# A Novel Sequence-to-Subgraph Framework for Diagnosis Classification

**Jun Chen**, **Quan Yuan**, **Chao Lu** and **Haifeng Huang**
Baidu Inc, Beijing 100193, China
{chenjun22, yuanquan02, luchao, huanghaifeng}@baidu.com

## Abstract

Text-based diagnosis classification is a critical problem in AI-enabled healthcare studies, which assists clinicians in making correct decision and lowering the rate of diagnostic errors. Previous studies follow the routine of sequence based deep learning models in NLP literature to deal with clinical notes. However, recent studies find that structural information is important in clinical contents that greatly impacts the predictions. In this paper, a novel sequence-to-subgraph framework is introduced to process clinical texts for classification, which changes the paradigm of managing texts. Moreover, a new classification model under the framework is proposed that incorporates subgraph convolutional network and hierarchical diagnostic attentive network to extract the layered structural features of clinical texts. The evaluation conducted on both the real-world English and Chinese datasets shows that the proposed method outperforms the state-of-the-art deep learning based diagnosis classification models.

## 1 Introduction

Diagnosis classification based on clinical notes, e.g. Electronic Medical Record (EMR) documents, is one of the most important but challenging tasks in AI-enabled healthcare studies. There are approximately 12 million adults misdiagnosed each year in the United States [Singh *et al.*, 2014], and simiarly, the misdiagnosis rate in primary healthcare facilities in China is estimated to be 27.8%[1]. Under such circumstances, with the advance of AI, diagnosis classification is expected to provide physicians with appropriate diagnostic suggestions to lower the rate of diagnostic errors. For example, once a junior physician has recorded the illness description of patient in Hospital Information System (HIS), the diagnostic suggestions will be generated and shown to assist this junior physician to give the final diagnosis.

Text-based diagnosis classification is mostly focused on understanding the EMR documents where certificated physicians record patient's illness with free texts, such as chief complaint, history of present illness, past history, etc. Therefore, similar to most of the *text-based* machine learning methods, attempts have been made to solve diagnosis classification with *sequential models* like CNN [Yang *et al.*, 2018; Li and Yu, 2020], GRU [Baumel *et al.*, 2018] and BiLSTM [Vu *et al.*, 2020] on clinical texts. However, unlike open-domain text classification, clinical documents are highly knowledge-based where professional terminology, phrases, medical relations and writing style are frequently used that distinguish themselves from other domains. Thus, external medical knowledge is incorporated in diagnosis classification, e.g. the detailed description of diseases on Wikipedia Pages [Wang *et al.*, 2020a]. Besides unstructured knowledge, recent studies found that the structured relations between diseases and symptoms are important in diagnosis classification [Chen *et al.*, 2020; Yuan *et al.*, 2020]. These methods *shallowly structuralize* the clinical notes by extracting medical entities, mapping them to the predefined disease-symptom bipartite graph from external medical knowledge, and performing flatten attention on the matched nodes towards generating feature representation. However, in real-world EMR documents, especially in long documents like admission notes, the illness of patient is too complicated to be represented only by a flatten pool of medical entities. Instead, the structure of illness is much more sophisticated. For example, a patient may simultaneously suffer from gastroenteritis, hypertension and pneumonia, and the description of these diseases are mixed together. Diagnosis classification results can be inaccurate if symptoms are not distinguished from each other to understand the structure of patient's illness.

To bring out the most of the structural features of clinical texts, in this paper, we propose a novel **Sequence-to-Subgraph** (Seq2Subgraph) framework to *structuralize* the clinical notes, namely EMR documents in later experiments, into subgraphs and combine with structured external knowledge towards diagnosis classification. Seq2Subgraph reforms the stereotype of managing texts with token-based sequential models in traditional NLP studies. Unlike other text classification problems, diagnostic decision is both data-driven and knowledge-driven. The way to incorporate and manage medical knowledge in diagnostic decision is critical. Therefore, under Seq2Subgraph framework, we propose a novel diagnostic model consisting of **S**ubgraph convolutional network and **Hi**erarchical **D**iagnostic **A**ttentive **N**etwork, namely **SHi-**

[1] https://www.cma.org.cn

**DAN** model, to address the diagnosis classification problem. The main contributions of this paper are summarized as:

- A novel sequence-to-subgraph (**Seq2Subgraph**) framework is proposed to transform clinical texts into structured subgraphs to take advantage of structured medical knowledge towards diagnosis classification, which changes the paradigm of processing clinical texts with conventional sequential models.
- Under the framework, we bring forward a novel diagnostic model, called **SHiDAN**, which mainly consists of subgraph convolutional network and hierarchical diagnostic attentive network. SHiDAN structuralizes the clinical texts into densely connected subgraphs based on medical knowledge and combines with layered structural feature extraction towards diagnosis classification.
- Evaluation on both the real-world Chinese and English EMR documents demonstrates that the proposed method outperforms state-of-the-art deep learning based diagnostic models, which validates the effectiveness of the proposed method.

## 2 Related Work

In this section, we briefly discuss the related work of graph convolutional network and hierarchical attention network in *diagnosis classification* studies.

### 2.1 Graph Convolutional Network (GCN)

GCN has been introduced to extract the structural information of data on a graphical structure with convolutions [Yao *et al.*, 2019]. Specifically, the node embeddings are updated by the properties of neighbors. Such technique has been widely applied in the medical domain. Naturally, the representation learning of diseases [Choi *et al.*, 2017] as well as electronic health records [Choi *et al.*, 2020] can take advantage of GCN upon disease hierarchy and other medical relations to obtain effective representations. In the online self-diagnosis problem, Wang [2020b] applies GCN on the user-symptom and the disease-symptom graphs to simultaneously obtain the next suggested question (symptom) to ask patient and inference diagnosis based on the retrieved symptoms. To the best of our knowledge, the GMAN model [Yuan *et al.*, 2020] is the most related work to ours. The GMAN model encodes the embeddings of diseases and symptoms by convolving the features of their neighbors on the knowledge graph, after which, the final feature towards diagnosis predictions is obtained by performing attention mechanism on the updated embeddings. The major difference of the proposed method to [Yuan *et al.*, 2020] is the clustering of subgraphs and subgraph convolutional network as well as the hierarchical attention network.

### 2.2 Hierarchical Attention Network (HAN)

Different from the single-layer attention for automatic diagnosis [Mullenbach *et al.*, 2018; Yuan *et al.*, 2020], HAN processes data with multi-level attention mechanism in a bottom-up manner from the fine-grained level to the coarse-grained level. Yang [2016] introduces HAN in text classification where sentence is represented by the word-level attention and document is further represented by the sentence-level attention. In medical studies, HAN is utilized in the same way towards classifying the ICD codes of clinical notes [Baumel *et al.*, 2018]. In the diagnosis classification problem, Sha [2017] applies HAN where the entity-level attention is used to get the representation of a particular visit, while the visit-level attention is used to obtain the representation of a particular patient, which is further used in diagnosis classification. The novelty of the proposed method is the framework that combines subgraph convolutional network with HAN to extract the layer-wise structural features of patient's illness.

## 3 Problem Definition

This work studies the problem of diagnosis classification based on EMR documents written with free texts. Formally, given the free-texts $\mathcal{X}$ describing the illness of a patient, which is a sequence of words $\{x_1, ..., x_i, ...\}$, output *the most probable diagnosis code* via generating a probability distribution $\mathbf{p}(d|\mathcal{X})$ over the set of all diseases $d \in \mathcal{D}$. For multi-label classification, the problem is to output the per-label probability on each diagnosis code. For input as EMR document, $\mathcal{X}$ is obtained by stringing the major free-text contents together like chief complaint, history of present illness and physical examination. Please note that the problem definition is general, which does not limit the applications of the proposed method in diagnosis classification with EMR documents. Besides, the conventional patient data like gender and age group are auxiliary inputs, which are processed as one-hot features appended with text features in classification.

## 4 The Sequence-to-Subgraph Framework

In this paper, we propose a novel framework, called Sequence-to-Subgraph (Seq2Subgraph), to deal with plain medical texts, which can be described as:

**Definition 4.1** (**Sequence-to-Subgraph**). *Structuralize text sequence by: 1) Extracting critical entities and relations in it, 2) Inducing graph(s) using extracted results in combination with external knowledge, 3) Generating locally and densely connected subgraph(s) via graph clustering, 4) Performing layered representation learning from inside subgraph to that between subgraphs to obtain the ultimate representation of input. The focus of feature representation has been shifted from sequential text features to structural features.*

Seq2Subgraph poses a novel paradigm to study the diagnosis classification problem. Different from the open-domain text classification, diagnosis classification is heavily knowledge-driven where knowledge can be explictly expressed as relations between medical concepts like the causal relations *disease–symptom* as well as the hierarchy of concepts like *respiratory system* (呼吸系统) — *upper respiratory tract* (上呼吸道) — *acute upper respiratory tract infection* (急性上呼吸道感染). Therefore, different from the methods that incorporate unstructured knowledge as free texts like those from Wikipedia Pages in diagnostic models [Wang *et al.*, 2020a], under the Seq2Subgraph framework, we propose a novel structure-aware diagnosis classification model, called **SHiDAN**, that benefits from transforming the

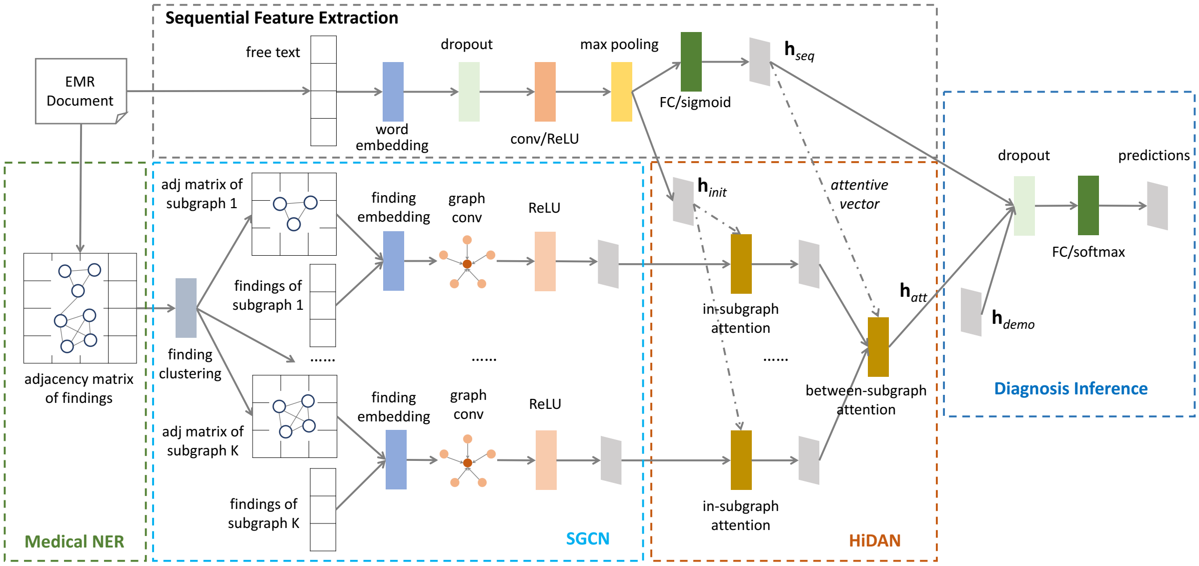


Figure 1: The architecture of the proposed **SHiDAN** model. SHiDAN is composed of sequential feature extraction, structural feature extraction and diagnosis inference. Specifically, medical NER, subgraph convolutional network (SGCN) and hierarchical diagnostic attentive network (HiDAN) consist of the structural feature extraction.

plain texts into structural medical features via generating subgraphs to enhance the representation of EMR documents for diagnosis classification.

Figure 1 shows the architecture of **SHiDAN**. Generally, there are two lanes of feature extraction: a conventional *sequential feature extraction* and a novel *structural feature extraction*, which lead to the compound representation of EMR document for disease inference. Specifically, structural feature extraction consists of medical NER, SGCN and HiDAN.

## 4.1 Sequential Feature Extraction

The original texts are represented by the sequence of words: $\mathcal{X} = \{x_1, ..., x_i, ...\}$. After embedding, the word sequence is represented by $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_i, ...\}$ where $\mathbf{x}_i \in \mathbb{R}^k$ is the $k$-dimensional embedding of word $x_i$. A dropout layer is stacked after word embedding to reduce the impact of overfitting. Then, the initial feature representation of the original texts is obtained by feeding the sequence into a convolutional layer and a max pooling layer:

$$\mathbf{h}_{init} = MaxPool(Conv(Dropout(\mathbf{X}))) \in \mathbb{R}^m, \qquad (1)$$

where $MaxPool$ is performed kernel-wise and $m$ is the number of CNN kernels. On one hand, $\mathbf{h}_{init}$ is used as an external feature in later *in-subgraph attention* module. On the other hand, $\mathbf{h}_{init}$ is passed through a fully-connected layer to get the ultimate sequential feature representation of texts:

$$\mathbf{h}_{seq} = \sigma(\mathbf{W}^{(0)}\mathbf{h}_{init} + \mathbf{b}^{(0)}) \in \mathbb{R}^m, \qquad (2)$$

where $\sigma$ is the sigmoid function, and $\mathbf{W}^{(0)} \in \mathbb{R}^{m \times m}$ and $\mathbf{b}^{(0)} \in \mathbb{R}^m$ are the trainable parameters.

## 4.2 Structural Feature Extraction

Apart from the previous diagnostic methods that only process EMR documents [Yang *et al.*, 2018; Girardi *et al.*, 2018] or external medical knowledge [Wang *et al.*, 2020a] as unstructured plain texts, recent studies prove that the structural information of medical concepts in the texts, e.g. the causal relations between diseases and symptoms, is important to understand the illness [Yuan *et al.*, 2020; Chen *et al.*, 2020].

The illness of a patient can be described in different forms as plain texts in an EMR, including symptoms, vitals and other important findings in the reports of medical films and laboratory tests, all of which are denoted by *findings* in general in this paper. Findings are caused to be present on patient by diseases. Basically, a disease may cause multiple findings to be present, and alternatively, a finding can be caused to be present by multiple diseases.

### Medical NER

The first step of “structuralizing” plain texts is **N**amed **E**ntity **R**ecognition (NER) which is an important direction of research in Natural Language Processing [Cetoli *et al.*, 2017; Gregoric *et al.*, 2018; Cao *et al.*, 2018; Wu *et al.*, 2018]. As a sub-direction, medical NER [Wu *et al.*, 2018; Wang *et al.*, 2019; Zhao *et al.*, 2019; Dai *et al.*, 2019] extracts from texts the type-aware medical entities such as findings and diseases, e.g. coughing, fever, asthma and lung cancer. Please note that *disease* can also be mentioned in the plain texts of illness description, for instance, the past history of illness. Let $\mathcal{F}$ and $\mathcal{D}$ denote the set of findings and that of diseases (diagnosis), respectively. Let $\mathcal{V} = \mathcal{F} \cup \mathcal{D}$ denote the vocabulary of medical entities in this paper. The extracted entities from medical NER are within the scope of $\mathcal{V}$.

In this study, given original text $\mathcal{X}$, NER extracts a list of entities $\mathcal{V}^{(\mathcal{X})} = \mathcal{F}^{(\mathcal{X})} \cup \mathcal{D}^{(\mathcal{X})} \subseteq \mathcal{V}$, where $\mathcal{F}^{(\mathcal{X})}$ and $\mathcal{D}^{(\mathcal{X})}$ are the findings and the diseases recognized in $\mathcal{X}$. $\forall v \in \mathcal{V}^{(\mathcal{X})}$ can either be a finding entity or a disease entity. Similar to [Chen *et al.*, 2020], only the positive findings that are actual present on patient in $\mathcal{X}$ are preserved while negative findings (denied

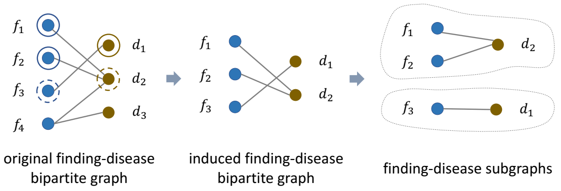

$f_1$
$f_2$
$f_3$
$f_4$
$d_1$
$d_2$
$d_3$
original finding-disease bipartite graph
$f_1$
$f_2$
$f_3$
$d_1$
$d_2$
induced finding-disease bipartite graph
$f_1$
$f_2$
$d_2$
$f_3$
$d_1$
finding-disease subgraphs

Figure 2: The process of finding subgraph generation from a given EMR. The findings and diseases recognized in the given EMR are circled with solid lines in the original finding-disease bipartite graph, and their direct neighbors are circled with dashed lines. The induced finding-disease bipartite graph contain both the recognized findings and diseases as well as their direct neighbors. The finding-disease subgraphs are generated by performing graph clustering on the induced bipartite graph.

on the patient) are removed. Since NER is not the focus of this study, we experiment with the existing NER package in later evaluation.

### Subgraph Convolutional Network (SGCN)

Based on medical NER, there is an easy approach to extract the relations between findings and diseases by simply counting the co-occurrences of findings from texts and diseases from diagnosis in the same EMR documents based on a large EMR corpus. We use the extracted relations where *finding* and *disease* co-occur for more than a threshold of times, e.g. 5, to approximate[2] the causal relations between findings and diseases, where the threshold can be determined with grid search in K-fold cross validation.

Let $\mathcal{E}$ denote the set of all extracted *finding-disease* relations, i.e. $\forall(f,d) \in \mathcal{E}$, $f \in \mathcal{F}$ and $d \in \mathcal{D}$. Then, let $G(\mathcal{V},\mathcal{E})$ denote the *finding-disease* bipartite graph. Let $\mathbf{A} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$ be the adjacency matrix of $G(\mathcal{V},\mathcal{E})$, and we use $\mathbf{A}(i,j) \in \{0,1\}$ to denote the existence of connection between the $i$-th and the $j$-th entities in $\mathcal{V}$.

For a given $\mathcal{V}^{(\mathcal{X})}$, the induced bipartite graph in this paper is defined as: $G(\hat{\mathcal{V}}^{(\mathcal{X})}, \hat{\mathcal{E}}^{(\mathcal{X})})$:

$$\hat{\mathcal{V}}^{(\mathcal{X})} = \mathcal{V}^{(\mathcal{X})} \bigcup_{f\in\mathcal{F}^{(\mathcal{X})}} N(f,G) \bigcup_{d\in\mathcal{D}^{(\mathcal{X})}} N(d,G), \tag{3}$$

$$\hat{\mathcal{E}}^{(\mathcal{X})} = \bigcup_{f\in\hat{\mathcal{V}}^{(\mathcal{X})}, d\in\hat{\mathcal{V}}^{(\mathcal{X})}, (f,d)\in\mathcal{E}} (f,d), \tag{4}$$

where $N(f,G)$ and $N(d,G)$ are the sets of all direct (one-hop) neighbors of $f$ in $G$ and those of $d$ in $G$, respectively.

We perform graph clustering with LPA algorithm[3] [Raghavan *et al.*, 2007] on $G(\hat{\mathcal{V}}^{(\mathcal{X})}, \hat{\mathcal{E}}^{(\mathcal{X})})$ to get a list of subgraphs: $G_1$, ..., $G_K \subseteq G(\hat{\mathcal{V}}^{(\mathcal{X})}, \hat{\mathcal{E}}^{(\mathcal{X})})$. Our intuition is that, there might be different groups of findings because a patient may suffer from multiple diseases at the same time, each of which shows different findings, or one disease may cause symptoms of multiple organs and systems to be present on the patient. Via clustering, the findings are supposed to be distinguished by forming different subgraphs, which are supposed to capture the characteristics of diseases. LPA is an algorithm of community detection that aims at discovering densely connected communities (subgraphs) on a large graph. Figure 2 shows the process of finding subgraph generation.

To extract the structural feature of subgraphs, we propose the **SGCN** to perform graph representation learning. Specifically, let $\mathbf{V} \in \mathbb{R}^{|\mathcal{V}|\times k}$ be the embeddings of all nodes in $\mathcal{V}$, and let $\mathbf{I}_i \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ be the identity matrix w.r.t. the nodes in the $i$-th subgraph $G_i$. We use $\mathbf{V}_i = \mathbf{I}_i\mathbf{V}$ to denote the embedding matrix w.r.t. $G_i$.

Similar to [Yuan *et al.*, 2020; Fu *et al.*, 2019], the node embeddings in the subgraph are updated by convolving the features of their neighbors based on GCN:

$$\hat{\mathbf{V}}_i = ReLU((\mathbf{I}+\mathbf{D})^{-1}(\mathbf{I}+\mathbf{A})\mathbf{V}_i\mathbf{W}^{(1)}), \tag{5}$$

where $\mathbf{A}$ is the adjancency matrix, $\mathbf{D}$ is the degree matrix, $\mathbf{I}$ is the identity matrix and $\mathbf{W}^{(1)} \in \mathbb{R}^{k\times m}$ are the trainable parameters shared by subgraphs. In SGCN, the feature of a node is impacted partially by its direct neighbors as well as partially by itself according to Eq. (5).

### Hierarchical Diagnostic Attentive Network (HiDAN)

Besides the finding-disease relations captured by SGCN, the hierarchy of medical concepts is another important structural feature. Naturally, the output of SGCN makes it possible to automatically understand the illness of patient in a hierarchical bottom-up manner from the individual fine-grained entities to the whole coarse-grained EMR.

Hierarchical attention network [Yang *et al.*, 2016] has been proposed in NLP studies to represent a document from word-level embeddings to sentence-level representations, and upwards to document-level representations, which mimics the process of human to read and understand a document. Similar study [Baumel *et al.*, 2018] has been performed in the ICD coding problem on clinical notes. However, the structure of entities has not been investigated in hierarchical attention network before. Therefore, we propose the **Hi**erarchical **D**iagnostic **A**ttentive **N**etwork (HiDAN) to capture more structural feature of EMR resulting in hierarchical feature representations of *entity-level* → *subgraph-level* → *EMR-level*. Figure 3 illustrates the detail of HiDAN, which consists of two attention layers:

**In-subgraph Attention.** The subgraph-level representations are obtained by aggregating the entity-level representations from SGCN:

$$\mathbf{a}_i = softmax(\hat{\mathbf{V}}_i\mathbf{h}_{init}), \tag{6}$$

$$\mathbf{c}_i = \mathbf{a}_i^\top\hat{\mathbf{V}}_i, \tag{7}$$

where $\mathbf{h}_{init}$ is the intermediate sequential feature of free texts in Eq. (1) and $\mathbf{a}_i \in \mathbb{R}^{|\mathcal{V}|}$ is the per-entity attention weights in this layer. Then, $\mathbf{c}_i$ is the subgraph-level representation of $G_i$.

**Between-subgraph Attention.** Based on the subgraph-level representations of subgraphs $G_1$, ..., $G_K$, a between-

[2] Not strictly the causal relations because the findings of multiple diagnosis may be mixed together in the same EMR. With truncated co-occurrences, the causal relations are statistically approximated.

[3] Codes: https://github.com/zzz24512653/CommunityDetection/tree/master/algorithm

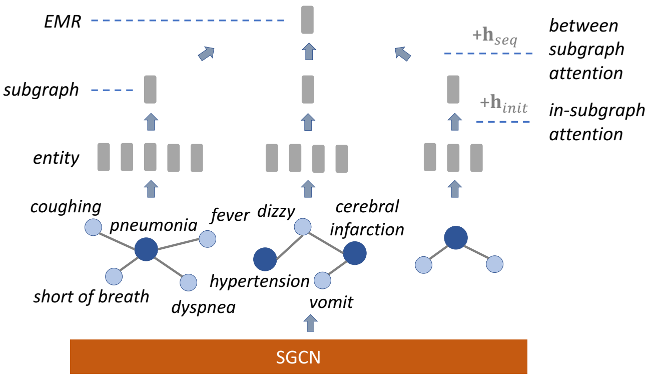


Figure 3: The illustration of HiDAN. The network hierarchically generates feature representations from entity-level upwards to subgraph-level till EMR-level. $\mathbf{h}_{seq}$ and $\mathbf{h}_{init}$ are external attention vectors generated from sequential feature extraction.

subgraph attention layer is used to obtain the compound structural feature of EMR:

$$\hat{\mathbf{a}} = softmax(\begin{bmatrix} \mathbf{c}_1 \\ \dots \\ \mathbf{c}_K \end{bmatrix} \mathbf{h}_{seq}), \tag{8}$$

$$\mathbf{h}_{att} = \hat{\mathbf{a}}^\top \begin{bmatrix} \mathbf{c}_1 \\ \dots \\ \mathbf{c}_K \end{bmatrix}, \tag{9}$$

where $\mathbf{h}_{seq}$ is the sequential feature of texts from Eq. (2), $\hat{\mathbf{a}}$ is the per-subgraph attention weights and $\mathbf{h}_{att}$ is the ultimate structural feature of EMR. Thus, $\hat{\mathbf{a}}$ and $\mathbf{a}_i$ indicate the importance of each subgraph as well as the importance of each entity in the subgraph in determining the classification, which introduces interpretability of results to some degree.

To obtain a valid neural network structure, we fix the number of subgraphs $K$ in the experiments, e.g. $K = 3$. That is, we preverse the top-$K$ largest subgraphs in the finding subgraph generation to balance between computation cost and accuracy of diagnosis prediction. If there exist less than $K$ subgraphs after clustering, there will be “placeholder” subgraphs where the identity matrix $\mathbf{I}_i$ is filled with all zeros in Eq. (5). Accordingly, the subgraph-level representation $\mathbf{a}_i$ as well as its corresponding entry in the per-subgraph attention weight are zeros. Thus, it does not impact the ultimate representation $\mathbf{h}_{att}$. Besides, since the trainable parameters in HiDAN are shared across subgraphs and the weighted sum of subgraph representations is used in Eq. (8) and (9), the proposed method is not *subgraph-order-sensitive*. That is, the change of the order of subgraphs in the input does not change the classification results.

### 4.3 Diagnosis Inference

Besides $\mathbf{h}_{seq}$ and $\mathbf{h}_{att}$, we append gender and age group as one-hot demographic feature $\mathbf{h}_{demo}$ in classification:

$$\mathbf{p} = softmax(\mathbf{W}^{(3)} Dropout(\begin{bmatrix} \mathbf{h}_{seq} \\ \mathbf{h}_{att} \\ \mathbf{h}_{demo} \end{bmatrix}) + \mathbf{b}^{(3)}), \tag{10}$$

where $\mathbf{W}^{(3)}$ and $\mathbf{b}^{(3)}$ are the trainable parameters. The disease of the largest probability is predicted as diagnosis. For multi-label classification, *softmax* is replaced by per-label *sigmoid* in Eq. (10).

| Metrics | MIMIC-III-50 | CHS-AD-200 |
|---|---|---|
| # of diagnosis codes | 50 | 200 |
| # of EMRs | 11368 | 50254 |
| avg # of diagnosis per EMR | 5.8 | 1 |
| avg # of entities per EMR | 36.8 | 18.1 |
| avg # of subgraphs per EMR | 9.5 | 3.8 |

Table 1: The statistics of datasets in the evaluation. # represents “number”. The average number of clustered subgraphs per EMR is calculated based on the output of LPA algorithm. Please note that the $K$ largest subgraphs (by number of nodes) are input into SGCN.

## 5 Evaluation

In this section, the evaluation results of the proposed method on real-world datasets are discussed.

### 5.1 Expreimental Settings

The proposed method is evaluated on real-world datasets:

*MIMIC-III-50*: A public English EMR dataset consisting of the Top-50 frequent diagnosis codes[4] [Mullenbach *et al.*, 2018]. Each EMR has one or more diagnosis codes. Thus, we use **MIMIC-III-50** to evaluate the performance of the proposed method on multi-label classification.

*CHS-AD-200*: We collected a new Chinese datasets of admission notes in collaboration with a Top-Tier Chinese hospital. The Top-200 frequent diagnosis codes are selected, which cover more than 97% of all admission notes. Unlike **MIMIC-III-50**, we only have a single diagnosis code for each admission note. Thus, we use **CHS-AD-200** to evaluate the performance of the proposed model on single-label classification.

More detail about the datasets is shown in Table 1.

Since medical NER is not the focus of this study, we experiment with the existing packages. For English dataset, we use CliNER[5] [Boag *et al.*, 2018], which reports 83.8% F1 score in the original paper. For Chinese dataset, we collaborate with a Chinese medical NER service provider[6], whose F1 score is reported about 91% in an offline evaluation on admission notes [Yuan *et al.*, 2020].

### 5.2 Performance Comparison

The performance of the proposed method on diagnosis classification is compared with the state-of-the-art deep learning based diagnosis methods, including **CNN** [Yang *et al.*, 2018], **ACNN** (CNN with attention) [Girardi *et al.*, 2018], **CAML** (CNN with label-wise attention) [Mullenbach *et al.*, 2018], **MultiResCNN** (Multi-filter CNN with ResNet) [Li and Yu, 2020] and **GMAN** (GCN with mutual attention) [Yuan *et al.*, 2020]. We use micro F1 and macro F1 as the evaluation metrics. For the fairness of comparison, we directly cite the numbers reported in the original paper for MultiResCNN [Li and Yu, 2020]. Since CHS-AD-200 is used

[4] https://github.com/jamesmullenbach/caml-mimic
[5] https://github.com/text-machine-lab/CliNER
[6] https://ai.baidu.com/solution/mtp

| Methods | MIMIC-III-50 | | CHS-AD-200 | |
|---|---|---|---|---|
| | micro F1 | macro F1 | micro F1 | macro F1 |
| CNN | 62.52% | 56.75% | 71.22% | 71.10% |
| ACNN | 63.13% | 57.00% | 74.67% | 74.70% |
| CAML | 63.39% | 57.11% | 72.56% | 72.48% |
| MultiResCNN | 67.00%† | 60.60%† | 74.64% | 74.61% |
| GMAN | 66.04% | 62.38% | 80.19% | 80.23% |
| SHiDAN | **69.19%** | **64.69%** | **81.50%** | **81.65%** |

Table 2: The evaluation results of performance on all comparative methods. † The results of MultiResCNN on MIMIC-III-50 are directly cited from the original paper since the same training and the testing datasets as well as the metrics are used.

to evaluate single-label classification performance, F1 equals precision and recall on this dataset.

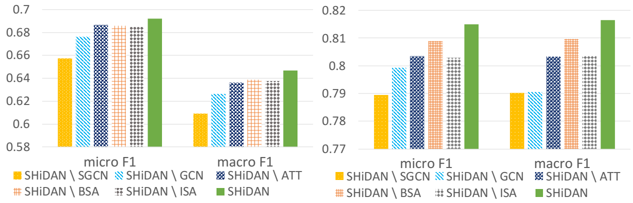
By default, the number of dimensions of word embeddings and entity embeddings is 100. The number of dimensions of latent feature $m$ is 128. The dropout rate is empirically 0.2. On MIMIC-III-50, each model is trained 12 epochs with batch size 16, and the maximum number of subgraphs $K = 15$. On CHS-AD-200, each model is trained 35 epochs with batch size 64, and the maximum number of subgraphs $K = 6$.

Table 2 shows the comparative results. On both the English and the Chinese datasets, the proposed method outperforms the state-of-the-art methods with considerable improvement. CNN, ACNN and CAML process clincal texts with sequential models only, and their results are much inferior to the rest methods. MultiResCNN improves performance of sequential models by stacking deeper residual blocks, but it generally reports poorer performance than the models incorporating structural features. The proposed method defeats all comparative methods including GMAN in the evaluation. From the results, we come to a conclusion that, by exploring the structural feature of texts, especially by applying the Seq2Subgraph framework, the performance of the downstream classification task can be remarkably improved.

### 5.3 Ablation Studies and Parameter Sensitivity

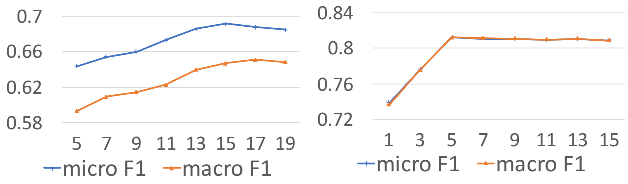
Figure 4 illustrates the results of ablation studies on the proposed method. For comparative groups without a certain attention layer, it is replaced by simple average of features. For comparative groups without GCN, the features of entities are directly passed to downstream layers without update. From the results, we can see that the most significant performance drop is found on "SHiDAN \SGCN" when no subgraph is generated and the single induced graph is directly used. This means, the clustering of nodes upon which SGCN is applied, can capture important signals that improve the classification performance. Similarly, the results also show that the modules of GCN and layered attention add to the performance gain of the proposed method with different degrees.

Figure 5 shows the sensitivity of the maximum number $K$ of subgraphs in SGCN. Interestingly, we notice a drop of performance when $K$ exceeds 15 on MIMIC-III-50 and 5 on CHS-AD-200. Although the drop is slight, it challenges our intuition that truncated number of subgraphs would lead to loss of information from input and further decrease the overall performance. After analyzing the EMRs in the datasets, it can be attributed to the fact that EMR documents are noisy, especially for long documents like admission notes, and the description of texts may contain information from past medical history that barely influences the current illness situation. Thus, with truncated subgraphs, the less important noise can be removed, which improves the performance.



(a) MIMIC-III-50 (b) CHS-AD-200

Figure 4: The results of ablation studies on SHiDAN. "SHiDAN \X" means SHiDAN **without** module X. \ATT means the removal all attention mechanism and replace with average. Similarly, \BSA and \ISA mean the removal of between-subgraph attention and in-subgraph attention, respectively. \SGCN means it does not perform finding subgraph generation and use the induced graph directly.



(a) MIMIC-III-50 (b) CHS-AD-200

Figure 5: The sensitivity of the maximum subgraph number $K$.

## 6 Conclusion

This paper studies the problem of text-based diagnosis classification, which is an important application in AI-enabled healthcare studies. It proposes a novel sequence-to-subgraph framework that transforms clinical notes into subgraphs where the structural features of contents are investigated. Moreover, a new classification model under the framework is proposed, which introduces subgraph convolutional network and hierarchical diagnostic attentive network that deeply explores the structural features of texts. The evaluation conducted on the real-world EMR datasets proves the effectiveness of the proposed method in diagnosis classification. This study will be extended by considering the issue of imbalance classes so that the model performs well on rare diseases.

## Acknowledgments

Our work is supported by the National Key Research and Development Program of China No.2020AAA0109400.

## References

[Baumel *et al.*, 2018] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad.

Multi-label classification of patient notes: Case study on ICD code assignment. In *Proceedings of the Workshops of AAAI*, pages 409–416, 2018.

[Boag *et al.*, 2018] Willie Boag, Elena Sergeeva, Saurabh Kulshreshtha, Peter Szolovits, Anna Rumshisky, and Tristan Naumann. CliNER 2.0: Accessible and accurate clinical concept extraction. In *arXiv:1803.02245*, 2018.

[Cao *et al.*, 2018] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In *EMNLP*, pages 182—-192, 2018.

[Cetoli *et al.*, 2017] Alberto Cetoli, Stefano Bragaglia, Andrew D O'Harney, and Marc Sloan. Graph convolutional networks for named entity recognition. In *International Workshop on Treebanks and Linguistic Theories*, pages 37–45, 2017.

[Chen *et al.*, 2020] Jun Chen, Xiaoya Dai, Quan Yuan, Chao Lu, and Haifeng Huang. Towards interpretable clinical diagnosis with Bayesian network ensembles stacked on entity-aware CNNs. In *ACL*, pages 3143–3153, 2020.

[Choi *et al.*, 2017] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: Graph-based attention model for healthcare representation learning. In *KDD*, pages 787–795, 2017.

[Choi *et al.*, 2020] Edward Choi, Zhen Xu, Yujia Li, Michael W. Dusenberry, Gerardo Flores, Emily Xue, and Andrew M. Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. In *AAAI*, pages 606–613, 2020.

[Dai *et al.*, 2019] Dai Dai, Xinyan Xiao, Yajuan Lyu, Shan Dou, Qiaoqiao She, and Haifeng Wang. Joint extraction of entities and overlapping relations using position-attentive sequence labeling. In *AAAI*, pages 6300–6308, 2019.

[Fu *et al.*, 2019] Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In *ACL*, pages 1409–1418, 2019.

[Girardi *et al.*, 2018] Ivan Girardi, Pengfei Ji, An phi Nguyen, Nora Hollenstein, Adam Ivankay, Lorenz Kuhn, Chiara Marchiori, and Ce Zhang. Patient risk assessment and warning symptom detection using deep attention-based neural networks. In *EMNLP Workshop*, pages 139–148, 2018.

[Gregoric *et al.*, 2018] Andrej Zukov Gregoric, Yoram Bachrach, and Sam Coope. Named entity recognition with parallel recurrent neural networks. In *ACL*, pages 69—-74, Melbourne, Australia, 2018.

[Li and Yu, 2020] Fei Li and Hong Yu. Icd coding from clinical text using multi-filter residual convolutional neural network. In *AAAI*, pages 8180–8187, 2020.

[Mullenbach *et al.*, 2018] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *NAACL*, pages 1101—1111, 2018.

[Raghavan *et al.*, 2007] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(036106), September 2007.

[Sha and Wang, 2017] Ying Sha and May D. Wang. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In *ACM BCB*, pages 233–240, 2017.

[Singh *et al.*, 2014] Hardeep Singh, Ashley N D Meyer, and Eric J Thomas. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving us adult populations. *BMJ Quality and Safety*, 23(9):727–731, September 2014.

[Vu *et al.*, 2020] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for icd coding from clinical text. In *IJCAI*, pages 3335–3341, 2020.

[Wang *et al.*, 2019] Qi Wang, Yangming Zhou, Tong Ruan, Daqi Gao, Yuhang Xia, and Ping He. Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition. *Journal of Biomedical Informatics*, 92, April 2019.

[Wang *et al.*, 2020a] Ke Wang, Xuyan Chen, Ning Chen, and Ting Chen. Automatic emergency diagnosis with knowledge-based tree decoding. In *IJCAI*, pages 3407–3414, 2020.

[Wang *et al.*, 2020b] Zifeng Wang, Rui Wen, Xi Chen, Shilei Cao, Shao-Lun Huang, Buyue Qian, and Yefeng Zheng. Online disease self-diagnosis with inductive heterogeneous graph convolutional networks. In *arXiv:2009.02625*, 2020.

[Wu *et al.*, 2018] Yonghui Wu, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. Clinical named entity recognition using deep learning models. In *AMIA Annual Symposium*, pages 1812–1819, 2018.

[Yang *et al.*, 2016] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *NAACL-HLT*, pages 1480–1489, 2016.

[Yang *et al.*, 2018] Zhongliang Yang, Yongfeng Huang, Yiran Jiang, Yuxi Sun, Yu-Jin Zhang, and Pengcheng Luo. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Scientific Reports*, 8(1), April 2018.

[Yao *et al.*, 2019] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *AAAI*, pages 7370–7377, 2019.

[Yuan *et al.*, 2020] Quan Yuan, Jun Chen, Chao Lu, and Haifeng Huang. The graph-based mutual attentive network for automatic diagnosis. In *IJCAI*, pages 3393–3399, 2020.

[Zhao *et al.*, 2019] Shan Zhao, Zhiping Cai, Haiwen Chen, Ye Wang, Fang Liu, and Anfeng Liu. Adversarial training based lattice lstm for chinese clinical named entity recognition. *Journal of Biomedical Informatics*, 99, 2019.