less since diagnosis itself is evident enough. Instead, *complication* relation in the medical knowledge graph should be attached more importance because serious complications are fatal.

To measure the weight of different relation types, we first assign an embedding vector r_i for the i-th relation. Stage-aware multi-relational attention module operates on these relation embeddings and the hidden state vector f_{t-1}. In order to obtain sufficient expressive power, following GAT [15], we adopt a linear layer to transform these embeddings and hidden state vectors into high-level features. After that, inner product parameterized by w_r, W_r is employed to compute the importance of different relation types as follows:

$$a_{r_i} = w_r^T [W_r r_i || W_r f_{t-1}], \tag{2}$$

where $||$ means vector concatenation and f_{t-1} is hidden state vector returned from downstream prediction models in the last time step. Leaky ReLu activation and Softmax layer are adopted for further normalization among all relation types as follows:

$$\alpha_{r_i} = \frac{\exp(\text{LeakyReLu}(a_{r_i}))}{\sum_{j=1}^{|\mathcal{R}|} \exp(\text{LeakyReLu}(a_{r_j}))}, \tag{3}$$

in which $\mathcal{R}$ is the set of all the relations in the KG.

4.4.2 Stage-Aware Node-Level Attention

In the medical knowledge graph, a concept usually has dozens of neighbors most of which provide little help in diagnosis prediction task. As a result, neighboring nodes should be attached importance to different degree in message passing and aggregation. Three factors should be take into consideration in calculating the attention for a target node: (1) the representation of the node itself, (2) representations for neighbor nodes, (3) current stage of the patient. To take all the three factors into consideration, we design a stage-aware node-level attention mechanism which operates on source node embedding, target node embedding and state vector together.

Note that source node embedding, target node embedding and state vector may lie in different feature space and have different distribution, so the first step is to map them

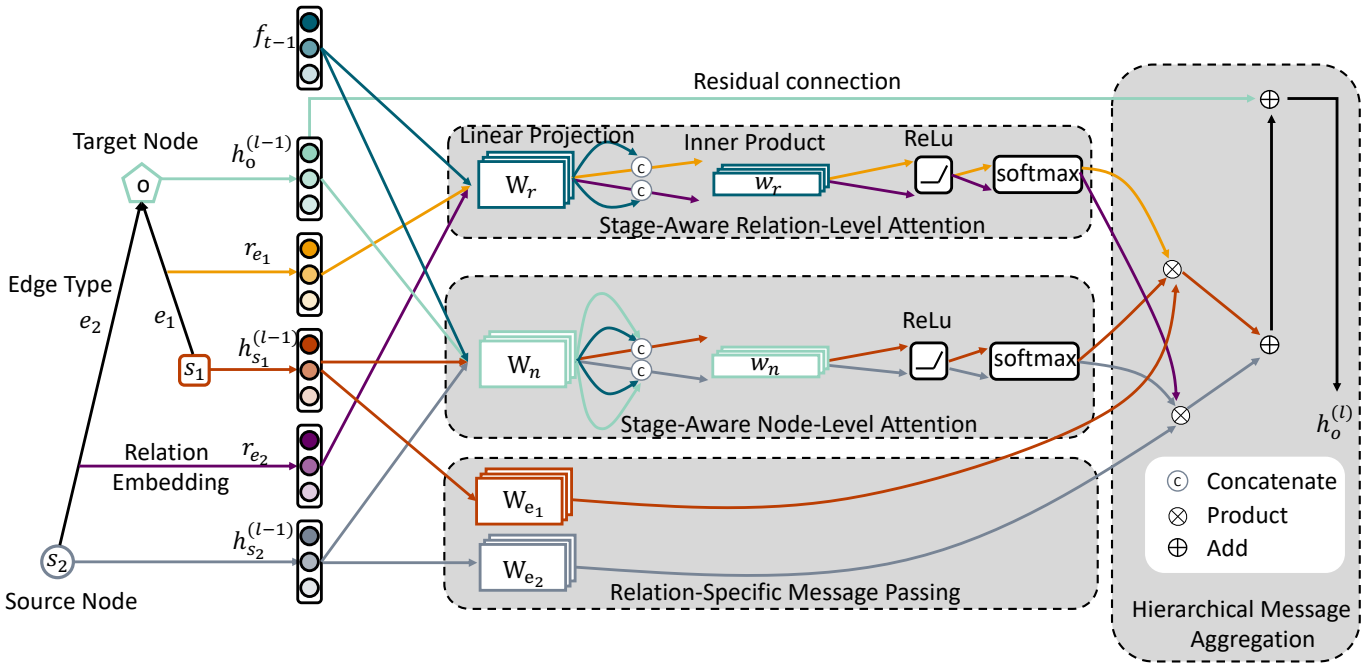

Fig. 3: Illustration of the computation process of both visit-level and code-level attention, relation-specific message passing and hierarchical aggregation between node o and s_1, s_2. Hidden state f_{t-1} returned by the prediction model participates in the computation of attention at two levels.

into same feature space. To maximize parameter sharing while maintaining feature space mapping, we propose to parameterize feature space transformation through a weight matrix W_n. Then transformed embeddings are concatenated together and experience the following computation process similar to relation-level attention:

$$a_n(o, s_i) = w_n^T[W_n h_o || W_n h_{s_i} || W_n f_{t-1}], \tag{4}$$

in which h_o, h_{s_i} represent embeddings for node target node o and source node s_i.

$$\alpha_n(o, s_i) = \frac{\exp(\text{LeakyReLu}(a_n(o, s_i)))}{\sum_{s_j \in \mathcal{N}_o} \exp(\text{LeakyReLu}(a_n(o, s_j)))} \tag{5}$$

4.4.3 Relation-Specific Message Passing

In parallel with the computation of relation-level attention and node-level attention, information is passed from source nodes to target nodes. Inspired by RGCN [52], we introduce relation-specific transformations, i.e., depending on the type of edges. This could be seen as an extension of regular graph convolution:

$$msg(o, s_i) = W_{e_i} h_{s_i}, \tag{6}$$

in which, e_i is the relation type of edge between s_i and o.

4.4.4 Hierarchical Message Aggregation

With stage-aware relation-level and node-level attention calculated and multi-relational message passed, the importance of different relations and neighbor nodes are measured. Based on the two-level attention mechanisms, we design a hierarchical message aggregation module:

$$\widetilde{h}_o^{(l)} = \sum_{r \in \mathcal{R}} \alpha_r \cdot \left(\sum_{s_i \in \mathcal{N}_o^r} \alpha_n(o, s_i) \cdot msg(o, s_i) \right), \tag{7}$$

where $\mathcal{N}_o^r$ denotes the set of neighbor indices of node o under relation $r \in \mathcal{R}$. Two levels of summation in Equation 7 reflects the hierarchy of aggregation. The inner summation reflects message aggregation from nodes under each relation. The outer summation reflects message aggregation among multiple relations.

To ensure that the representation of a target node at layer $l+1$ can also be informed by the corresponding representation at the l-th layer, a self-loop connection is added. Hence, the refined node representation for target node o becomes

$$h_o^{(l)} = \lambda h_o^{(l-1)} + (1-\lambda) \sum_{r \in \mathcal{R}} \alpha_r \cdot \left(\sum_{s_i \in \mathcal{N}_o^r} \alpha_n(o, s_i) \cdot msg(o, s_i) \right), \tag{8}$$

where hyper-parameter λ controls the weight of self-loop message.

Note that in Eq 8, no non-linear activation function is included in node representation updating, since we find that non-linear activation has no positive effect on diagnosis prediction empirically. Similar phenomenon has been observed in recommendation systems [**?**]. This can be explained as: each node in the knowledge graph only has an ID which contains no rich semantics. In this case, performing multiple nonlinear transformations will not contribute to learn better features; even worse, it may add difficulties to the optimization.

4.5 Diagnosis Prediction

As mentioned before, for each visit, HAR refines node embeddings of sub-graph extracted from knowledge graph according to historical visits. Since existing temporal prediction models cannot accept multiple node embeddings as

Algorithm 1: The HAR framework (forward propagation)

Input : EHR visists $v_1, v_2, \ldots, v_{t-1}$
Medical knowledge graph G
Sub-graph order k
HAR layer number L
Downstream Temporal Prediction Model P
Output: The next diagnosis predictions $\widetilde{v}_t$

// Personalized Graph Extraction.
$g_t \leftarrow k$-order subgraph from the neighborhood of N_{v_t} in G
// The update of node embeddings by HAR.
for $l \in [1:L]$ **do**
  Calculate relation-level attention coefficients a_{r_i} according to Equation 2 and 3.
  Calculate node-level attention coefficients a_{r_i} according to Equation 4 and 5.
  Obtain relation-specific message $msg(o, s_i)$ between node s_i and o according to Equation 6
  Conduct hierarchical message aggregation following Equation 7.
  Update the representation of node o following Equation 8.
// The prediction of next item.
Prediction model P feeds state vector f_t back to HAR.
$\hat{\mathbf{x}}_\mathbf{t} \leftarrow P(H^{(1)}, \ldots, H^{(t)})$

input, we adopt a sum readout-function to obtain a graph-level representation as follows:

$$h_{g_t} = \sum_{u \in N_{v_t}} h_u^{(L)}, \tag{9}$$

in which N_{v_t} is the set of nodes corresponding to medical codes in visit v_t. Then, graph-level representation $h_{g_1}, h_{g_2}, \ldots, h_{g_T}$ are fed into downstream temporal prediction model P. After that, model P make predictions for next time step. Most predictors contain RNN or similar modules to model disease progression, hence hidden states are returned back to HAR to represent current state of the patient. These hidden states helps HAR absorb useful information from external medical knowledge graph discriminatively.

4.6 End-to-End Training with Existing Prediction Models

HAR model is trained together with downstream prediction model P in an end-to-end manner. Diagnosis prediction task is essentially a multi binary-classification task, hence we adopt cross-entropy loss. The prediction loss for all time steps from the second visit is calculated as follows:

$$\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) = -\frac{1}{T}\sum_{t=2}^{T}(\mathbf{x}_t \log(\hat{\mathbf{x}}_t) + (1-\mathbf{x}_t)\log(1-\hat{\mathbf{x}}_t)). \tag{10}$$

With the above loss function, HAR model and predictor model P are optimized jointly through gradient descent methods such as [56], [57].

TABLE 2: Basic statistics of MIMIC-III and MIMIC-IV datasets. Those ICD-9 codes which have no responding CUIs in SemMed knowledge graph are ignored. For all datasets, we choose patients who made at least two visits.

Dataset	MIMIC-III	MIMIC-IV
Number of Patients	7,177	40,149
Number of Visits	19,203	153,419
Unique ICD-9 codes	1,086	1,271
Average Visits per Petient	2.68	3.82
Maximum Visits per Patient	42	120
Average Medical Codes per Visit	7.48	6.46
Maximum Medical Codes per Visit	31	29

4.7 Interpretation of Prediction

Interpretation of prediction results is of great importance for clinical applications. In HAR, the stage-aware hierarchical attentive relational network could provide explicit interpretations for predictions. As long as the knowledge graph is large and complete enough, diagnoses in the prediction would appear in the extracted sub-graph. Then, the corresponding node would participate the computation of relation-specific message passing and hierarchical message aggregation. Through both node-level and relation-level attention coefficients, we could reveal relationships between different disease symptoms.

5 Experiments

In this section, we first provide details of experimental settings including datasets, the public knowledge graph, baseline methods, evaluation metric and implementation. Then, we provide experimental results and analysis. In addition, we investigate the design of our framework through an ablation study. Finally, a case study is presented for an intuitive understanding of HAR's interpretability. To make our results fully reproducible, source codes have been made public at https://github.com/lipingcoding/HAR/tree/main.

5.1 Experimental Setup

5.1.1 Datasets

In this paper, experiments are conducted on two publicly available EHR datasets: MIMIC-III [58] and MIMIC-IV, which include thousands of patients' health records from ICU. For both datasets, only diagnosis codes are used for prediction task. It's worthy note that medical codes in MIMIC-III and MIMIC-IV are recored in ICD-9 while ones in SemMed knowledge graph are recored in CUI. Hence, we ignore those ICD-9 codes which have no responding CUIs in SemMed. After the filtering, we choose those patients who made at least two visits. We provide basic statistics of datasets in Table 2.

5.1.2 SemMed: A Public Knowledge Graph

In this paper, the medical knowledge graph that we use is SemMed [55] [25] which is a large-scale multi-relational medical knowledge graph with more than 150,000 entities

TABLE 3: Visit-Level Precision@k comparison on MIMIC-III and MIMIC-IV datasets between four existing prediction models and their variants by adding our HAR. Average results for multiple values of k and relative improvement of HAR variants compared with base models are included.

Dataset	MIMIC-III Visit-Level Precision @ k						MIMIC-IV Visit-Level Precision @ k					
Model	10	15	20	25	30	Average	10	15	20	25	30	Avgerage
LSTM	43.15	47.67	52.41	56.57	59.89	51.94	53.27	57.51	61.53	64.74	67.30	60.87
HAR-LSTM	43.75	49.03	54.28	58.42	61.44	53.38 (**2.78%**↑)	54.25	58.85	62.95	66.17	68.84	62.21 (**2.20%**↑)
Dipole	42.54	47.45	52.67	56.80	60.30	51.95	53.07	57.53	61.59	64.91	67.55	60.92
HAR-Dipole	42.04	47.48	52.83	57.17	60.70	52.04(**0.17**%↑)	53.10	57.59	61.82	65.14	67.92	61.11(**0.31**%↑)
RETAIN	44.51	49.39	54.12	58.12	61.29	53.49	54.44	58.80	62.92	66.14	68.78	62.22
HAR-RETAIN	45.41	50.15	55.51	59.51	63.06	54.73 (**2.32%**↑)	54.73	59.32	63.43	66.82	69.48	62.76 (**0.86%**↑)
RAIM	37.00	42.03	47.48	51.61	55.40	46.70	51.77	56.04	60.12	63.41	66.13	59.49
HAR-RAIM	41.07	45.86	51.12	55.26	58.53	50.37 (**7.85%**↑)	53.45	57.92	61.99	65.31	68.01	61.34 (**3.10%**↑)
StageNet	42.96	47.82	52.75	56.98	60.07	52.12	53.43	58.05	62.22	65.55	68.16	61.48
HAR-StageNet	43.81	48.69	53.76	58.14	61.51	53.18 (**2.05%**↑)	53.70	58.22	62.41	65.74	68.45	61.70 (**0.36%**↑)
HiTANet	45.50	49.90	55.25	59.19	62.61	54.49	57.35	61.81	65.96	69.18	71.77	65.21
HAR-HiTANet	46.70	51.66	57.09	61.31	64.86	56.32 (**3.36**↑)	57.91	62.91	67.28	70.65	73.36	66.42(**1.86**↑)

TABLE 4: Code-Level Accuracy@20 on MIMIC-III and MIMIC-IV datasets between four existing prediction models and their variants by adding our HAR. Medical codes are divided to five groups according to their frequencies in the training dataset. For example, 0-20 means the rarest ones while 80-100 means the frequent ones. Each group's results and overall results are reported at the same time.

	MIMIC-III						MIMIC-IV					
Model	0-20	20-40	40-60	60-80	80-100	Overall	0-20	20-40	40-60	60-80	80-100	Overall
LSTM	6.71	24.23	46.31	84.03	96.27	50.59	18.79	35.68	54.90	87.96	97.16	58.89
HAR-LSTM	9.88	25.71	44.79	87.71	96.78	52.03 (**2.85%**↑)	20.91	37.16	55.97	89.82	97.23	60.22 (**2.26%**↑)
Dipole	7.46	24.15	46.00	84.59	96.51	50.82	18.75	35.13	54.36	88.52	97.22	58.78
HAR-Dipole	6.96	19.40	43.79	91.92	98.16	51.03 (**0.41**%↑)	17.75	34.73	54.73	90.51	98.31	59.20 (**0.71**%↑)
RETAIN	9.83	28.29	46.81	84.97	97.73	52.61	20.96	37.55	56.48	89.77	98.15	60.58
HAR-RETAIN	10.84	26.28	45.67	91.21	98.70	53.57 (**1.82%**↑)	21.37	37.05	56.50	91.46	98.53	60.98 (**0.67%**↑)
RAIM	0.88	10.55	35.08	92.92	97.89	46.31	13.83	32.93	55.53	89.45	97.85	57.94
HAR-RAIM	6.51	20.61	40.54	87.32	96.95	49.37 (**6.60%**↑)	19.85	35.61	53.93	89.07	97.32	59.14 (**2.07%**↑)
StageNet	6.63	24.40	45.34	86.12	97.13	50.98	18.45	34.93	55.95	90.14	97.59	59.43
HAR-StageNet	11.36	26.26	45.34	84.41	96.13	51.78 (**1.57%**↑)	19.47	35.46	56.19	89.62	97.47	59.65 (**0.37%**↑)
HiTANet	7.93	24.55	49.16	91.29	98.76	53.40	18.68	40.32	62.95	92.42	99.03	62.74
HAR-HiTANet	11.44	27.84	50.39	89.93	98.38	54.70(**2.43%**↑)	20.07	41.77	62.09	92.46	99.24	63.17(**0.68%**↑)

and 64 types of relation. SemMed consists of triplets extracted from the abstract part of medical publications on Pubmed [6].

5.1.3 Baselines

As we mentioned before, HAR is a general-purposed plug-in module, it can be combined with various temporal prediction models. To validate the effectiveness of the proposed HAR framework, without loss of generality, we choose four baseline models: LSTM, RETAIN [30], Dipole, [26], RAIM [59], StageNet [32], HiTANet [60]. On the one hand, they serve as baseline models, on the other hand, they can be used as base models on which the proposed HAR is built.

- **LSTM**: We adopt the same embedding method as Dipole [26]. Then, the embeddings of each visit are fed into a LSTM [11] layer. After that, all hidden states are added together to obtain a final feature vector. In the end, a linear classifier and Softmax Layer are employed to reach final predictions.
- **Dipole**: Dipole employs bidirectional recurrent neural networks to remember the information of both the past visits and the future visits and introduce three attention mechanisms to measure the influence of different visits.
- **RETAIN**: RETAIN is a competitive prediction model that adopts a two-level neural attention model that detects influential past visits and significant clinical variables with those visits.
- **RAIM**: RAIM introduces an efficient attention mechanism for continuous monitoring data, which is guided by discrete clinical events. With the guided multi-channel attention, high-density multi-channel signals are integrated with discrete events and prove very useful in risk prediction.
- **StageNet**: StageNet is a state-of-the-art model that extract disease stage information from patient data and integrate it into risk prediction.
- **HiTANet**: To leverage time information for risk prediction in a more reasonable way, HiTANet imitates the decision making process of doctors in risk prediction through a hierarchical time-aware attention network.

5.1.4 Evaluation Metric

Following KAME [35], we adopt two metrics to measure the performance of all methods for diagnosis prediction task,

6. https://pubmed.ncbi.nih.gov/

i.e. visit-level precision@k and code-level accuracy@k.

For a visit, its precision@k is defined as the correct ratio among the top-k highest-scoring prediction results:

$$precision@k = \frac{1}{min(k, |C|)} \sum_{i=1}^{|C|} \mathbb{1}_{\mathbf{x}[i]=1 \text{ and } \hat{\mathbf{x}}[i] \in \text{top}-k(\hat{\mathbf{x}})}, \tag{11}$$

where $\mathbb{1}$ is an indicator function and top-$k(\hat{\mathbf{x}})$ means the set of k greatest elements in $\hat{\mathbf{x}}$. The average value of precision@k among all the visits is reported.

For a medical code, its accuracy@k is defined as the ratio of the number that it is correctly predicted to the total number of its occurrences. What we call correctly predicted here means that the code are predicted with a top-k highest score, otherwise falsely. On the one hand, we report the average value of accuracy among all the medical codes. On the other hand, we sort the medical codes by their frequencies in the training dataset in non-decreasing order, and then divide them into five different groups. For example, 0-20 means the rarest ones while 80-100 means the most frequent ones. Average code-level accuracy in different groups reflects the prediction performance for codes with varying frequencies.

5.1.5 Implementation Details

In this paper, all the baselines and our models are implemented with PyTorch [7] [61] and DGL [8] [62]. Both MIMIC-III and MIMIC-IV datasets are randomly divided into training, validation and testing sets in a 0.7:0.1:0.2 ratio. Embedding size d is set 64 for all approaches. The same dropout strategy with a 0.5 drop rate is applied to all the methods. All methods are trained with Adam optimizer [56] with a mini-batch of 16 patients. Learning rate is fixed at 0.0005 for all methods.

5.2 Experimental Results

5.2.1 Performance Comparison

For each baseline method, we first test the vanilla model. Then, we build on the vanilla model and add our HAR. Comparison results at visit level are shown in Table 3, in which, precision results for different values of k are reported. For code level comparison, we report accuracy@20 results in Table 4. More specifically, in addition to the overall performance in code-level accuracy, we also report the results for each group which are obtained by dividing the medical codes according to the percentile of their frequencies in the training dataset. For example, 0-20 are the rarest diagnoses while 80-100 represent the most common ones. As Table 3 and 4 illustrate, compared with vanilla models, HAR improves the prediction performance on both visit and code level. Especially, the improvement is more significant for codes that appear less frequently, reflecting that HAR successfully incorporate outside medical knowledge to improve prediction accuracy of rare diseases or diagnoses. This also reveals the necessity of using outside medical knowledge in EHR domain.

To more fully assess the performance of the model, we compare HAR with other methods on MIMIC-III dataset which also utilize external medical knowledge and report results in Table 5. For fairness of comparison, all methods have access to the same knowledge graph, i.e. SemMed. Specifically, GCN and RGCN operate on the extracted sub-knowlege graph as HAR and are built on the same LSTM temporal prediction model with HAR. DGRNN [29] and MedPath [25] are the latest methods adopting external medical knowledge graphs for health prediction. From Table 5, we can observe that HAR obtains the best performance which reflects the effectiveness in utilizing external knowledge for diagnosis prediction.

TABLE 5: The comparison between HAR and methods utilizing medical knowledge on MIMIC-III dataset.

Model	Prec.@10	Acc.@10	Prec.@20	Acc.@20
GCN	34.45	30.77	45.46	44.27
RGCN	43.28	38.18	53.03	51.36
DGRNN	39.04	33.62	48.71	45.55
MedPath	43.04	37.87	53.35	51.81
HAR	**43.75**	**38.40**	**54.28**	**52.03**

5.2.2 Ablation Study

After validating the effectiveness of HAR framework, we want to figure out how the stage-aware hierarchical attentive relational network design can improve the performance. Hence, we investigate our framework design by an ablation study. HAR is essentially a sort of multi-relation graph neural network, hence we compare the performance of HAR with two representative models, i.e. GCN [63] and RGCN [52]. Meanwhile, we compare the performance with MedPath [25], which is also a pluggable module and incorporates outside medical knowledge. To validate the reasonableness of the stage-aware hierarchical attention mechanism, we evaluate three model variants: HAR-R, HAR-N, HAR-no-state. In HAR-R, node-level attention is omitted and relation-level attention is kept, while in HAR-N, the situation is just the opposite. For HAR-no-state model variant, stage vector returned by downstream predictors does not take part in the computation of attention. We build GCN, RGCN, MedPath and the three HAR variants on different prediction models, and report both the visit-level precision and code-level accuracy on MIMIC-III dataset in Table 6.

From Table 6, we can observe that for each vanilla model, HAR variant obtains the best performance. After removing state information from HAR, there will be an obvious accuracy decrease. This reflects the necessity of take state information into consideration for diagnosis prediction. Similar phenomenon happens when relation-level or node-level attention mechanisms is removed.

5.2.3 Sensitivity Analysis

Since the hyper-parameter λ controls the degree that outside medical knowledge participates, we conduct sensitivity analysis on the impact of different values of λ on the performance. As illustrated in Figure 4, when $\lambda = 1.0$, medical knowledge is not involved and HAR + base model degenerates to the base model. In this situation, there is a sharp decline in the performance, reflecting the effectiveness of the incorporation of existing medical knowledge. We can

7. https://pytorch.org/
8. https://www.dgl.ai/

TABLE 6: Visit-level precision@20 and code-level accuracy@20 for HAR and its variants built on each existing prediction model on MIMIC-III dataset.

Models	LSTM		RETAIN		RAIM		StageNet	
Metric	Prec.@20	Acc.@20	Prec.@20	Acc.@20	Prec.@20	Acc.@20	Prec.@20	Acc.@20
Vanilla	52.41	50.59	54.12	52.61	47.48	46.31	52.75	50.98
GCN	45.46	44.27	54.00	52.27	49.13	47.40	53.32	51.24
RGCN	53.03	51.36	54.79	53.09	50.87	49.30	50.27	48.61
MedPath	53.35	51.81	55.12	53.42	50.74	49.12	53.03	51.17
HAR-R	53.66	51.86	54.59	52.95	50.80	49.22	45.38	44.16
HAR-N	53.96	51.94	55.26	53.36	50.91	49.18	53.31	51.67
HAR-no-state	53.78	51.79	55.48	53.55	50.69	49.06	53.58	51.46
HAR	**54.28**	**52.03**	**55.51**	**53.57**	**51.12**	**49.37**	**53.76**	**51.78**

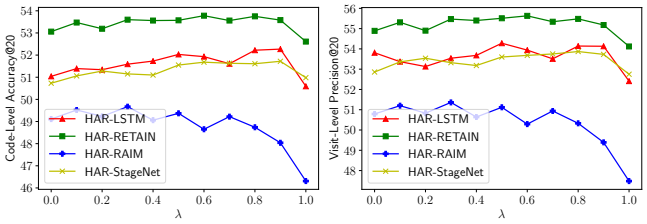

Fig. 4: Code-Level Accuracy@20 and Visit-Level Precision@20 on MIMIC-III dataset with λ varying from 0 to 1.

observe that HAR is relatively robust to the value of λ. In a wide range of value ($[0, 0.9]$) for λ, HAR can bring improvement to different degrees.

5.3 Model Interpretability

In clinical applications, interpretation of prediction results is of great importance. In HAR, the stage-aware hierarchical attentive relational network we propose not only improves the performance, but also provide explicit explanations to interpret prediction results and reveal the relationships between symptoms in disease progression. To verify our claim, we provide a case study on how to interpret the diagnosis prediction results in Figure 5. This case is randomly sampled from the test set and the patient ID is 46836. For easier understanding, we provide codes and their meanings in Table 7. For the last visit, the prediction results are also provided. Take code *038.8* as an example, in the prediction result for v_3, code *038.8* gets the highest prediction score. By searching the personalized sub-KGs, i.e. g_1, g_2, we found that in g_1, there is an edge of *preceded-by* between *038.8* and *486*. What's more, the attention coefficient (the product of relation-level attention and node-level attention) on that edge is the largest one among all the edges with *486* as target. This serves as an explicit interpretation for the prediction.

6 Conclusion

In this paper, we propose a novel stage-aware hierarchical attentive relational network, named HAR, for diagnosis prediction task based on EHR data. Our model is designed as a general-purpose plug-in module that can be built on various prediction models. We incorporate the pre-existing large-scale medical knowledge graph (KG) – SemMed into diagnosis prediction. For each visit, a personalized sub-KG is extracted, on which HAR conducts relation-specific message passing and hierarchical message aggregation. According to the current stage of the patient in the disease progression, HAR assigns different weights to relation types and neighboring nodes based on the current stage of the patient in disease progression. Extensive experimental results on two public benchmark datasets show that our model can improve the performance of existing prediction models in terms of both visit-level precision and code-level accuracy. An ablation study verifies the rationality of the network design and the effectiveness of each component in HAR. In addition, through the case study, we verify that interpretation can be provided by analyzing the attention coefficients generated by HAR.

TABLE 7: ICD-9 codes and their meanings appearing in Figure 5.

Code	Meaning
486	Pneumonia
401.9	Essential hypertension
428.0	Heart failure
424.0	Other diseases of endocardium
427.3	Cardiac dysrhythmias
244.9	Acquired hypothyroidism
578	Gastrointestinal hemorrhage
038.8	Septicaemia
599.0	Urinary tract infection
785.5	Symptoms involving cardiovascular system

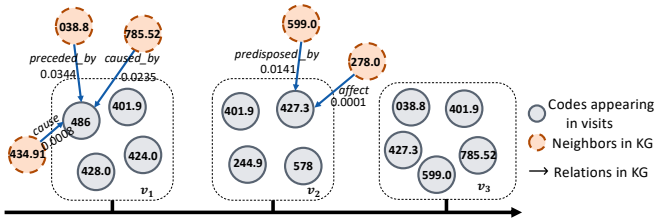

Fig. 5: Case study results for diagnosis prediction to show the participation of existing medical knowledge graph and model interpretability. Attention coefficients extracted in personalized sub-KG according to historical visits reflect their contribution to the prediction.

References

[1] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: A survey of recent advances in deep learning techniques for elec-

tronic health record (ehr) analysis," *IEEE J Biomed. Health Inform.*, pp. 1589—-1604, 2017.

[2] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary use of ehr: data quality issues and informatics opportunities," *Summit on Translational Bioinformatics*, vol. 2010, p. 1, 2010.

[3] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[4] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research," *Yearbook of medical informatics*, vol. 17, no. 01, pp. 128–144, 2008.

[5] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 601–606, 2011.

[6] S. Ebadollahi, J. Sun, D. Gotz, J. Hu, D. Sow, and C. Neti, "Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics," in *AMIA annual symposium proceedings*, vol. 2010. American Medical Informatics Association, 2010, p. 192.

[7] D. Zhao and C. Weng, "Combining pubmed knowledge and ehr data to develop a weighted bayesian network for pancreatic cancer prediction," *Journal of biomedical informatics*, vol. 44, no. 5, pp. 859–868, 2011.

[8] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes," *Journal of clinical epidemiology*, vol. 66, no. 4, pp. 398–407, 2013.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, p. 1735–1780, 1997.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *International Conference on Learning Representations*, 2018.

[16] F. Hu, L. Wang, S. Wu, L. Wang, and T. Tan, "Graph classification by mixture of diverse experts," *arXiv preprint arXiv:2103.15622*, 2021.

[17] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, 2019, pp. 165–174.

[18] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 639–648.

[19] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, and T. Tan, "Session-based recommendation with graph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 346–353.

[20] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports*, vol. 6, no. 1, pp. 1–10, 2016.

[21] A. N. Jagannatha and H. Yu, "Structured prediction models for rnn based sequence labeling in clinical text," in *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, vol. 2016. NIH Public Access, 2016, p. 856.

[22] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua, "Disease inference from health-related questions via sparse deep learning," *IEEE Transactions on knowledge and Data Engineering*, vol. 27, no. 8, pp. 2107–2119, 2015.

[23] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Medical concept representation learning from electronic health records and its application on heart failure prediction," *arXiv preprint arXiv:1602.03686*, 2016.

[24] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: a convolutional net for medical records (2016)," *ArXiv160707519 Cs Stat*.

[25] M. Ye, S. Cui, Y. Wang, J. Luo, C. Xiao, and F. Ma, "Medpath: Augmenting health risk prediction via medical knowledge paths," in *Proceedings of the Web Conference 2021*, 2021, p. 1397–1409.

[26] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2017, p. 1903–1911.

[27] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: Graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, p. 787–795.

[28] X. Zhang, B. Qian, Y. Li, C. Yin, X. Wang, and Q. Zheng, "Knowrisk: An interpretable knowledge-guided model for disease risk prediction," in *2019 IEEE International Conference on Data Mining (ICDM)*, 2019.

[29] C. Yin, R. Zhao, B. Qian, X. Lv, and P. Zhang, "Domain knowledge guided deep learning with electronic health records," in *2019 IEEE International Conference on Data Mining (ICDM)*, 2019, pp. 738–747.

[30] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "Retain: an interpretable predictive model for healthcare using reverse time attention mechanism," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 3512–3520.

[31] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware lstm networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, p. 65–74.

[32] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun, "Stagenet: Stage-aware neural networks for health risk prediction," in *Proceedings of The Web Conference 2020*, 2020, p. 530–540.

[33] K. Thulasiraman and M. N. Swamy, *Graphs: theory and algorithms*. John Wiley & Sons, 2011.

[34] M. Zhang, C. R. King, M. Avidan, and Y. Chen, "Hierarchical attention propagation for healthcare representation learning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, p. 249–256.

[35] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "Kame: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, p. 743–752.

[36] P. Ernst, C. Meng, A. Siu, and G. Weikum, "Knowlife: A knowledge graph for health and life sciences," in *2014 IEEE 30th International Conference on Data Engineering*, 2014.

[37] S. Ji, S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2021.

[38] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in neural information processing systems*, vol. 26, 2013.

[39] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[40] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, no. 1, 2014.

[41] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *International conference on machine learning*. PMLR, 2016, pp. 2071–2080.

[42] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," *arXiv preprint arXiv:1902.10197*, 2019.

[43] S. Zhang, Y. Tay, L. Yao, and Q. Liu, "Quaternion knowledge graph embeddings," *Advances in neural information processing systems*, vol. 32, 2019.
[44] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.
[45] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proceedings. 2005 IEEE international joint conference on neural networks*, vol. 2, no. 2005, 2005, pp. 729–734.
[46] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, vol. 29, 2016.
[47] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.
[48] C. Shang, Y. Tang, J. Huang, J. Bi, X. He, and B. Zhou, "End-to-end structure-aware convolutional networks for knowledge base completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3060–3067.
[49] D. Nathani, J. Chauhan, C. Sharma, and M. Kaul, "Learning attention-based embeddings for relation prediction in knowledge graphs," *arXiv preprint arXiv:1906.01195*, 2019.
[50] S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar, "Composition-based multi-relational graph convolutional networks," *arXiv preprint arXiv:1911.03082*, 2019.
[51] Z. Zhang, F. Zhuang, H. Zhu, Z. Shi, H. Xiong, and Q. He, "Relational graph neural network with hierarchical attention for knowledge graph completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9612–9619.
[52] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *The Semantic Web*, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds., 2018, pp. 593–607.
[53] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *The World Wide Web Conference*, 2019, p. 2022–2032.
[54] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," 2020, p. 2704–2710.
[55] T. C. Rindflesch, H. Kilicoglu, M. Fiszman, G. Rosemblat, and D. Shin, "Semantic medline: An advanced information management application for biomedicine," *Information Service & Use*, vol. 31, p. 15–21, Jan. 2011.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.
[57] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999.
[58] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
[59] Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sun, "Raim: Recurrent attentive and intensive model of multimodal patient monitoring data," in *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, 2018, pp. 2565–2573.
[60] J. Luo, M. Ye, C. Xiao, and F. Ma, "Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 647–656.
[61] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.
[62] M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, T. Xiao, T. He, G. Karypis, J. Li, and Z. Zhang, "Deep graph library: A graph-centric, highly-performant package for graph neural networks," *arXiv preprint arXiv:1909.01315*, 2019.
[63] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

Liping Wang received his B.S. degree from Nankai University, China, in 2019. He is currently pursuing the master degree in Center for Research on Intelligent Perception and Computing (CRIPAC) at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). His current research interests lie on machine learning, knowledge graph, and recommender systems.

Qiang Liu is an Associate Professor with the Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA). He received his PhD degree from CASIA. Currently, his research interests include data mining, recommender systems, text mining, knowledge graph, and graph representation learning. He has published papers in top-tier journals and conferences, such as IEEE TKDE, AAAI, IJCAI, NeurIPS, WWW, SIGIR, CIKM and ICDM.

Mengqi Zhang is currently pursuing the Ph.D. degree in computer application technology with the University of Chinese Academy of Sciences (UCAS), Beijing, China. His research interests include data mining, graph representation learning, and recommender systems. He has published several papers in such fields at international journals and conferences such as TKDE and ICDM.

Yaxuan Hu received the B.S. degree from the School of Information Sciences, Hunan University, China, in 2015. She is currently pursuing the master degree in the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing, China. Her research interests include natural language processing, graph data mining, and reasoning.

Shu Wu received his B.S. degree from Hunan University, China, in 2004, M.S. degree from Xiamen University, China, in 2007, and Ph.D. degree from Department of Computer Science, University of Sherbrooke, Quebec, Canada, all in computer science. He is an Associate Professor with Center for Research on Intelligent Perception and Computing (CRIPAC) at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). His research interests include data mining, information retrieval, and recommendation systems.

Liang Wang received both the BEng and MEng degrees from Anhui University in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2004. Currently, he is a full professor of the Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals such as IEEE TPAMI and IEEE TIP and leading international conferences such as CVPR, ICCV, and ICDM. He is an IEEE and IAPR Fellow.