Disease Prediction via Graph Neural Networks

Zhenchao Sun, Hongzhi Yin, Hongxu Chen, Tong Chen, Lizhen Cui, and Fan Yang

***Abstract*—With the increasingly available electronic medical records (EMRs), disease prediction has recently gained immense research attention, where an accurate classifier needs to be trained to map the input prediction signals (e.g., symptoms, patient demographics, etc.) to the estimated diseases for each patient. However, existing machine learning-based solutions heavily rely on abundant manually labeled EMR training data to ensure satisfactory prediction results, impeding their performance in the existence of rare diseases that are subject to severe data scarcity. For each rare disease, the limited EMR data can hardly offer sufficient information for a model to correctly distinguish its identity from other diseases with similar clinical symptoms. Furthermore, most existing disease prediction approaches are based on the sequential EMRs collected for every patient and are unable to handle new patients without historical EMRs, reducing their real-life practicality. In this paper, we introduce an innovative model based on Graph Neural Networks (GNNs) for disease prediction, which utilizes external knowledge bases to augment the insufficient EMR data, and learns highly representative node embeddings for patients, diseases and symptoms from the medical concept graph and patient record graph respectively constructed from the medical knowledge base and EMRs. By aggregating information from directly connected neighbor nodes, the proposed neural graph encoder can effectively generate embeddings that capture knowledge from both data sources, and is able to inductively infer the embeddings for a new patient based on the symptoms reported in her/his EMRs to allow for accurate prediction on both general diseases and rare diseases. Extensive experiments on a real-world EMR dataset have demonstrated the state-of-the-art performance of our proposed model.**

***Index Terms*—Disease prediction, big data health applications, data mining, graph embedding.**

Manuscript received April 12, 2020; revised May 25, 2020; accepted June 16, 2020. Date of publication June 22, 2020; date of current version March 5, 2021. This work was supported in part by NSFC under Grant 91846205, in part by National Key R&D Program under Grant 2017YFB1400100, in part by the Innovation Method Fund of China under Grant 2018IM020200, in part by Shandong Key R&D Program under Grants 2018YFJH0506 and 2019JZZY011007, and in part by Australian Research Council under Grant DP190101985. *(Corresponding author: Hongzhi Yin.)*

Zhenchao Sun and Lizhen Cui are with the School of Software, Shandong University, Jinan 250100, China (e-mail: zhenchao.sun@mail.sdu.edu.cn; clz@sdu.edu.cn).

Hongzhi Yin and Tong Chen are with the School of Information Technology & Electrical Engineering, The University of Queensland, Saint Lucia, QLD 4072, Australia (e-mail: h.yin1@uq.edu.au; tong.chen@uq.edu.au).

Hongxu Chen is with the School of Computer Science, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: hongxu.chen@uts.edu.au).

Fan Yang is with the School of Public Health & Institute for Medical Dataology, Shandong University, Jinan 250100, China (e-mail: fanyang@sdu.edu.cn).

Digital Object Identifier 10.1109/JBHI.2020.3004143

I. Introduction

As a widely-used data management scheme, electronic medical records (EMRs) are used to store the rich clinical data collected from different patients' visits to hospitals. Recently, with the prosperous advances in information technology and machine learning, the sheer volume of EMRs is becoming more manageable, and analyzing EMRs with machine learning and data mining techniques is becoming an emerging research direction to fulfill the goal of improving health care services [1].

An important application of machine learning in healthcare is disease prediction that aims to predict whether a patient suffers from a certain disease, where the task is commonly formulated as learning a classifier that infers the prediction results from EMRs [2], [3]. For example, Palaniappan and Awang applied a series of data mining techniques, namely Decision Trees [4], Naive Bayes [5] and Neural Networks [6], to build a heart disease prediction system [7]. With the power of Convolutional Neural Networks (CNNs), Suo *et al.* [2] firstly identified the similarity between patients based on their EMRs, and then performed personalized disease predictions. Ma *et al.* [3] further incorporated discrete prior medical knowledge into CNNs to improve the prediction performance.

Nevertheless, training these models requires a large amount of EMR data with respect to each particular disease, hindering existing models from generating accurate predictions when there are no sufficient disease-specific EMR records, e.g., predicting rare diseases. Rare diseases have a common characteristic of low prevalence and perception, while both the treatments and related medical research are inadequate [8]. This leads to critical challenges in the clinical diagnosis of a rare disease. According to the fact that up to 8% of the human population is affected by rare disease [9], the average time it takes to achieve a correct diagnosis for a rare disease case can be as much as 4.8 years.[1] The difficulty of identifying rare diseases is mainly associated with the high diversity of such diseases (more than 6,000 types discovered) and the lack of clinical experience [9], [10]. As a result, the majority of these patients are suffering from long-term illness, despair, and even wrong treatments caused by misdiagnosis. Therefore, on top of the conventional disease prediction, accurately predicting rare diseases at an early stage will help patients receive prevention treatments in a timely manner, thus significantly increasing their survival rates and minimizing the harm from such diseases. At the same time, most existing models make predictions in a sequential manner, where historical EMRs for a patient is an indispensable part of the model input. Consequently, these approaches are incompatible

[1][Online]. Available: https://globalgenes.org/rare-facts/

to new patients that are unseen during training, rendering them unable to make predictions for new patients.

Despite the importance of a machine learning model's sensitivity to rare diseases in the disease prediction task, as suggested by the name, it is, however, extremely difficult to collect abundant EMR data on these rare diseases to train a robust and reliable classifier. Moreover, in EMRs, some rare diseases may develop symptoms similar to the ones of common diseases, which offer counterfeit signals and are prone to be misclassified. In this regard, instead of solely relying on the EMRs, a classifier should be able to fully utilize various external information sources to guarantee the prediction performance in the presence of rare diseases. Fortunately, in addition to EMRs, a wide range of medical knowledge has been standardized by reputable institutions and published as well-organized ontologies, including the International Classification of Diseases (ICD)[2], the Human Phenotype Ontology (HPO)[3] and Orphanet.[4] Such information sources are essentially structured knowledge bases containing verified and valuable disease-related metadata (i.e., attributes), which bring immense potentials for improving the disease prediction accuracy in practice.

To this end, in light of the availability of both the public disease knowledge bases and EMRs, we respectively formulate two information sources as a medical concept graph and a patient record graph, and introduce a novel graph embedding-based model for disease prediction. Both constructed graphs are heterogeneous, where the medical concept graph links diseases with related symptoms and patient record graph extracts connections between patients and observed symptoms from the EMRs. Specifically, by investigating a disease's/patient's associations to different symptoms, we build a novel disease prediction model upon Graph Neural Networks (GNNs) [11] to encode the information of different symptoms, users and diseases into compact but representative latent vectors (a.k.a., embeddings). As such, the probability of observing an active patient-disease relationship (i.e., predicting a disease type for a patient) can be easily inferred via the similarity between the embeddings of a disease and the target patient.

In order to overcome the shortage of EMR data when learning to predict rare diseases, our model gains external medical knowledge on both the diseases and symptoms by aggregating the information of their connected neighbors from the medical concept graph. Hence, in the prediction stage, given an arbitrary patient that is a new visitor to the hospital (the patient's EMRs are unseen during training), our model can effectively generate the patient's embedding by merging the learned latent representations of symptoms reported in her/his EMRs. In addition, as our model is subsumed under a generic graph embedding framework that is not restricted to specific information aggregation schemes, we further explore two distinct neighborhood information aggregators, namely the Graph Attention Networks (GATs) [12] and Graph Isomorphic Networks (GINs) [13] via comparative studies.

[2][Online]. Available: http://www.who.int/
[3][Online]. Available: https://https://hpo.jax.org/app/
[4][Online]. Available: https://www.orpha.net/

The contributions of this paper are summarized as follows.

- We identify the challenges of predicting both common and rare diseases based on EMRs, and introduce a systematic solution by fusing expert knowledge with machine learning techniques.
- We propose a novel graph embedding-based model for disease prediction. The model inductively learns embeddings from the medical concept graph and patient record graph respectively extracted from the external knowledge base and EMR data, while being able to handle new patients and identify highly relevant symptoms to support accurate disease prediction.
- Extensive experiments on real-world EMR datasets have been conducted, and the results suggest that our model outperforms all baseline models in both common disease and rare disease prediction tasks.

II. Related Work

The creation and adoption of electronic medical records (EMRs) have ignited widespread interest and opened up abundant opportunities for clinical and translational research [14], thus motivating further studies on the prediction of risks, diagnoses, and diseases. Choi *et al.* [15] proposed GRaph-based Attention Model (GRAM) that supplements EMRs with hierarchical information extracted from medical ontologies. Choi *et al.* [16] also developed the REverse Time AttentIoN model (RETAIN) for utilizing the EMR data to improve both the accuracy and interpretability of predictive models. However, using Recurrent Neural Networks (RNNs) as the main building block, such approaches suffer from a severe performance drop when the length of the sequences becomes too large for RNNs to learn long-range dependencies. To address this issue, Ma *et al.* [17] proposed Dipole to predict patients' future health information, which employs bidirectional RNNs to memorize all information of both long-term and short-term patient status, and leverages three attention mechanisms to measure the contributions of different visits to the prediction. However, as typical variants of deep neural networks, these approaches invariably require a large amount of training data to learn the complex non-linear functions and data-driven patterns for accurate prediction. Consequently, these methods tend to underperform in EMR-related prediction tasks when sufficient EMR data is unavailable. To address this issue, Ma *et al.* [18] presented an end-to-end model named KAME to exploit external medical knowledge to improve the accuracy of diagnosis prediction. One recent work [19] proposed a meta-learning framework for clinic risk prediction with limited patient record data, which transfers knowledge from other closely related but information-intensive disease domains.

Among various EMR-related tasks, the disease prediction task is designed to predict whether a patient suffers from a certain disease based on the historical EMR data. Disease prediction is formulated as a classification task, where a wide range of traditional classification approaches have been applied to solve it. For instance, Palaniappan and Awang applied Decision Trees, Naive Bayes and Neural Network, and introduced a system for heart disease prediction [7]. Moreover, with the recent advances

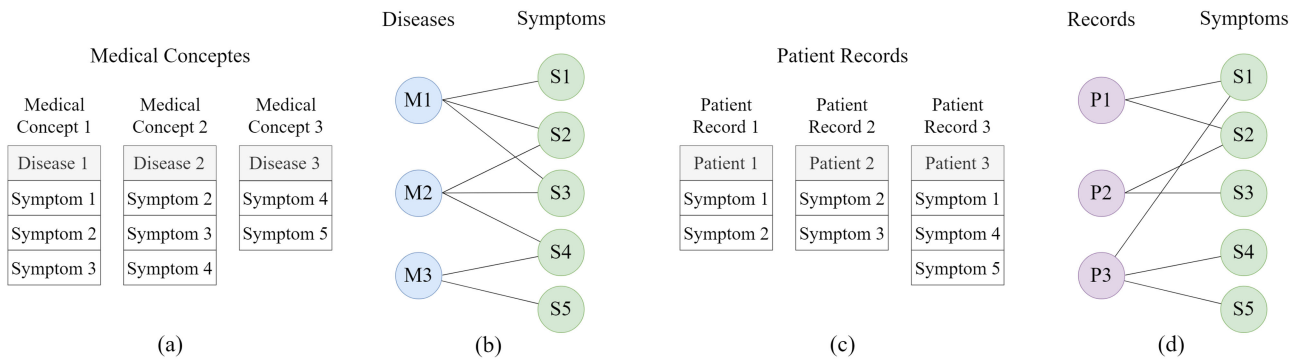

Fig. 1. Building the medical concept graph and patient record graph. (a) Shows three medical concepts. Each medical concept consists of one specific disease and several symptoms. (b) Shows the medical concept graph extracted from (a). (c) Shows three patient EMRs. Each EMR consists of the patient's identity and several symptoms. (d) Shows a patient record graph extracted from (c).

in deep attention networks and graph neural networks [20]–[25], there has been an increasing amount of applications in disease prediction tasks. Suo *et al.* [2] used CNNs to perform personalized disease predictions by identifying similar patients based on their historical EMRs. Also, Ma *et al.* [3] incorporated discrete prior medical knowledge into a CNN-based model to improve the prediction performance. Unfortunately, existing models mainly focus on general diseases with inherently high volume of relevant EMRs for training, and cannot adapt to cases of rare diseases due to higher data scarcity and more intricate relationships between symptoms and diagnoses.

Finding the right treatments for patients is a primary benefit of accurate disease prediction results, but rare diseases have been rather difficult to be identified among a large number of possible diagnoses. In the context of rare disease prediction, machine learning techniques have recently started to demonstrate more advantageous performance in terms of analyzing the latent patterns within EMRs. For example, Garg *et al.* [26] targeted a specific rare disease called cardiac amyloidosis and successfully automated the process of identifying potential patients with bootstrap machine learning algorithms. Along this line of research, MacLeod *et al.* [27] used self-reported behavioral data to distinguish people with rare diseases from people with more common chromic illnesses, while Hare *et al.* [28] used pattern recognition ensembles to improve the accuracy of identified rare disease patients. Additionally, genomic data was studied in [29], where a method adopting imbalance-aware learning strategies with a resampling algorithm was proposed for predicting rare and common diseases. In general, most studies make predictions based on longitudinal historical patient records, which means that they can only serve patients whose historical EMRs are used for model training.

III. The Proposed Method

A. Definitions

Definition 1 (Medical Concept Graph): For an arbitrary disease, its associated medical concepts include its name, diagnostic symptoms, and the category it belongs to. As the available medical concepts vary in different public medical knowledge bases, without loss of generality, we only consider a common concept, i.e., symptoms in this paper. We represent these medical concepts as a medical concept graph, denoted by $\mathcal{C} = (\mathcal{V}_M \cup \mathcal{V}_S, \mathcal{E}_{MS})$, where $\mathcal{V}_M$ is the node set of diseases, and $\mathcal{V}_S$ is the node set for symptoms extracted from medical knowledge bases, and $\mathcal{E}_M$ is the edge set. If a symptom $s \in \mathcal{V}_S$ is associated to $m \in \mathcal{V}_M$, then there is an edge between two types of nodes. The construction process of the medical concept graph is depicted in Fig. 1(a)–(b). Each disease $m \in \mathcal{V}_M$ has a $|\mathcal{V}_M|$-dimensional one-hot encoding $\mathbf{e}_m = \{0,1\}^{|\mathcal{V}_M|}$ with 1 at the m-th position as its unique indentifier. At the same time, each disease also has a binary label $r_m \in \{0,1\}$ indicating whether it is a rare disease or not ($r_m = 1$ if true and vice versa).

Definition 2 (Patient Record Graph): Similar to the medical concept graph, the key information for disease prediction can be represented as a patient record graph denoted by $\mathcal{P} = (\mathcal{V}_P \cup \mathcal{V}_S, \mathcal{E}_{PS})$. $\mathcal{P}$ is a graph-structured representation of EMRs, where $\mathcal{V}_P$ is the node set of all patients and $\mathcal{V}_S'$ contains all the symptoms occurred in EMRs ($\mathcal{V}_S' \subseteq \mathcal{V}_S$ in our case), and $\mathcal{E}_{PS}$ represents all observed edges between patient nodes and symptom nodes. An example of a patient record graph is illustrated in Fig. 1(c)–(d). Each record $p \in \mathcal{V}_P$ is assigned a multi-hot encoding $\mathbf{c}_p \in \{0,1\}^{|\mathcal{V}_M|}$ indicating the diseases this patient has (corresponding indexes are marked as 1).

Problem 1 ((Rare) Disease Prediction): Given a medical concept graph $\mathcal{C} = (\mathcal{V}_M \cup \mathcal{V}_S, \mathcal{E}_{MS})$, a patient record graph $\mathcal{P} = (\mathcal{V}_P \cup \mathcal{V}_S, \mathcal{E}_{PS})$ and the corresponding labels, our goal is to learn a Graph Convolutional Network-based model, which is able to predict the diseases for each new patient $p \notin \mathcal{V}_P$. Apart from general disease prediction, we will additionally perform rare disease prediction by evaluating the prediction performance exclusively on patients who are diagnosed with rare diseases in the real clinical dataset.

B. Neural Graph Encoder

We use Fig. 2 to demonstrate the workflow of our proposed framework for disease prediction. Our model takes the medical concept graph $\mathcal{C}$ and patient record graph $\mathcal{P}$ as its input, then embeds every node in both graphs by aggregating the information from their sampled neighbors. Eventually, for a given patient, we can form a vector representation by fusing the

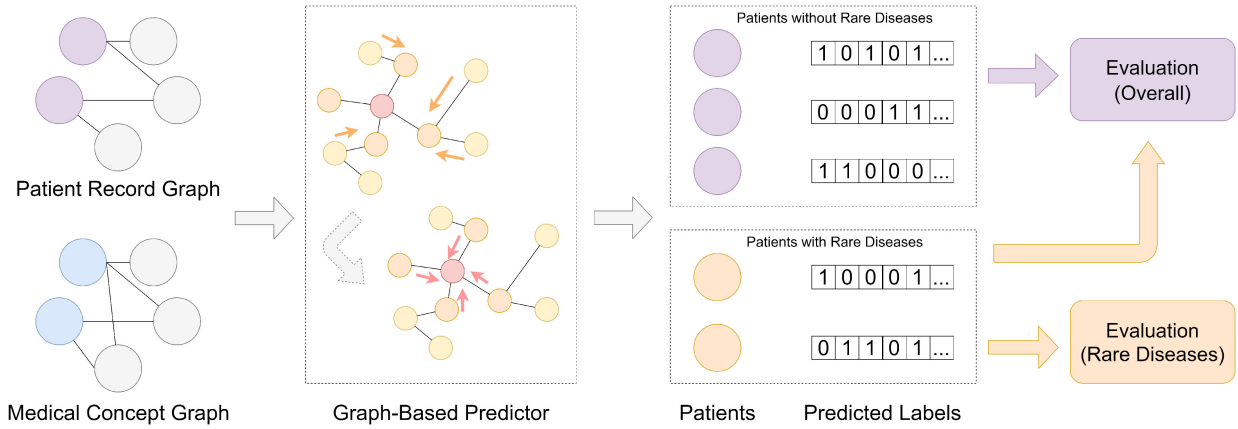

Fig. 2. The workflow of our graph neural network-based model for disease prediction.

learned embeddings of symptoms described in her/his EMRs. Then, by measuring the closeness between the embeddings of the patient and any disease, we can eventually estimate the likelihood of diagnosing patient p with disease m. In this section, we first describe a graph encoder model, which is responsible for producing a low-dimensional embedding $\mathbf{z} \in \mathbb{R}^d$ for each node in an arbitrary graph.

Inspired by recent advances in graph convolutional networks [11], we define the calculation of embeddings of each node (a disease, a symptom, or a patient) via an aggregation scheme on the features of its directly connected neighbors. The defined aggregation function takes into account the first-order neighbors of a node and applies the same transformation across all nodes in the graph. In this way, each node in the graph defines its own filed of computational input, but different nodes' computational procedures reuse the same set of parameters that define how information is shared and propagated. This setting makes efficient use of information shared across regions in the graph, and allows embeddings to be generated for previously unseen nodes during training, e.g., a newly joined patient in the EMR dataset.

To begin with, for a graph $\mathcal{G} = \{\mathcal{C}, \mathcal{P}\}$, we uniformly represent a disease, symptom, or patient node as $v \in \mathcal{G}$ to be succinct. Then, at the l-th information propagation layer, the embedding $\mathbf{h}_v^l$ of node v is calculated as:

$$\begin{aligned} \mathbf{h}_{\mathcal{N}(v)}^l &= AGGREGATE(\{\mathbf{h}_{v'}^{l-1}, \forall v' \in \mathcal{N}(v)\}) \\ \mathbf{h}_v^l &= \sigma(\mathbf{W}^l \cdot [\mathbf{h}_v^{l-1}; \mathbf{h}_{\mathcal{N}(v)}^l]) \end{aligned} \tag{1}$$

where $\mathbf{W}^l$ is the weight matrix to be learned at the l-th layer, $\mathbf{h}_v^{l-1}$ is node v's embedding at the previous layer, and we denote the total layer size as L. We use $[\cdot\,;\cdot]$ to represent the concatenation of two vectors, and use $\mathcal{N}(v)$ to denote the set of evenly sampled neighbor nodes of v. Note that for $l = 0$, the node embedding $\mathbf{h}_v^0 \in \mathbb{R}^d$ is initialized via either randmonized values or side information from the data (subject to availability). For instance, given a patient node, with the available patient demographics and medical profiles in the EMR data, then $\mathbf{h}_v^0$ will be initialized as a real-valued dense feature vector, and each digit in $\mathbf{h}_v^0$ represents the observed value of a feature dimension (e.g., age). $\mathbf{h}_{\mathcal{N}(v)}^l$ is the synergic representation resulted from the aggregation function, which is designed to aggregate the embeddings of node v's neighbors at the $(l-1)$-th layer. σ is a non-linear activation function (e.g., tanh), and the aggregator can be chosen as mean, max pooling, RNNs, etc. By default, we deploy $mean(\cdot)$ in our model for information aggregation.

Then, we take a normalization step before reaching the final embedding for all nodes at the last layer L:

$$\mathbf{h}_v = \frac{\mathbf{h}_v^L}{||\mathbf{h}_v^L||_2}, \ \forall v \in \mathcal{G} \tag{2}$$

In this paper, it is worth mentioning that the representations learned for the symptom nodes p in both the medical concept graph $\mathcal{C}$ and patient record graph $\mathcal{P}$ share the same embedding space. That is to say, for each symptom p, its embeddings remain the same in both graphs, thus serving as an effective bridge between patient and disease nodes from different graphs. Meanwhile, as different types of nodes, i.e., diseases, patients, and symptoms are learned from three separate embedding spaces, we further align their contexts by projecting all node embeddings onto the same space, followed by a non-linear activation:

$$\mathbf{z}_v = \sigma(\mathbf{W}\mathbf{h}_v), \ \forall v \in \mathcal{G} \tag{3}$$

where $\mathbf{W}$ is the learnable projection weight, and $\mathbf{z}_v$ is the final embedding for node v.

C. Varying Graph Encoder Kernels

To fully investigate the effectiveness of different neural architectures in disease prediction, we introduce two variants of the graph encoder. In this section, we replace the graph encoder kernels described in Eq. (1) with two widely-used GNNs. We adopt two encoder architectures: the Graph Attention Network [12] and the Graph Isomorphic Network [13]. Both variants also follow the paradigm of neighborhood-based information aggregation, but each encoder employs a specific message passing rule focusing on different nuances of the graph structural information.

Graph Attention Networks (GATs): GATs apply attention mechanisms to selectively encode the information from neighbors according to their importance to the target node v. This is achieved by taking a weighted sum of the representations of all

v's neighbor nodes:

$$\mathbf{h}_v^l = \sum_{v' \in \mathcal{N}(v)} \alpha_{v'v} \mathbf{M} \mathbf{h}_{v'}^{l-1} \tag{4}$$

where $\mathbf{M}$ is the transformation weight matrix, and $\alpha_{v'v}$ is the attentive weights indicating the importance of neighbor node $v' \in \mathcal{N}(v)$ when calculating $\mathbf{h}_v^l$. Each $a_{v'v}$ is computed via the following attention network:

$$\alpha_{v'v} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{N}\mathbf{h}_v^{l-1} || \mathbf{N}\mathbf{h}_{v'}^{l-1}]))}{\sum_{k \in \mathcal{N}(v)} \exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{N}\mathbf{h}_v^{l-1} || \mathbf{N}\mathbf{h}_k^{l-1}]))} \tag{5}$$

with a projection vector $\mathbf{a}$ and the weight matrix $\mathbf{N}$. Essentially, the learned attentive weights allows the aggregator to lay more emphasis on neighbor nodes having more contributions to the message passing process, thus being able to generate highly expressive node embeddings.

Graph Isomorphic Networks (GINs): GINs are claimed effective in representing isomorphic and non-isomorphic graphs with discrete attributes. The computation process of its l-th layer is defined as:

$$\mathbf{h}_v^l = MLP\left(\sum_{u \in \mathcal{N}(v)} \mathbf{h}_u^{l-1}\right) \tag{6}$$

where MLP is a multi-layer perceptron. In contrast to other graph encoder kernels which combine the information from both node v itself and its neighbors, a GIN-based graph encoder forms the embedding for v purely based on the embeddings of neighbor nodes.

D. Graph Decoder for Disease Prediction

Disease Prediction for Patients: The purpose of the graph decoder in our model is to translate the information contained in the symptom, disease, and patient node embeddings into predictions of possible diseases associated to a given patient. In summary, given the embedding $\mathbf{z}_p$ of a patient p, the graph decoder maps it to a vectorized output $\hat{\mathbf{c}}_p \in \{0,1\}^{|\mathcal{C}|}$ which approximates this patient's multi-hot disease label $\mathbf{c}_p \in \{0,1\}^{|\mathcal{C}|}$ (i.e., the ground truth in patient record graph $\mathcal{P}$). Specifically, in this decoding process, every element in $\hat{\mathbf{c}}_p$ is computed via:

$$\hat{\mathbf{c}}_{p,n} = \text{sigmoid}(\mathbf{z}_p^T \mathbf{Q}_n), \ \forall m_n \in \mathcal{V}_M \tag{7}$$

where $\hat{\mathbf{c}}_{p,n}$ is the n-th element in $\hat{\mathbf{c}}_p$, while $n \leq |\mathcal{V}_M|$ is used for indexing all the diseases m. The closer $\hat{\mathbf{c}}_{p,n} \in \hat{\mathbf{c}}_p$ is to 1, the more likely patient p is diagnosed with disease m_n. $\mathbf{Q} \in \mathbb{R}^{|\mathcal{V}_M| \times d}$ carries the corresponding regression weights for all diseases, and $\mathbf{Q}_n$ is the n-th column of it. To train our model, we quantify the prediction error via the following negative log likelihood loss function:

$$\mathcal{L}_{\mathcal{P}} = -\sum_{n=1}^{|\mathcal{V}_M|} \mathbf{c}_{p,n} \log(\hat{\mathbf{c}}_{p,n}) \tag{8}$$

Handling New Patients: With a fully trained neural graph encoder, the message passing schema and the latent correlations between diseases and symptoms are uncovered. However, compared with existing patient nodes in $\mathcal{P}$, before we can predict the diseases for a newly joined patient p, we need to first infer her/his embedding vector $\mathbf{z}_p$. In this regard, our approach shows its advantage in inductively generating node representations. The intuition is that the knowledge mined from the medical concept graph $\mathcal{C}$ about all symptom nodes is stored in the learned parameters (e.g., weight matrices), which can be additionally applied to a newly arrived patient node connecting to several well-represented symptom nodes. In short, based on the symptoms reported in a new patient's EMRs, our model can effectively produce an expressive representation for the patient. To be specific, with the weight matrices $\{\mathbf{W}^l\}_{l=1}^L$ for the aggregation function at each layer, we consolidate the trained aggregators and apply them to the newly added patients. One of the crucial features of the defined aggregation scheme is that the calculation of embeddings of a node only relies on its first-order neighbors, making the embeddings of a newly added patient easily computable by aggregating the embeddings of its neighbor symptoms according to the graph encoder defined in Eq. (1).

Supplementary Node Classification Task: To thoroughly learn the embeddings and network parameters, we leverage the available disease label information to design a supplementary task of node classification. Specifically, with the embedding $\mathbf{z}_m$ for each disease m, we decode the latent representation to approximate its one-hot label $\mathbf{c}_m$ defined in the medical concept graph $\mathcal{C}$. Similar to Eq. (7), the estimated label $\hat{\mathbf{c}}_m$ of disease identity for a given embedding $\mathbf{z}_m$ is as follows:

$$\hat{\mathbf{c}}_{m,n} = \text{sigmoid}(\mathbf{z}_m^T \mathbf{G}_n), \ \forall m_n \in \mathcal{V}_M \tag{9}$$

where $\mathbf{G} \in \mathbb{R}^{d \times |\mathcal{V}_M|}$ is a trainable weight matrix, and $\mathbf{G}_n$ is the n-th column of $\mathbf{G}$. We define the following negative log likelihood for this supplementary node classification task:

$$\mathcal{L}_{\mathcal{M}} = -\sum_{n=1}^{|\mathcal{V}_M|} \mathbf{c}_{m,n} \log(\hat{\mathbf{c}}_{m,n}) \tag{10}$$

Loss Function: To retain both the external knowledge in the medical concept graph and patients diagnostic information in the patient record graph, we define the combined loss function as the following:

$$\mathcal{L} = \mathcal{L}_{\mathcal{M}} + \mathcal{L}_{\mathcal{P}} \tag{11}$$

which can be easily optimized via Stochastic Gradient Descent (SGD) algorithms.

IV. Experiments

In what follows, we present detailed experimental analysis on our proposed disease prediction model. To support easy implementation, we have released the source code of our model at: https://github.com/zhchs/Disease-Prediction-via-GCN.

TABLE I
MAJOR STATISTICS OF THE EMR DATASET

number of patients	806
number of patients with rare diseases	451
number of rare diseases	71
number of symptoms	131
average number of diseases	1.49
maximum number of diseases	5

A. Datasets

In our experiments, we utilize the **Proprietary EMR** dataset for constructing the patient record graph, which is our private real-world patient clinical record dataset collected from local hospitals. It contains 806 patients, while 451 among them were diagnosed with at least one rare disease. Each patient has an average of 1.49 diagnosed diseases. The main statistics of the Proprietary EMR dataset are shown in Table I.

Besides, to formulate external medical knowledge into the medical concept graph, we choose **Human Phenotype Ontology** (HPO) in our experiments. HPO provides an ontology of medically relevant phenotypes, disease-phenotype annotations, and the corresponding algorithms. The HPO is mainly used for computational deep phenotyping, precision medicine, as well as the integration of clinical data into translational research [30]. It currently contains over 13,000 terms arranged in a directed acyclic graph and are connected by "is-a" (subclass-of) edges, such that a term represents a more specific or limited instance of its parent term(s). The annotation file of the HPO contains manual and semi-automated annotations of *OMIM*, *Orphanet*, and *DECIPHER* entries. Here we mark the diseases annotated in the Orphanet database as rare diseases. We extract 71 diseases with Orphanet annotation and 669 linked phenotypes for our experiments.

B. Baseline Methods

We compare our model with three well-established classifiers, namely **Support Vector Machine** (SVM) [31], **Decision Tree** (DT) [4] and **Random Forest** (RF) [32], as well as four state-of-the-art graph embedding-based models, which are introduced below.

DeepWalk [33] is an approach for learning latent node representations in a graph. It learns node embeddings by maximizing the co-occurrence probability of nodes on the sequences generated by random walks.

LINE [34] is a graph embedding method that is well suited to heterogeneous graphs. It has an objective function that preserves both the first-order and second-order proximities.

SDNE [35] is able to map graph-structured data to a highly non-linear latent space to preserve the both the global and local network structures and is robust to data sparsity.

Struc2Vec [36] uses a hierarchy to measure node similarities at different scales and constructs a multi-layer graph to encode structural contexts into node embeddings.

C. Experimental Settings and Evaluation Protocols

We evaluate our model in terms of performance on both the general disease prediction and rare disease prediction. For our model, we set the learning rate and batch size respectively to 0.3 and 200, the aggregator layer size $L = 1$. The dimensions of initialization (i.e., $\mathbf{h}_v^0$) and output ($\mathbf{h}_v^L$) embeddings are set to 10,000 and 1,000, respectively. For each node, we uniformly sample 5 of its neighbors for information aggregation (i.e., $|\mathcal{N}(v)| = 5$). In classic baselines, i.e., SVM, DT and RF, we take symptoms in each patient's EMRs as the input features for classification. In all graph embedding-based peer methods, we transform the EMRs into a patient record graph as their input. We randomly split the patients in the Proprietary EMR dataset with a ratio of 7:3 for training and evaluation, respectively.

We utilize three widely-used metrics, namely **recall**, **precision**, and **F1 score** of the top-K diseases in each patient's prediction result. In short, recall reflects how accurately a model can predict the right disease for a patient, precision indicates how well a model distinguishes the true diseases from the false ones, while F1 is the trade-off between two terms by taking the harmonic mean of recall and precision. Here we choose $K = \{1, 2, 3, 4, 5\}$ based on the average and maximum number of diagnosed diseases in the dataset.

D. Overall Prediction Effectiveness

We showcase all models' prediction results on the general disease prediction task (i.e., considering all diseases in the dataset) in Table II. Apparently, in most cases ($K = 2, 3, 4, 5$), our model outperforms all state-of-the-art baselines by a significant margin. This verifies our model's effectiveness in inductively representing a disease by aggregating the learned representations of its neighbor nodes (i.e., diseases). With an increasing K, the general trend shows an increasing recall and a decreasing precision, whilst all models achieved the highest F1 score when $K = 2$ or $K = 3$. It is because that, when more possible diseases are predicted, more actually diagnosed diseases will be covered in the result, but the percentage of correct results also drops in the prediction.

In the case of $K = 1$, the SVM performs the best. However, as each patient is diagnosed with an average of more than one disease, the setting of generating only one prediction for each patient can be a mismatch for the real situations and misses important opportunities for identifying other diseases. Furthermore, in real-world scenarios, it is more practical and realistic if a disease prediction model can provide a few possible results to assist the doctors with accurate diagnoses.

When $K > 1$, SVM, DT and RF obtain similar results with F1 scores over 0.41 when $K = 2$ and around 0.37 when $K = 3$. For the graph embedding-based models, SDNE performs the best among other graph embedding-based models and it is comparable to our model. As an extension to the LINE model, SDNE focuses more on the first-order and second-order proximity between nodes, making it able to learn highly representative node embeddings. Also, Struc2Vec did not perform as good as other models, and it is possibly due to the fact that Struc2Vec mainly models the structural similarity between the nodes instead of the neighbor nodes' features, which are rather important when inferring the representation of a disease from its related symptoms.

TABLE II
OVERALL PREDICTION PERFORMANCE ON ALL DISEASES. ENTRIES IN BOLD FACE ARE THE BEST RESULTS

Method	K=1			K=2			K=3			K=4			K=5		
	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision
SVM	**0.430**	**0.373**	**0.508**	0.428	0.507	0.370	0.370	0.560	0.277	0.355	0.656	0.244	0.321	0.703	0.208
DT	0.393	0.348	0.450	0.410	0.506	0.345	0.362	0.569	0.266	0.302	0.580	0.205	0.271	0.617	0.174
RF	0.414	0.361	0.483	0.427	0.517	0.364	0.386	0.596	0.285	0.343	0.645	0.233	0.299	0.671	0.193
DeepWalk	0.343	0.289	0.421	0.379	0.443	0.331	0.368	0.565	0.273	0.344	0.631	0.237	0.299	0.653	0.194
LINE	0.345	0.290	0.426	0.404	0.473	0.353	0.378	0.580	0.281	0.354	0.652	0.243	0.314	0.678	0.204
SDNE	0.409	0.350	0.492	0.440	0.518	0.382	0.423	0.638	0.317	0.372	0.687	0.255	0.337	0.732	0.219
Struc2Vec	0.284	0.245	0.339	0.284	0.341	0.244	0.266	0.409	0.197	0.240	0.451	0.163	0.246	0.547	0.159
Ours	0.406	0.347	0.488	**0.457**	**0.547**	**0.393**	**0.427**	**0.648**	**0.318**	**0.379**	**0.702**	**0.259**	**0.346**	**0.759**	**0.224**

TABLE III
PREDICTION PERFORMANCE ON RARE DISEASES. ENTRIES IN BOLD FACE ARE THE BEST RESULTS

Method	K=1			K=2			K=3			K=4			K=5		
	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision
SVM	0.332	0.271	0.430	0.344	0.377	0.316	0.303	0.419	0.237	0.337	0.571	0.239	0.306	0.618	0.204
DT	0.374	0.321	0.447	0.410	0.486	0.355	0.386	0.577	0.289	0.328	0.601	0.226	0.286	0.618	0.186
RF	**0.388**	**0.329**	**0.474**	0.433	0.499	0.382	0.404	0.591	0.307	0.343	0.621	0.237	0.318	0.670	0.209
DeepWalk	0.196	0.148	0.289	0.336	0.364	0.311	0.330	0.470	0.254	0.308	0.520	0.219	0.273	0.546	0.182
LINE	0.172	0.126	0.272	0.360	0.392	0.333	0.348	0.493	0.269	0.313	0.531	0.221	0.289	0.575	0.193
SDNE	0.264	0.209	0.360	0.383	0.421	0.351	0.388	0.543	0.301	0.341	0.584	0.241	0.318	0.637	0.212
Struc2Vec	0.123	0.092	0.184	0.184	0.200	0.171	0.182	0.248	0.143	0.182	0.307	0.129	0.199	0.393	0.133
Ours	0.329	0.267	0.430	**0.442**	**0.503**	**0.395**	**0.408**	**0.578**	**0.316**	**0.375**	**0.652**	**0.263**	**0.336**	**0.681**	**0.223**

It is worth noting that our model can scale up to more complex EMR datasets with heterogeneous information. As symptoms and diagnosed diseases are two fundamental types of clinical information available in almost all EMRs, our proposed model is compatible and can easily generalize to other EMR datasets containing such information. Furthermore, as described in Section III-B, by transforming auxiliary side information into initial node features, our model can fully incorporate the available knowledge from complex EMR data to achieve optimal performance.

E. Rare Disease Prediction

In this section, we exclusively evaluate the prediction performance on rare diseases. We apply the same training settings described in Section IV-C and only make predictions on patients diagnosed with at least on rare disease in the test set. The performance of all models in this task is reported in Table III. Similar to the general disease prediction task, our model outperforms all baselines in this task for $K > 1$. Also, when $K = 2$ or $K = 3$, every model reaches its highest F1 score. Generally, the results indicate that our model can capture the latent relationship between symptoms and diseases to better distinguish the types of rare diseases.

Compared with the overall performance, the results of most models in rare disease prediction slightly drops. It is reasonable because the information and number regarding rare diseases and diagnosed patients are insufficient for models to thoroughly learn the patterns. In terms of $K = 1$, the traditional methods like the RF method yields better performance than all graph embedding-based models, including ours. Though graph embedding-based method shows advantageous performance when $K > 1$ on general disease prediction, it is interesting that the traditional machine learning methods can perform better on rare disease prediction when K grows, especially when $K = 2$. One possible reason might be that the graphs constructed from our EMR dataset is not large enough for models to capture the relation between nodes. In this situation, directly using the symptoms as the feature vectors and utilize traditional classifiers can already provide good prediction performance.

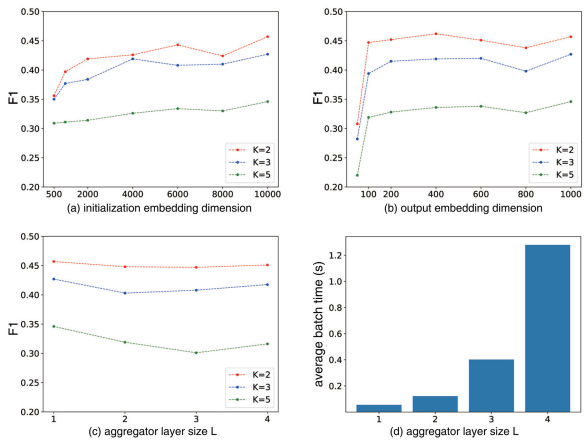

Fig. 3. Results obtained with: (a) different initial embedding dimensions; (b) different output embedding dimensions; and (c) different aggregator layer sizes. (d) The average batch time cost with different aggregator layer sizes.

F. Impact of Hyperparameters

We further study our model's sensitivity to different values of key hyperparameters, namely the initialization embedding dimension of $\mathbf{h}_v^0$, the output embedding dimension of $\mathbf{z}_v$ (i.e., d), and the aggregator layer size L. Specifically, we test the overall disease prediction performance with different hyperparameters on the full EMR dataset following the evaluation protocols as in Section IV-C. We report the F1 scores with $K = 2, 3, 5$ for demonstration.

Initialization Embedding Dimension: In this test, we vary the initialization embedding dimension in $\{500, 1{,}000, 2{,}000, 4{,}000, 6{,}000, 8{,}000, 10{,}000\}$ and record the performance of our model accordingly in Fig. 3(a). In

TABLE IV
PREDICTION PERFORMANCE WITH DIFFERENT GRAPH ENCODER KERNELS. ENTRIES IN BOLD FACE ARE THE BEST RESULTS

Kernel	K=1			K=2			K=3			K=4			K=5		
	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision
GAT (overall)	0.394	0.330	0.488	0.411	0.479	0.360	0.385	0.574	0.289	0.352	0.646	0.242	0.310	0.672	0.202
GIN (overall)	0.274	0.233	0.331	0.308	0.362	0.269	0.276	0.423	0.205	0.270	0.515	0.183	0.238	0.531	0.154
Ours (overall)	**0.406**	**0.347**	**0.488**	**0.457**	**0.547**	**0.393**	**0.427**	**0.648**	**0.318**	**0.379**	**0.702**	**0.259**	**0.346**	**0.759**	**0.224**
GAT (rare)	0.287	0.223	0.404	0.361	0.394	0.333	0.352	0.491	0.275	0.346	0.606	0.242	0.292	0.582	0.195
GIN (rare)	0.131	0.102	0.184	0.206	0.221	0.193	0.204	0.289	0.158	0.191	0.348	0.132	0.197	0.406	0.130
Ours (rare)	**0.329**	**0.267**	**0.430**	**0.442**	**0.503**	**0.395**	**0.408**	**0.578**	**0.316**	**0.375**	**0.652**	**0.263**	**0.336**	**0.681**	**0.223**

general, the higher the dimension is, the better performance will be achieved. The most obvious performance gain from higher initialization embedding dimension is observed at $K = 2$ and $K = 3$. When the dimension exceeds 2,000, the performance of our model becomes stable. Notably, though the values of node embedding vectors are randomly initialized, it is the carrier of the graph topology structure information, i.e., the crucial disease-symptom relationships, and higher embedding dimension means that more latent structural information is passed into our model.

Output Embedding Dimension: We also explore the effect of different out embedding dimensions from the graph encoder. The embedding dimension of the encoder is adjusted in $\{50, 100, 200, 400, 600, 800, 1{,}000\}$, and Fig. 3(b) shows corresponding results. A dramatical performance boost appears as the embedding dimension increases from 50 to 100. We can observe that when the dimension of embedding is over 100, the performance increase becomes negligible, which indicates that our proposed model can preserve the node and structural information well for disease prediction with a relatively compact output dimension.

Aggregator Layer Size L: Our model's performance with layer size $L \in \{1, 2, 3, 4\}$ is reported in Fig. 3(c). The embedding dimension of each layer is 1,000 and the number of sampled neighbors is 5. As increasing the layer size directly affects the running time of our model, we also report the time cost per batch in Fig. 3(d). On one hand, the results show that adding extra deep layers in our graph neural network does not incur further performance gain. As our constructed bipartite graphs contain only patient-disease and disease-symptom links, so setting $L = 1$ is already sufficient for the model to learn representative node embeddings for disease prediction, while propagating more information with additional layers may infuse noise into node embeddings and lead to inferior performance. On the other hand, the time cost increases when L increases from 1 to 4 because it takes more computational steps to generate the embedding for each target node, but our proposed graph neural network is highly efficient with $L = 1$ as the average running time per batch is less than 0.1 s.

G. Comparing Different Graph Encoder Kernels

In graph neural networks, the selection of an appropriate kernel function for information aggregation is largely associated with the characteristics of the data. As we have described in Section III-C, we utilize two GNN-based variants as the graph encoder kernel in our model. In particular, we choose GAT and GIN as two variant kernels for modelling our medical concept graph and patient record graph, because GAT focuses on the most relevant parts of the input to make decisions [12], while GIN generalizes the Weisfeiler-Lehman test and is hence able to produce discriminative node embeddings [13].

The comparison results are illustrated in Table IV, where we use "overall" and "rare" to mark the prediction results in general and rare disease prediction tasks, respectively. Both GAT and GIN are deployed with a single-layer structure and the same hyperparameters introduced in Section IV-C. In our assumption, utilizing complex kernel methods can improve the model's expressiveness for disease prediction. However, in our experiments, both GAT and GIN kernels are not as good as our default mean aggregator function. Firstly, an obvious performance drop is observed with the GIN kernel. One possible reason is that, GIN neglects the important information from the target node v itself when generating embedding $\mathbf{h}_v^l$. Secondly, GAT kernel shows slightly lower prediction accuracy than our default model, and the cause is largely related to its excessive model parameters and the insufficiency of the available EMR data. GAT introduces substantially more parameters, making it prone to overfitting and require a lot more training data and iterations to effectively optimize all model parameters and thoroughly capture the complex patterns within the data. In contrast, our model only aggregates the information of first-order neighbors, so it can directly learn the disease-symptom relationships to ensure better performance.

V. CONCLUSION

In this work, we present a GNN-based model for disease prediction with EMRs, which novelly leverages the external graph-structured medical knowledge to learn the latent node embeddings, thus enabling accurate disease prediction for new patients under sparse training data in an inductive manner. The experimental results on our real-world EMR dataset shows promising effectiveness of our proposed model, especially in multi-label disease classification settings. To conclude, our model offers an intuitive yet accurate solution to disease prediction, tackling the data scarcity problem and the hardship in diagnosing rare diseases at the same time.

REFERENCES

[1] R. Hillestad *et al.*, "Can electronic medical record systems transform health care? potential health benefits, savings, and costs," *Health Affairs*, vol. 24, no. 5, pp. 1103–1117, 2005.
[2] Q. Suo *et al.*, "Personalized disease prediction using a CNN-based similarity learning method," in *Proc. IEEE Int. Conf. Bioinformatics Biomedicine*, 2017, pp. 811–816.
[3] 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. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 1910–1919.
[4] J. Quinlan, "Simplifying decision trees," *Int. J. Man-Mach. Stud.*, vol. 27, no. 3, pp. 221–234, 1987.

[5] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.
[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[7] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appli.*, 2008, pp. 108–115.
[8] R. C. Griggs *et al.*, "Clinical research for rare disease: Opportunities, challenges, and solutions," *Molecular Genetics Metabolism*, vol. 96, no. 1, pp. 20–26, 2009.
[9] H. J. Dawkins *et al.*, "Progress in rare diseases research 2010–2016: An IRDIRC perspective," *Clin. Translational Sci.*, vol. 11, no. 1, pp. 11–20, 2018.
[10] S. N. Wakap *et al.*, "Estimating cumulative point prevalence of rare diseases: Analysis of the orphanet database," *Eur. J. Human Genetics*, vol. 28, no. 2, pp. 165–173, 2020.
[11] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1025–1035.
[12] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2019.
[13] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *Proc. Int. Conf. Learn. Representations*, 2019.
[14] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Rev. Genetics*, vol. 13, no. 6, pp. 395–405, 2012.
[15] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: graph-based attention model for healthcare representation learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 787–795.
[16] 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. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3504–3512.
[17] 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. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 1903–1911.
[18] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "KAME: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proc. Conf. Inf. Knowl. Manage.*, 2018, pp. 743–752.
[19] X. S. Zhang, F. Tang, H. H. Dodge, J. Zhou, and F. Wang, "MetaPred: Meta-learning for clinical risk prediction with limited patient electronic health records," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 2487–2495.
[20] H. Chen, H. Yin, W. Wang, H. Wang, Q. V. H. Nguyen, and X. Li, "PME: Projected metric embedding on heterogeneous networks for link prediction," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 1177–1186.
[21] Y. Wang, H. Yin, H. Chen, T. Wo, J. Xu, and K. Zheng, "Origin-destination matrix prediction via graph convolution: A new perspective of passenger demand modeling," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 1227–1235.
[22] T. Chen, H. Yin, H. Chen, R. Yan, Q. V. H. Nguyen, and X. Li, "AIR: Attentional intention-aware recommender systems," in *Proc. Int. Council Open Distance Educ.*, 2019, pp. 304–315.
[23] H. Chen, H. Yin, T. Chen, W. Wang, X. Li, and X. Hu, "Social boosted recommendation with folded bipartite network embedding," *IEEE Trans. Knowl. Data Eng.*, to be published, doi: 10.1109/TKDE.2020.2982878.
[24] H. Chen, H. Yin, T. Chen, Q. V. H. Nguyen, W.-C. Peng, and X. Li, "Exploiting centrality information with graph convolutions for network representation learning," in *Proc. Int. Council Open Distance Educ.*, 2019, pp. 590–601.
[25] H. Chen, H. Yin, X. Sun, T. Chen, B. Gabrys, and K. Musial, "Multi-level graph convolutional networks for cross-platform anchor link prediction," 2020, *arXiv:2006.01963*.
[26] R. P. Garg, S. Dong, S. J. Shah, and S. R. Jonnalagadda, "A bootstrap machine learning approach to identify rare disease patients from electronic health records," 2016, *arXiv:1609.01586*.
[27] H. MacLeod, S. Yang, K. Oakes, K. Connelly, and S. Natarajan, "Identifying rare diseases from behavioural data: A machine learning approach," in *Proc. CHASE*, 2016, pp. 130–139.
[28] T. Hare, P. Sharan, E. J. Kleczyk, and D. Evans, "Improving accuracy in rare disease patient identification using pattern recognition ensembles," *J. Pharmaceutical Manage. Sci. Assoc.*, vol. 6, no. 6, pp. 41–59, 2018.
[29] M. Schubach, M. Re, P. N. Robinson, and G. Valentini, "Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants," *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, 2017.
[30] S. Köhler *et al.*, "The human phenotype ontology in 2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D865–D876, 2016.
[31] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[32] T. K. Ho, "Random decision forests," in *Proc. Int. Conf. Document Anal. Recognit.*, 1995, pp. 278–282.
[33] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 701–710.
[34] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: Large-scale information network embedding," in *Proc. WWW*, 2015, pp. 1067–1077.
[35] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1225–1234.
[36] L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo, "*struc2vec:* Learning node representations from structural identity," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 385–394.