IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 August 2023, accepted 23 September 2023, date of publication 26 September 2023, date of current version 4 October 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3319502*

RESEARCH ARTICLE

# Enhancing Diagnosis Prediction in Healthcare With Knowledge-Based Recurrent Neural Networks

**HUA SHEN**
College of Artificial Intelligence, Anshan Normal University, Anshan 114007, China
e-mail: huashen.cn@gmail.com

This work was supported in part by the National Natural Science Foundation of China under Grant 61272374 and Grant 61972065, and in part by the Foundation of Department of Education of Liaoning Province under Grant LJKMZ20221812.

**ABSTRACT** The objective of diagnosis prediction involves foreseeing the potential diseases/conditions according to analyzing patients' historical Electronic Health Records (EHRs). The primary challenge in this task is to develop a predictive model that is both sturdy and accurate, while also being interpretable. The most advanced models usually take recurrent neural networks (RNNs) as backbones and then utilize other techniques, such as attention mechanisms, to address this challenge. However, the effectiveness of these models heavily relies on having ample EHR data. Consequently, when the data is insufficient, the performance of these models declines significantly. Recently, graph-based attention models have been proposed to mitigate the issues caused by insufficient data, although they do not fully capitalize on the knowledge present in medical ontologies. To address these problems, **k**nowledge-b**a**sed **r**ecurrent **n**eural network**s** (named KARNS) are introduced, which is an end-to-end, robust, and accurate deep learning-based architecture designed to predict patients' future health information. KARNS explicitly leverages the high-level representations of medical codes within the medical ontologies to enhance the accuracy of predictions. Experimental outcomes demonstrate that the proposed KARNS outperforms existing approaches on three real-world medical datasets. It ensures robustness even with limited training data and learns disease representations that are interpretable.

**INDEX TERMS** Healthcare informatics, diagnosis prediction, deep learning models, recurrent neural networks.

## I. INTRODUCTION

The prediction of patients' future health status using their historical Electronic Healthcare Records (EHRs) has garnered significant attention from healthcare providers and researchers alike [1], [2], [3], [4], [5], [6], [7], [8], and [9]. Specifically, the field of *diagnosis prediction*, which focuses on forecasting future diagnoses based on patients' sequential EHR data, has become a popular yet challenging area of research. The challenges within diagnosis prediction encompass two main aspects: 1) effectively modeling sequential EHR data to capture their unique characteristics, such as high-dimensionality and noise existence, and 2) ensuring reliable predictive performance even in the presence of limited training data.

Various deep learning approaches have been proposed to achieve this objective [10], [11], [12], [13], [14]. One such approach, Med2Vec [14], simulates work embedding techniques [15] to learn embeddings for medical codes in a low-dimensional space, which are used to predict future visits or potential diseases for patients. However, Med2Vec only models the codes in a short visit window and treats each visit as independent. This approach ignores the importance of modeling the sequential nature of EHR data. To capture sequential dependencies within healthcare records, state-of-the-art diagnosis prediction methods commonly employ

The associate editor coordinating the review of this manuscript and approving it for publication was Filbert Juwono.

recurrent neural networks (RNNs) [11], [12], [13], [16], [17], [18]. For instance, the reverse time attention model (RETAIN) [10] utilizes two time-ordered reverse RNNs with attention mechanisms to further boost the prediction performance. Besides, the design of RETAIN makes it output the attention weights of each medical code during the prediction, which increases the interpretability of the diagnosis prediction task. Another RNN-based approach, Dipole, [13] uses a bidirectional RNN (BRNN) with different attention mechanisms, leading to improved prediction accuracy. However, these models often require substantial amounts of training data to ensure reliable predictive performance. Yet, there are cases where certain medical codes associated with rare diseases appear infrequently in EHR data. Consequently, training a robust and accurate predictive model for these rare codes becomes a more challenging yet crucial task.

To address this challenge, the graph-based attention model (GRAM) was introduced [11]. It leverages the International Classification of Diseases (ICD) ontology, which is a tree structure, as extra knowledge to enhance prediction performance and interpretability. Correspondingly, GRAM proposes a graph-based attention mechanism to learn robust representations of medical codes. GRAM exhibits good performance even with limited data availability. However, it does not demonstrate any performance improvement when abundant EHR data is present compared to RNN variants. Furthermore, GRAM solely utilizes ontology information for learning medical code representations, which indirectly influences the prediction outcomes. Thus, *directly incorporating high-level representations of medical codes into the prediction process* can enhance the accuracy of predictive models while preserving the interpretability of medical code representations.

In this paper, a novel deep learning architecture called KARNS is introduced, which utilizes knowledge-based recurrent neural networks for predicting patients' future diagnoses, as depicted in Figure 1. The main idea behind KARNS is to select high-level or general representatives for each medical code from a given medical ontology or knowledge graph, specifically by considering codes in their ancestor set. This approach allows us to create representative vectors for each visit. The knowledge graph used in KARNS can be obtained from reliable sources such as the ICD ontology or the Clinical Classifications Software (CCS).

To learn the embeddings of medical codes and their ancestors, KARNS employs a graph-based attention mechanism. Subsequently, KARNS learns the visit-level representation $\mathbf{v}_t$ using the learned medical code embeddings. To further use the extra knowledge, KARNS utilizes the learned ancestor embeddings to aggregate a knowledge-based representation $\mathbf{q}_t$ for the visit. These vectors are then separately inputted into recurrent neural networks (RNNs) to generate hidden state representations, i.e., $\mathbf{h}_t$ and $\mathbf{k}_t$. $\mathbf{h}_t$ represents the output of the original input visit, while $\mathbf{k}_t$ captures the high-level representation of the visit. The concatenation of these two hidden states is passed through a classification layer to predict what diagnosis codes will appear in the next visit.

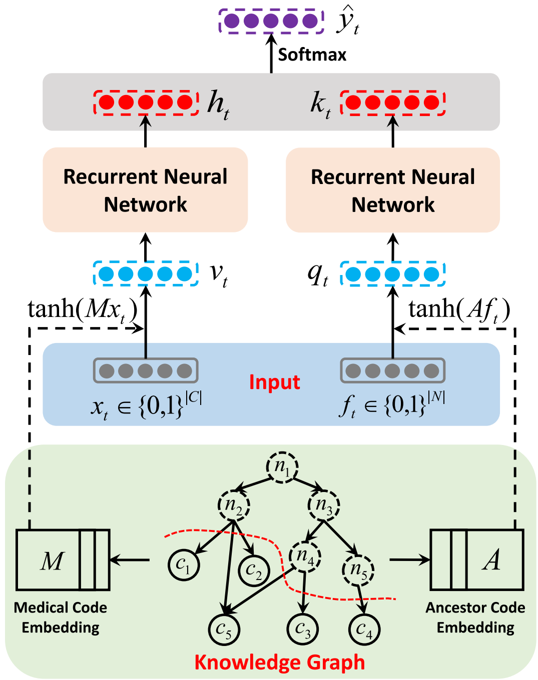

**FIGURE 1.** The proposed KARNS model.

In the experiments, three real-world medical datasets are used to evaluate the efficacy of KARNS, compared with state-of-the-art baselines. The quantitative study demonstrates the effectiveness and robustness of KARNS in scenarios with both sufficient and insufficient data. Furthermore, the qualitative analysis illustrates the interpretability of KARNS by visualizing the generated medical code embeddings.

Overall, the proposed KARNS architecture shows remarkable performance in the diagnosis prediction task, outperforming existing approaches. Through extensive experimentation and analysis, the effectiveness, robustness, and interpretability of KARNS is demonstrated using real-world medical datasets.

The designed model holds significant potential as a crucial component of a smartphone or web application. Once users register on the application, they can upload their Electronic Healthcare Records (EHR) data. The application will automatically encrypt and securely store this data in a database. By analyzing the uploaded data, the application can provide prediction results regarding the user's health status.

Whenever a user inputs new visit information, the application will be triggered to automatically re-predict the user's health status based on both the historical and new EHR data using the well-trained model. To ensure the ongoing effectiveness of the model, it will be regularly retrained when the number of available data reaches a specified threshold.

Equipping the application with the designed model allows users to conveniently monitor their health status in real time. By providing timely predictions, users can take proactive measures to prevent potential diseases and maintain their well-being. The application serves as a valuable tool for individuals to stay informed about their health and take appropriate actions for early disease prevention.

## II. THE PROPOSED MODEL

In this section, some basic notations of medical ontology and EHR data are first introduced, and then preliminary concepts used in the model are described. Finally, the proposed knowledge-based recurrent neural networks KARNS model is presented.

### A. NOTATIONS

In the model design, a medical ontology is used to enhance model performance and interpretability. Let $\mathcal{G}$ represent the medical ontology, a tree structure as shown in Figure 1, which allows us to use a directed acyclic graph (DAG) to represent the hierarchical relationships between various medical concepts. In this tree-based knowledge graph, there are two kinds of medical codes: leaves and their ancestors.

- **Leaves**. $c_i$ $(1 \leq i \leq |\mathcal{C}|)$ is used to represent a leaf node, and all leaf nodes can be represented by a set, $\mathcal{C} = \{c_1, c_2, \cdots, c_{|\mathcal{C}|}\}$, where $|\mathcal{C}|$ represents the total number of unique medical codes.
- **Ancestors**. $n_j$ $(1 \leq j \leq |\mathcal{N}|)$ is used to denote an ancestor node, and the set of ancestor nodes is represented by $\mathcal{N} = \{n_1, n_2, \cdots, n_{|\mathcal{N}|}\}$, where $|\mathcal{N}|$ represents the total number of ancestor codes in $\mathcal{G}$.

For a certain leaf node $c_i$, it has a set of ancestors, which is denoted as $\phi(c_i) \subset \mathcal{N}$. Take $c_1$ in Figure 1 as an example, its ancestor set is $\phi(c_1) = \{n_1, n_2\}$. All of these nodes in $\phi(c_i)$ can be seen as the high-level or general representatives of $c_i$. The closer they are to $c_i$, the stronger the ability of representation for $c_i$ is. However, the high-level representative of $c_i$ should be neither too general nor too specific. For example, the medical code "*250.10: Diabetes with ketoacidosis, type II or unspecified type, not stated as uncontrolled*" has four ancestors in the CCS-multi-level diagnoses hierarchy. They are "*Endocrine; nutritional; and metabolic diseases and immunity disorders*", "*Diabetes mellitus with complications [50.]*", "*Diabetes with ketoacidosis or uncontrolled diabetes*", and the virtual root of $\mathcal{G}$. Among them, "*Diabetes mellitus with complications [50.]*" may be the best or moderate representative of the code "250.10". Mathematically, $n_j \in \phi(c_i)$ $(1 \leq j \leq |\mathcal{N}|)$ is used to denote the high-level representative of $c_i$. Note that many medical codes may have the same high-level representative $n_j$. In this paper, the node, which appears second in the hierarchy from the root in the CCS multi-level diagnoses, is used as the high-level representative.

Assume that there are $P$ patients in the EHR dataset, and for each patient, there are $T$ visits. It is worth noting that for each patient, $T$ may be different. The $T$ visits can form a sequence $\{V_1, \cdots, V_T\}$. For each visit $V_t$, it contains a subset of leaf codes. In other words, $V_t \subseteq \mathcal{C}$. As mentioned before, each leaf code $c_i$ has an ancestor code set $\phi(c_i)$. For all codes in $V_t$, the union set of their ancestors is denoted as $Q_t$. Two binary vectors are used to simplify the representation of each visit and its ancestor code set. $\mathbf{x}_t$ represents the binary representation of $V_t$. If the $i$-th code appears in $V_t$, then $\mathbf{x}_t[i] = 1$; otherwise, $\mathbf{x}_t[i] = 0$. Note that the length of $\mathbf{x}_t$ is $|\mathcal{C}|$. Similarly, $\mathbf{f}_t$ denotes the binary vector of the ancestor codes $Q_t$. If $n_j$ is in the ancestor code set, then $\mathbf{f}_t[j] = 1$; otherwise, $\mathbf{f}_t[j] = 0$, and the size of $\mathbf{f}_t$ is $|\mathcal{N}|$.

The task of diagnosis prediction is defined as using the graph $\mathcal{G}$, the visit sequence $\{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$ and the corresponding ancestor sequence $\{\mathbf{f}_1, \cdots, \mathbf{f}_T\}$, to predict the set of diagnosis codes that will appear in $V_{T+1}$.

### B. THE PROPOSED KARNS

The proposed knowledge-based Recurrent Neural Networks (KARNS) architecture is depicted in Figure 1. This model utilizes an external knowledge graph $\mathcal{G}$ to obtain the embedding matrices $\mathbf{M}$ for medical codes and $\mathbf{A}$ for ancestor codes, employing a graph-based attention mechanism.

The learned medical code embeddings are then used to generate the representation of each visit. Since in the proposed KARNS, there are two RNNs, and each RNN needs the corresponding input data. For the left RNN, $\mathbf{M}$ and $\mathbf{x}_t$ are used to generate the original visit representation $\mathbf{v}_t$ as its input. For the right RNN, $\mathbf{A}$ and $\mathbf{f}_t$ are used to generate the input $\mathbf{q}_t$. Each RNN will output the hidden state $\mathbf{h}_t$ (left) or $\mathbf{k}_t$ (right). These two hidden states are concatenated to generate a final aggregated representation $\mathbf{s}_t$, which is used for the diagnosis prediction.

#### 1) KNOWLEDGE GRAPH EMBEDDING

To encode the medical ontology structure, the graph-based attention mechanism is employed, inspired by the work of **GRAM** [11]. This allows us to obtain medical code embeddings $\mathbf{M}$ and an ancestor embedding matrix $\mathbf{A}$.

Each medical code $c_i$ in the knowledge graph $\mathcal{G}$ is associated with a learnable basic embedding vector $\mathbf{e}_i \in \mathbb{R}^d$, where $d$ represents the dimensionality. Similarly, each ancestor code $n_j$ is associated with a learnable embedding vector $\mathbf{a}_j \in \mathbb{R}^d$. Here, $1 \leq i \leq |\mathcal{C}|$ represents the index for medical codes, and $1 \leq j \leq |\mathcal{N}|$ represents the index for ancestor codes.

The final representation of a medical code $c_i$, denoted by $\mathbf{m}_i \in \mathbb{R}^d$, is obtained by leveraging the graph-based attention mechanism, which combines the basic embedding $\mathbf{e}_i$ with its ancestors. The specific mechanism for combining the embeddings is described below:

$$\mathbf{m}_i = \alpha_{ii}\mathbf{e}_i + \sum_{j \in \phi(c_i)} \alpha_{ij}\mathbf{a}_j, \tag{1}$$

where $\alpha_{ii} + \sum_{j \in \phi(c_i)} \alpha_{ij} = 1$, and

$$\alpha_{ii} = \frac{\exp(\theta(\mathbf{e}_i, \mathbf{e}_i))}{\exp(\theta(\mathbf{e}_i, \mathbf{e}_i)) + \sum_{j \in \phi(c_i)} \exp(\theta(\mathbf{e}_i, \mathbf{a}_j))},$$
$$\alpha_{ij} = \frac{\exp(\theta(\mathbf{e}_i, \mathbf{a}_j))}{\exp(\theta(\mathbf{e}_i, \mathbf{e}_i)) + \sum_{j \in \phi(c_i)} \exp(\theta(\mathbf{e}_i, \mathbf{a}_j))}. \tag{2}$$

$\theta(\cdot,\cdot)$ is a scalar value and defined as

$$\theta(\mathbf{e}_i, \mathbf{a}_j) = \mathbf{u}_a^\top \tanh(\mathbf{W}_a \begin{bmatrix} \mathbf{e}_i \\ \mathbf{a}_j \end{bmatrix} + \mathbf{b}_a). \tag{3}$$

To compute $\theta(\mathbf{a}_j, \mathbf{a}_j)$, $\mathbf{a}_j$ will be replaced by $\mathbf{e}_i$ in Eq. (3). Therefore, $\theta(\mathbf{e}_i, \mathbf{e}_i)$ can be obtained.

Consequently, the medical code embedding matrix is generated as $\mathbf{M} = \mathbf{m}_1, \mathbf{m}2, \cdots, \mathbf{m}|\mathcal{C}| \in \mathbb{R}^{d\times|\mathcal{C}|}$, where $\mathbf{m}_i$ represents the $i$-th column of $\mathbf{M}$. Similarly, the ancestor embeddings are represented by $\mathbf{A} = \mathbf{a}_1, \mathbf{a}2, \cdots, \mathbf{a}|\mathcal{N}| \in \mathbb{R}^{d\times|\mathcal{N}|}$, where $\mathbf{a}_j$ denotes the $j$-th column of $\mathbf{A}$.

In the original GRAM model, only the medical code embedding matrix $\mathbf{M}$ is utilized in the final prediction, neglecting the significance of ancestor codes. However, ancestor codes possess more general or high-level information about medical codes. Taking into account the high-level representations of medical codes allows us to enhance the predictive performance for diagnosis prediction tasks. Hence, the proposed model KARNS leverages both the medical code embedding matrix $\mathbf{M}$ and the ancestor embedding matrix $\mathbf{A}$ in its architecture.

### 2) VISIT AND REPRESENTATIVE EMBEDDING

Taking into account the importance of high-level representations of medical codes, the proposed model KARNS considers both the leaf code feature matrix $\mathbf{M}$ and the ancestor code feature matrix $\mathbf{A}$. To embed the binary input vectors $\mathbf{x}_t \in 0, 1^{|\mathcal{C}|}$ and $\mathbf{f}_t \in 0, 1^{|\mathcal{N}|}$, the linear transformation with a non-linear activation function is used, which is defined as follows:

$$\mathbf{v}_t = \tanh(\mathbf{M}\mathbf{x}_t), \tag{4}$$

$$\mathbf{q}_t = \tanh(\mathbf{A}\mathbf{f}_t), \tag{5}$$

where $\mathbf{v}_t \in \mathbb{R}^d$ and $\mathbf{q}_t \in \mathbb{R}^d$.

### 3) KNOWLEDGE-BASED TWO-WAY RNNs

In the KARNS architecture, two GRUs with shared parameters are employed. The first GRU is responsible for learning visit-level hidden states, while the second GRU is utilized to capture the high-level information of visits.

To obtain the visit-level hidden state $\mathbf{h}_t \in \mathbb{R}^g$ given the visit-level vector $\mathbf{v}_t$, the following expression can be used:

$$\mathbf{h}_t = \text{GRU}(\mathbf{v}_t; \Omega). \tag{6}$$

Similarly, the high-level hidden state $\mathbf{k}_t \in \mathbb{R}^g$ of the $t$-th visit can be obtained as follows:

$$\mathbf{k}_t = \text{GRU}(\mathbf{q}_t; \Omega). \tag{7}$$

Here, GRU denotes Gated Recurrent Unit (GRU) [19], which belongs to the family of RNN. Note that in the model design, GRU can be replaced by RNN variants, such as Long-Short Term Memory (LSTM) [20] and T-LSTM [21]. $\Omega$ denotes all the parameters of the GRU model.

### 4) DIAGNOSIS PREDICTION

To generate the aggregated representation $\mathbf{s}_t \in \mathbb{R}^{2g}$ in the proposed KARNS, the current hidden state $\mathbf{h}_t$ and the high-level hidden state $\mathbf{k}_t$ are then concatenated using a simple concatenation layer, resulting in:

$$\mathbf{s}_t = [\mathbf{h}_t; \mathbf{k}_t]. \tag{8}$$

The goal of KARNS is to predict the set of diagnosis codes appearing in the $(t+1)$-th visit. To this end, the aggregated representation $\mathbf{s}_t$ is passed through the non-linear activation function softmax to generate the output $\hat{\mathbf{y}}_t$ as follows:

$$\hat{\mathbf{y}}_t = \text{Softmax}(\mathbf{W}_c \mathbf{s}_t + \mathbf{b}_c), \tag{9}$$

where $\mathbf{W}_c \in \mathbb{R}^{|\mathcal{C}|\times 2g}$ and $\mathbf{b}_c \in \mathbb{R}^{|\mathcal{C}|}$ are learnable parameters.

### 5) OBJECTIVE FUNCTION

The loss for each patient across all timestamps is calculated by employing the cross-entropy between the ground truth visit $\mathbf{y}_t$ and the predicted visit $\hat{\mathbf{y}}_t$, based on Eq. (9). The loss can be expressed as:

$$\mathcal{L}_p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T; \mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_T) = -\frac{1}{T-1}\sum_{t=1}^{T-1}\left(\mathbf{y}_t^\top \log(\hat{\mathbf{y}}_t) + (\mathbf{1}-\mathbf{y}_t)^\top \log(\mathbf{1}-\hat{\mathbf{y}}_t)\right). \tag{10}$$

Note that this loss is only for an individual patient. The average of the individual cross-entropy loss for all $P$ patients is calculated as follows:

$$\mathcal{L} = \frac{1}{P}\sum_{p=1}^{P}\mathcal{L}_p. \tag{11}$$

## III. EXPERIMENTS

This section begins by presenting three real-world datasets utilized in the experiments. Subsequently, the experimental setup is described. Finally, the performance of the proposed KARNS model on these three real-world datasets is analyzed. The results indicate that KARNS outperforms state-of-the-art predictive models across multiple evaluation strategies.

### A. REAL-WORLD DATASETS

In the experiments, three datasets are used to evaluate the effectiveness of KARNS. One is a publicly available dataset, and the other two datasets are private.

The first one is the Medical Information Mart for Intensive Care (MIMIC-III) dataset [22], which is a publicly available electronic health record dataset comprising medical records of ICU patients over 11 years and has been used by existing study [10], [23]. Since the goal of this task is to predict the next visit's diagnosis code set, if the patient only has one recorded visit, then it will be removed from the dataset. This dataset is characterized by short visit lengths and a relatively small number of patients, allowing us to assess the performance of KARNS in scenarios with limited training data.

The Medicaid dataset, an insurance claim dataset, contains records of 99,159 patients and 2,034,485 visits spanning the years 2011 and 2012, and the selected patients had at least ten visits, enabling us to evaluate KARNS on datasets with longer visit records.

Furthermore, another dataset focusing on one specific disease Diabetes is used, which is a subset of Medicaid. In this dataset, all patients had a diagnosis of diabetes, as indicated by the presence of the ICD9 diagnosis code 250.xx in their claims.

In the context of diagnosis prediction, the objective is to predict the diagnosis information for the next visit. However, instead of directly predicting the specific diagnosis codes, the experiments follow the approach used in [11] and [13] and predict diagnosis categories.

There are several benefits to predicting category information. Firstly, it improves training speed and predictive performance compared to predicting individual diagnosis codes. Secondly, it ensures an adequate level of granularity for all diagnoses. In the experiments, the nodes in the second hierarchy of the ICD9 codes are used as category labels. It is worth noting that the hierarchy of CCS (Clinical Classifications Software) can also be utilized as category labels, and both grouping methods yield similar predictive performance [11].

More detailed information about the three real-world datasets is presented in Table 1.

**TABLE 1.** Statistics of MIMIC-III, Medicaid, and Diabetes Dataset.

| Dataset | MIMIC-III | Medicaid | Diabetes |
|---|---|---|---|
| # of patients | 7,499 | 99,159 | 17,584 |
| # of visits | 19,911 | 2,034,485 | 466,024 |
| Avg. visits per patient | 2.66 | 20.52 | 26.50 |
| # of unique ICD9 codes | 4,880 | 9,701 | 7,437 |
| Avg. # of codes per visit | 13.06 | 2.78 | 3.39 |
| Max # of codes per visit | 39 | 41 | 37 |
| # of category codes | 171 | 157 | 155 |
| Avg. # of codes per visit | 10.16 | 2.30 | 2.92 |
| Max # of codes per visit | 30 | 23 | 22 |

## B. EXPERIMENTAL SETUP

In order to conduct a fair evaluation of the proposed KARNS, a list of state-of-the-art baselines is presented first. These baselines serve as reference models for comparison. Next, the evaluation metrics are introduced and used to assess the predictive performance of KARNS. These metrics provide a comprehensive evaluation of the model's effectiveness. Lastly, detailed information regarding the implementation, including dataset preprocessing, model architecture, hyperparameter settings, and the optimization algorithm employed, is introduced. These implementation details ensure the transparency and reproducibility of the experimental setup.

**Baselines**

The following four state-of-the-art approaches are used:

- RNN serves as a naive baseline that directly embeds visit information into vector representations using a GRU. In other words, it does not use any extra knowledge. Each visit will generate a hidden state, which is then used to make predictions.
- RNN+ extends the RNN model by incorporating a location-based attention model, similar to Dipole. The main difference is that RNN+ uses a unidirectional GRU for predictions, while Dipole employs a bidirectional GRU.
- Dipole [13] uses BRNNs and attention mechanisms for predicting future visit information. It achieves the best performance among diagnosis prediction approaches that do not employ medical ontologies. The visit sequence is embedded using a multilayer perceptron (MLP), and the bidirectional GRUs with attention mechanisms generate latent vectors for predictions.
- GRAM [11] is different from the previous baselines, which is the first work to incorporate a medical knowledge ontology and recurrent neural networks for diagnosis prediction. It embeds each visit in a time-ordered sequence using a medical code embedding matrix learned from the knowledge graph. The embedded visit vectors are then fed into a GRU to predict the next visit information.

### 1) EVALUATION MEASURES

The diagnosis prediction task aims to predict a set of diagnosis codes, which is different from classification tasks. To evaluate the performance of all approaches, two evaluation metrics are used: visit-level accuracy@$k$ and code-level accuracy@$k$. The accuracy@$k$ metric measures the correctness of the predicted medical codes. Next, the details of these two metrics with an example are introduced.

For example, there are two visits, and their ground truth labels are $G_1^p = \{c_1, c_3, c_4\}$ and $G_2^p = \{c_1, c_2\}$. The predictions are $\hat{G}_1^p = \{[c_1, c_2, c_3, c_4, c_7, c_8]\}$ and $\hat{G}_2^p = \{[c_1, c_2, c_3, c_4]\}$.

- **Visit-level accuracy**@$k$ evaluates the average accuracy of predicting the correct medical codes among the top $k$ guesses. It is calculated by dividing the number of correct medical codes in the top $k$ predictions by the minimum value between $k$ and the total number of category labels in the $(t+1)$-th visit, which is defined as follows:

$$\text{V-Acc@}k = \frac{1}{|\mathcal{P}_t|}\sum_{p=1}^{|\mathcal{P}_t|}\frac{1}{T_p}\sum_{t=1}^{T_p}\frac{|G_t^p \cap \hat{G}_t^p[1,k]|}{\min\{k, |G_t^p \cap \hat{G}_t^p[1,k]|\}}, \tag{12}$$

where $|\mathcal{P}_t|$ denotes the number of patients in the testing data, $T_p$ is the number of visits in the $p$-th patient's record, $G_t^p$ is the set of the ground truth codes of the

**TABLE 2. Performance in terms of the metric accuracy@*k*.**

| Dataset | Model | Visit-Level Accuracy@$k$ | | | | | | Code-Level Accuracy@$k$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 5 | 10 | 15 | 20 | 25 | 30 | 5 | 10 | 15 | 20 | 25 | 30 |
| MIMIC-III | KARNS | **0.708** | **0.658** | **0.696** | **0.758** | **0.812** | **0.853** | **0.316** | **0.511** | **0.637** | **0.724** | **0.787** | **0.833** |
| | GRAM | 0.699 | 0.645 | 0.685 | 0.744 | 0.801 | 0.842 | 0.312 | 0.503 | 0.629 | 0.714 | 0.779 | 0.827 |
| | Dipole | 0.622 | 0.584 | 0.631 | 0.695 | 0.756 | 0.806 | 0.277 | 0.456 | 0.580 | 0.667 | 0.735 | 0.790 |
| | RNN+ | 0.616 | 0.580 | 0.624 | 0.691 | 0.754 | 0.802 | 0.276 | 0.455 | 0.575 | 0.665 | 0.735 | 0.787 |
| | RNN | 0.658 | 0.618 | 0.664 | 0.725 | 0.784 | 0.827 | 0.294 | 0.484 | 0.611 | 0.696 | 0.763 | 0.812 |
| Medicaid | KARNS | **0.598** | **0.731** | **0.801** | **0.846** | **0.880** | **0.906** | **0.543** | **0.698** | **0.774** | **0.825** | **0.862** | **0.891** |
| | GRAM | 0.583 | 0.719 | 0.790 | 0.837 | 0.872 | 0.898 | 0.528 | 0.684 | 0.763 | 0.815 | 0.853 | 0.882 |
| | Dipole | 0.594 | 0.723 | 0.789 | 0.834 | 0.868 | 0.894 | 0.541 | 0.690 | 0.764 | 0.813 | 0.850 | 0.879 |
| | RNN+ | 0.596 | 0.721 | 0.792 | 0.839 | 0.875 | 0.901 | 0.540 | 0.687 | 0.764 | 0.817 | 0.855 | 0.885 |
| | RNN | 0.545 | 0.674 | 0.750 | 0.804 | 0.843 | 0.874 | 0.494 | 0.630 | 0.720 | 0.778 | 0.822 | 0.856 |
| Diabetes | KARNS | **0.578** | **0.724** | **0.798** | **0.847** | **0.882** | **0.908** | **0.517** | **0.698** | **0.779** | **0.831** | **0.869** | **0.897** |
| | GRAM | 0.559 | 0.705 | 0.782 | 0.833 | 0.868 | 0.896 | 0.496 | 0.678 | 0.762 | 0.816 | 0.855 | 0.885 |
| | Dipole | 0.569 | 0.702 | 0.776 | 0.827 | 0.864 | 0.892 | 0.511 | 0.677 | 0.758 | 0.812 | 0.852 | 0.882 |
| | RNN+ | 0.568 | 0.701 | 0.777 | 0.828 | 0.865 | 0.894 | 0.509 | 0.674 | 0.757 | 0.812 | 0.852 | 0.884 |
| | RNN | 0.552 | 0.685 | 0.764 | 0.818 | 0.858 | 0.888 | 0.498 | 0.661 | 0.746 | 0.802 | 0.844 | 0.876 |

$t$-th visit, $\hat{G}_t^p$ is the predicted category code set, and $\hat{G}_t^p[1,k]$ means the top-$k$ predictions. In this example, if $k = 2$, $\hat{G}_1^p[1,2] = \{c_1, c_2\}$ and $\hat{G}_2^p[1,2] = \{c_1, c_2\}$, and visit-level accuracy@2 $= \frac{1}{2}(\frac{1}{\min\{2,3\}} + \frac{2}{\min\{2,2\}}) = \frac{1}{2}(0.5 + 1) = 0.75$.

- **Code-level accuracy@$k$** evaluates the accuracy of predicting the correct category labels within a visit. If the target label is among the top $k$ predicted labels for a visit that contains multiple category labels, it is considered as a correct prediction. Code-level accuracy@$k$ is determined by dividing the number of correct label predictions by the total number of label predictions, which is defined as follows:

$$\text{C-Acc@}k = \frac{1}{|\mathcal{P}_t|}\sum_{p=1}^{|\mathcal{P}_t|}\frac{1}{T_p}\sum_{t=1}^{T_p}\frac{|G_t^p \cap \hat{G}_t^p[1,k]|}{|\hat{G}_t^p|}, \qquad (13)$$

where $|\hat{G}_t^p|$ denotes the number of category codes in the $t$-th predicted visit. Still in this example, when $k = 2$, accuracy@2 $= \frac{1}{2}(\frac{1}{6} + \frac{2}{4}) = \frac{1}{2}(0.17 + 0.5) = 0.33$.

Both visit-level and code-level accuracy@$k$ measures provide insights into the performance of approaches at different levels of granularity. Higher values indicate better performance for all four measures. In the experiments, $k$ is varied from 5 to 30 to assess the performance across different prediction scenarios.

#### 2) IMPLEMENTATION DETAILS

In the experiments, the CCS-multi-level ontology is utilized as the knowledge graph, following the approach proposed by Choi et al. [11]. For medical codes with multiple ancestors, the codes appearing in the second hierarchy as their representatives are selected. A code has only one ancestor, which is treated as its representative.

The implementations of all approaches are based on PyTorch 2.0. Each dataset is randomly divided into three parts in a 0.75:0.10:0.15 ratio as training, validation, and testing sets at the patient level. All baselines and the proposed approach use the same training, validation, and testing sets. The lowest loss on the validation set is stored as the optimal parameter set during the training stage. For training the models, the Adadelta optimization algorithm [24] with a minibatch size of 50 patients is employed. Regularization is applied using $\ell_2$ norm with a coefficient of 0.001, and dropout with a rate of 0.5 is employed for all approaches.

To ensure a fair performance comparison, the dimension $d$ is set to 128 for all methods. Additionally, for GRAM and the proposed KARNS, the attention size $l$ is set to 128 as well.

### C. RESULTS OF DIAGNOSIS PREDICTION

To fairly compare all the approaches, different values of $k$ are used in the experiments on the three real-world datasets. The experimental results in terms of the accuracy@$k$ measure are shown in Table 2, where set $k$ from 5 to 30, and the proposed KARNS achieves the superior performance.

On the MIMIC-III dataset, the performance of GRAM is much better than that of other baselines. However, KARNS significantly outperforms GRAM in terms of both visit-level and code-level accuracy. This indicates that leveraging high-level recurrent neural networks can greatly improve predictive performance. It is worthy noting that both Dipole and RNN+ perform worse than the naive baseline, RNN. This is likely due to the limited number of visits per patient in this dataset. Insufficient data prevents Dipole and RNN+ from accurately learning attention modules, resulting in poor prediction performance. In contrast, KARNS leverages high-level information from previous visits to learn general representations for future visits, mitigating the impact of limited training data.

Regarding the values of $k$, it can be observed that the values generally increase as $k$ becomes larger, except for the visit-level accuracy on the MIMIC-III dataset. This can be attributed to the lack of training data for certain labels. The lower probabilities assigned to these labels in the predictions lead to fewer correct predictions when $k$ is larger. However,

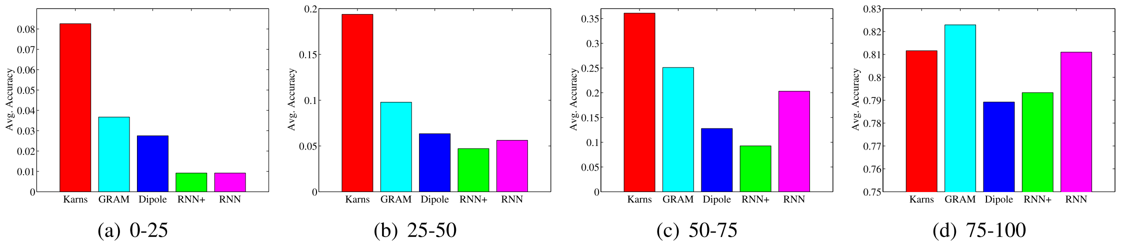

**FIGURE 2.** Results of different label frequency groups on the MIMIC-III dataset.

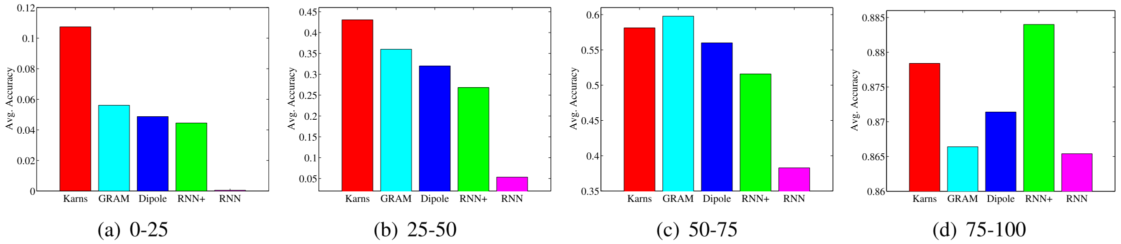

**FIGURE 3.** Results of different label frequency groups on the Medicaid dataset.

the number of correct predictions is divided by a larger value of $\min(k, |\mathbf{y}_t|)$, resulting in lower average performance compared to the case when $k = 5$.

On the Medicaid dataset, both Dipole and RNN+ outperform GRAM for small values of $k$, indicating the superior ability of attention mechanisms to improve predictive accuracy when sufficient data is available. However, the accuracy of KARNS remains higher than both Dipole and RNN+, confirming the effectiveness of considering general or high-level information in improving prediction performance.

On the Diabetes dataset, the accuracy of KARNS surpasses all the baseline approaches. This can be attributed to the high relevance of most medical codes to diabetes in this dataset. The proposed KARNS effectively captures the relationships among medical codes using high-level information, leading to accurate predictions.

It is worth noting that the performance of GRAM is comparable to that of both Dipole and RNN+ on the Diabetes dataset, indicating that models utilizing knowledge graphs can achieve similar performance to models employing attention mechanisms.

In summary, the results in Table 2 demonstrate the robustness and effectiveness of KARNS across different types of datasets, supporting its superiority over the baseline methods.

### D. DATA SUFFICIENCY ANALYSIS

The results of different diagnosis prediction approaches are indeed sensitive to the number of available training instances, as observed in Table 2. To further analyze the impact of training dataset sizes, additional experiments were conducted on the MIMIC-III and Medicaid datasets.

In this experiment, the data are divided into four groups, according to the frequency of category labels appearing in the training data. The frequency is ranked in an ascending order. The first 25% data are as the first group, and the group name is 0-25. The next 25% data are in the second group, named 25-50. Similarly, the following two groups are 50-75 and 75-100. Such a division method can clearly distinguish the rarest labels from the most common ones. The value of the accuracy of labels in each group is calculated.

Figure 2 shows the results in terms of the code-level accuracy@20 on the MIMIC-III dataset. The x-axis represents the different approaches, while the y-axis denotes the average accuracy of the approaches. Figure 3 depicts the code-level accuracy@20 on the Medicaid dataset.

From Figure 2 and Figure 3, it can be observed that the proposed KARNS consistently outperforms the baselines in the 0-25 and 25-50 groups. This demonstrates the effectiveness of KARNS in predicting codes with limited training data, emphasizing the importance of incorporating high-level information.

In Figure 2, among all the baselines, GRAM achieves the best performance, indicating that even with insufficient data, GRAM can still learn reasonable medical code representations based on the knowledge graph, which improves the predictions. However, the accuracy of GRAM is lower compared to the proposed KARNS, highlighting the significance of incorporating high-level information in the diagnosis prediction task.

On the other hand, in the Medicaid dataset where there is sufficient training data, attention mechanisms play a more important role. Figure 2 shows that RNN+ and Dipole exhibit relatively better performance in the 75-100 group. In comparison, the performance of GRAM is inferior to that

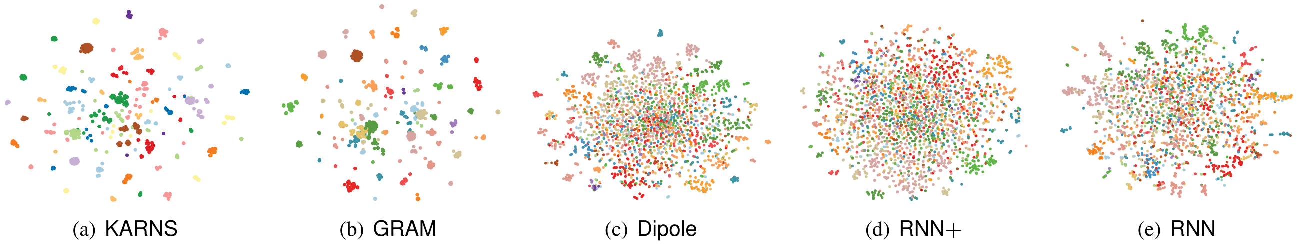
(a) KARNS (b) GRAM (c) Dipole (d) RNN+ (e) RNN

**FIGURE 4. Visualization of the learned medical code representations on the diabetes dataset using *t*-SNE scatterplots.**

of RNN+ and Dipole. This suggests that when training data is sufficient, calculating attention weights for past visits can achieve similar performance to incorporating a medical knowledge graph. However, KARNS goes beyond these approaches by integrating both the medical knowledge graph and high-level information, leading to the best overall performance, as shown in Table 2.

In conclusion, the experiments provide further evidence of the effectiveness of KARNS in predicting codes with limited training data, as well as the importance of incorporating high-level information in improving prediction performance.

### *E. INTERPRETABLE REPRESENTATION ANALYSIS*

Interpretability of medical code representations is indeed crucial for the diagnosis prediction task. To provide a qualitative demonstration of the learned representations by all the models on the Diabetes dataset, a random selection of 2000 medical codes was made, and these codes were plotted in a 2-D space using *t*-SNE (*t*-Distributed Stochastic Neighbor Embedding) [25], as shown in Figure 4. The color of the dots in the plot represents the first disease categories in the CCS multi-level hierarchy.

From the figure, it can be observed that both KARNS and GRAM have the ability to cluster diagnosis codes into interpretable clusters. This demonstrates that these models have learned representations that capture meaningful relationships among the medical codes. However, it is worth noting that the predictive performance of KARNS is significantly better than that of GRAM, as shown in Table 2. This observation highlights the fact that KARNS not only maintains the advantages of GRAM in terms of interpretability but also improves the prediction accuracy.

Furthermore, from Figure 4(c), Figure 4(d), and Figure 4(e), it can be concluded that without a knowledge graph, relying solely on co-occurrences or supervised predictions cannot easily learn interpretable medical code representations. This further emphasizes the importance of incorporating medical knowledge into the models, as done in KARNS and GRAM, to obtain both interpretable and accurate representations.

In summary, the qualitative analysis of the learned representations supports the effectiveness of KARNS in generating interpretable medical code representations while achieving superior predictive performance compared to GRAM and other baselines.

## IV. RELATED WORK

This section presents an overview of related work in the field, focusing on mining electronic healthcare records (EHR) using deep learning techniques and the specific area of diagnosis prediction.

### *A. DEEP LEARNING FOR EHR*

Deep learning techniques have been widely applied to extract valuable medical knowledge from EHR data, encompassing both structured and unstructured information [5]. Convolutional neural networks (CNNs) have been utilized for predicting unplanned readmission [6], diseases [7], and risk [4], [26] based on EHR data. Stacked denoising autoencoders (SDAs) have been employed to identify characteristic physiological patterns in clinical time series data [2]. To capture the temporal dynamics in EHR data, recurrent neural networks (RNNs) have been extensively used for modeling disease progression [8], [27], handling time series healthcare data with missing values [28], [29], and performing diagnosis classification [30] and prediction [10], [13], [16], [17], [18], [23], [31], [32], [33], [34].

### *B. DIAGNOSIS PREDICTION*

Diagnosis prediction, a core task in EHR data mining, involves predicting future visit information based on historical records. **Med2Vec** [14] is an unsupervised method that learns interpretable embeddings of medical codes. Although it can be used for predicting future health information, it disregards long-term dependencies among medical codes across visits. **RETAIN** [10] incorporates a reverse time attention mechanism in an RNN to interpret the contribution of each medical code to the current prediction, focusing on binary prediction tasks. Dipole [13] is a state-of-the-art diagnosis prediction model that employs bidirectional recurrent neural networks (BRNNs) and different attention mechanisms to predict patient visit information. **GRAM** [11] utilizes a graph-based attention mechanism on a medical ontology to learn robust medical code embeddings, even in scenarios with limited training data, and employs an RNN to model patient visits.

Among the mentioned diagnosis prediction models, the most relevant one to the proposed KARNS is GRAM. Compared to GRAM, the proposed KARNS not only applies a graph-based attention mechanism to learn interpretable medical code embeddings but also incorporates high-level representations of medical codes. This integration significantly improves prediction accuracy and ensures robustness compared to state-of-the-art approaches.

## V. DISCUSSION

Although the proposed KARNS outperforms baselines, it still suffer from the following issues:

- **Integration of temporal dynamics**: Currently, KARNS captures information from previous visits using recurrent neural networks. Future research can explore more advanced models that effectively capture the temporal dynamics within EHR data, such as transformer-based models [35].
- **Enhancement of medical code embeddings**: Although KARNS incorporates accurate embeddings of medical codes and their ancestors, there is room for improvement. Future studies can explore techniques to refine medical code embeddings, such as incorporating external knowledge sources or leveraging pre-trained language models specifically trained on medical text [23].
- **Handling missing data**: EHR data often contain missing values, which can impact prediction accuracy. Future work can focus on developing strategies to handle missing data in KARNS, such as imputation techniques or attention mechanisms that dynamically weigh the available information.
- **Interpretability and explainability**: While KARNS demonstrates interpretability through visualizing learned medical code representations, further research can delve into enhancing the interpretability and explainability of the model's predictions. This could involve techniques such as attention mechanisms, saliency maps, or rule extraction methods to provide insights into the decision-making process.
- **Generalization across diverse populations**: It is crucial to evaluate the generalizability of KARNS across diverse patient populations. Future studies can investigate the model's performance on datasets representing different demographic groups, ensuring fairness and avoiding bias in diagnosis predictions.
- **Real-time prediction and clinical deployment**: Consider exploring the feasibility of implementing KARNS in real-time prediction systems within clinical settings. Conducting studies that validate the model's performance in real-time scenarios can provide valuable insights into its practical usability and potential integration with existing healthcare systems.

## VI. CONCLUSION

Diagnosis prediction poses a significant challenge in healthcare informatics due to limitations in existing approaches when handling various types of electronic healthcare record (EHR) datasets and incorporating high-level representations of medical codes. To address these challenges and accurately predict patients' future visit information, this paper introduces a novel end-to-end model called KARNS.

The proposed KARNS model utilizes a given medical ontology to identify high-level representatives for medical codes and generates representative vectors for each visit. By incorporating precise embeddings of medical codes and their ancestors, the model captures both specific and general knowledge. To leverage information from previous visits and the knowledge graph, KARNS employs two recurrent neural networks (RNNs). By combining the hidden states of these RNNs into a novel vector representation, KARNS significantly enhances prediction accuracy.

Experimental results on three real-world medical datasets demonstrate the effectiveness and robustness of KARNS for diagnosis prediction. The proposed model outperforms state-of-the-art approaches, regardless of the availability of abundant or insufficient EHR data. Additionally, the learned medical code representations in KARNS exhibit interpretability, as demonstrated through visualization.

In summary, the KARNS model offers a powerful and robust solution for diagnosis prediction in healthcare informatics. It leverages high-level representations of medical codes, incorporates knowledge from medical ontologies, and achieves state-of-the-art performance across different types of EHR datasets.

## REFERENCES

[1] H. Malik, N. Fatema, and J. A. Alzubi, *AI and Machine Learning Paradigms for Health Monitoring System: Intelligent Data Analytics*. Berlin, Germany: Springer, 2021.

[2] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, "Deep computational phenotyping," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 507–516.

[3] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 787–792.

[4] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," in *Proc. SIAM Int. Conf. Data Mining*, Jun. 2016, pp. 432–440.

[5] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings Bioinf.*, vol. 19, no. 1, pp. 1236–1246, Jan. 2018.

[6] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: A convolutional net for medical records," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 22–30, Jan. 2017.

[7] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, A. Zhang, and J. Gao, "Personalized disease prediction using a CNN-based similarity learning method," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 811–816.

[8] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "DeepCare: A deep dynamic memory model for predictive medicine," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2016, pp. 30–41.

[9] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, and P. Sundberg, "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, no. 1, p. 18, 2018.

[10] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3504–3512.

[11] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 787–795.

[12] Q. Suo, F. Ma, G. Canino, J. Gao, A. Zhang, P. Veltri, and A. Gnasso, "A multi-task framework for monitoring health conditions via attention-based recurrent neural networks," in *Proc. AMIA Annu. Symp.*, 2017, p. 1665.

[13] 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 *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1903–1911.

[14] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1495–1504.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[16] M. Gupta, T.-L.-T. Phan, H. T. Bunnell, and R. Beheshti, "Obesity prediction with EHR data: A deep learning approach with interpretable elements," *ACM Trans. Comput. Healthcare*, vol. 3, no. 3, pp. 1–19, Jul. 2022.

[17] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–10.

[18] F. Yang, J. Zhang, W. Chen, Y. Lai, Y. Wang, and Q. Zou, "DeepMPM: A mortality risk prediction model using longitudinal EHR data," *BMC Bioinf.*, vol. 23, no. 1, p. 423, Oct. 2022.

[19] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[21] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware LSTM networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 65–74.

[22] A. E. W. 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," *Sci. Data*, vol. 3, no. 1, pp. 1–9, May 2016.

[23] F. Ma, Y. Wang, H. Xiao, Y. Yuan, R. Chitta, J. Zhou, and J. Gao, "A general framework for diagnosis prediction via incorporating medical code descriptions," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 1070–1075.

[24] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701*.

[25] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[26] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, "Risk prediction on electronic health records with prior medical knowledge," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1910–1919.

[27] C. Che, C. Xiao, J. Liang, B. Jin, J. Zho, and F. Wang, "An RNN architecture with dynamic temporal matching for personalized predictions of Parkinson's disease," in *Proc. SIAM Int. Conf. Data Mining*, 2017, pp. 198–206.

[28] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," 2016, *arXiv:1606.01865*.

[29] Z. C. Lipton, D. C. Kale, and R. Wetzel, "Modeling missing data in clinical time series with RNNs," in *Proc. 1st Mach. Learn. Healthcare Conf.*, 2016, pp. 253–270.

[30] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with LSTM recurrent neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–18.

[31] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. 1st Mach. Learn. Healthcare Conf.*, 2016, pp. 301–318.

[32] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *Npj Digit. Med.*, vol. 4, no. 1, p. 86, May 2021.

[33] E. Choi, C. Xiao, W. Stewart, and J. Sun, "MiME: Multilevel medical embedding of electronic health records for predictive healthcare," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4547–4557.

[34] J. G. D. Ochoa and F. E. Mustafa, "Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses," *Artif. Intell. Med.*, vol. 131, Sep. 2022, Art. no. 102359.

[35] J. Luo, M. Ye, C. Xiao, and F. Ma, "HiTANet: Hierarchical time-aware attention networks for risk prediction on electronic health records," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 647–656.

**HUA SHEN** received the Ph.D. degree from the Dalian University of Technology, China. She is currently an Associate Professor with the College of Artificial Intelligence, Anshan Normal University, China. Her research interests include data mining and machine learning.