LGFat-RGCN: Faster Attention with Heterogeneous RGCN for Medical ICD Coding Generation

Zhenghan Chen*
Peking University
Beijing, China
1979282882@pku.edu.cn

Changzeng Fu*
Northeastern University
Qinhuangdao, China
changzeng.fu@irl.sys.es.osaka-u.ac.jp

Ruoxue Wu†*
Worcester Polytechnic Institute
Worcester, United States
rochelle.wu820@gmail.com

Ye Wang
Peking University
Beijing, China
wangye2111@gmail.com

Xunzhu Tang†
University of Luxemburg
Luxemburg, Luxemburg
xunzhu.tang@uni.lu

Xiaoxuan Liang
University of Massachusetts Amherst
Amherst, United States
xiaoxuanlian@umass.edu

ABSTRACT

With the increasing volume of healthcare data, automated International Classification of Diseases (ICD) has become increasingly relevant and is frequently regarded as a medical multi-label prediction problem. Current methods struggle to accurately classify medical diagnosis texts that represent deep and sparse categories. Unlike these works that model the label with code hierarchy or description for label prediction, we argue that the label generation with structural information can provide more comprehensive knowledge based on the observation that label synonyms and parent-child relationships in vary from their context in clinical contexts. In this study, we introduce `LGFat-RGCN`, a heterogeneous graph model with improved attention for automated ICD coding. Notably, our approach represents the model to consider this task as a labelled graph generation problem. Our enhanced attention mechanism boosts the model's capacity to learn from multi-relational heterogeneous graph representations. Additionally, we propose a discriminator for labelled graphs (LG) that computes the reward for each ICD code in the labelled graph generator. Our experimental findings demonstrate that our proposed model significantly outperforms all existing strong baseline methods and attains the best performance on three benchmark datasets.

CCS CONCEPTS

• **Computing methodologies → Information extraction**; **Natural language generation**; **Discourse, dialogue and pragmatics**.

KEYWORDS

neural networks, gaze detection, text tagging

*Both authors contributed equally to this research.
†Corresponding author.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Zhenghan Chen, Changzeng Fu, Ruoxue Wu, Ye Wang, Xunzhu Tang, and Xiaoxuan Liang. 2018. LGFat-RGCN: Faster Attention with Heterogeneous RGCN for Medical ICD Coding Generation. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Automated International Classification of Diseases (ICD) coding, which involves assigning ICD codes to patient visits, has garnered considerable attention owing to its potential to reduce the time and labor required for billing [24, 21, 25]. Historically, healthcare institutions have been compelled to engage the services of specialized coders for the execution of the International Classification of Diseases (ICD) coding process. This approach is associated with significant drawbacks, such as high financial costs, lengthy time investments, and susceptibility to errors. Consequently, numerous alternative methodologies aimed at automating the ICD coding process have been proposed and explored since the 1990s [7].

Recent approaches to this task predominantly frame it as a multi-label classification problem [32, 11, 38]. These methods employ deep learning techniques to extract representations of Electronic Medical Records (EMRs) using Recurrent Neural Network (RNN) or Convolutional Neural Network (CNN) encoders, and subsequently predict ICD codes using multi-label classifiers. State-of-the-art methodologies have introduced label attention, which utilizes code representations as attention queries to extract code-related representations [15]. In addition, numerous studies have proposed leveraging the hierarchical structure of ICD codes [8, 33, 4] and integrating code descriptions to enhance label representations and improve the overall performance of the automated ICD coding process.

Through our analysis of the International Classification of Diseases (ICD) codes, we discovered that only 122 of the 9,219 codes correspond to the most common top 50, indicating a severe imbalance in the distribution of codes and a predominance of inactive codes in clinical texts. Moreover, the majority of prior methods neglect or undervalue the relationships between ICD codes, such as parent-child, sibling, and mutually exclusive relationships [15, 32]. Lastly, existing approaches rely on a single training method to update parameters [28, 26, 9], which may result in failure for some clinical texts covering uncommon disorders.

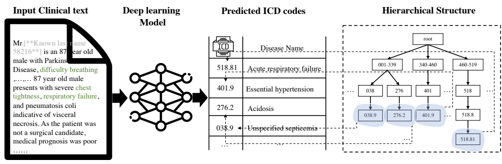


Figure 1: A hierarchical diagram of the International Classification of Diseases, Ninth Revision (ICD-9) codes, along with an example of the automatic ICD coding task. In this task, the model takes clinical text as input and outputs the predicted ICD codes.

To address the difficulties discussed above, we propose a novel approach for automated ICD coding that formulates the task as a labeled graph generation problem along the ICD code graph. The majority of neural network methods treat automated coding as a multi-label prediction problem [14, 3]. In contrast to the majority of preceding methodologies, which address this challenge as a multi-label prediction issue, we approach it as a labeled graph generation problem. Our proposed method, `LGFat-RGCN`, comprises several components, including a Labeled Graph Generator (LGG), a Labeled Graph Discriminator (LGD), and a Message Integration Module (MIM). When provided with clinical text, the text encoder generates an input representation, which is then fed to MIM to model the relationships between clinical text and ICD codes. Specifically, the LGG aims to generate graph labels that are indistinguishable from original ICD labels, while the LGD aims to differentiate between original and generated ICD labels.

We conduct extensive experiments on the MIMIC-III benchmark dataset [10] to empirically demonstrate the effectiveness of our proposed method, `LGFat-RGCN`. Our experimental results show that `LGFat-RGCN` outperforms state-of-the-art techniques by a significant margin. In summary, the key contributions of this paper include:

- We propose a novel approach that formulates automatic ICD coding as a labeled graph generation task and introduce a multi-algorithm model named `LGFat-RGCN`. Notably, we design a Labeled Graph Discriminator (LGD) that evaluates intermediate rewards as supervision signals for `LGFat-RGCN`.
- We introduce a Message Integration Module (MIM) that models the parent-child, sibling, and mutually exclusive relationships among ICD codes in order to improve the accuracy of automatic ICD coding.
- We demonstrate the effectiveness of proposed `LGFat-RGCN` in generating ICD codes by achieving superior performance over several baseline models on three benchmark datasets.

2 RELATED WORK

2.1 Automatic ICD Coding

The automatic EHR coding task has garnered significant attention in recent years, with a multitude of studies exploring various approaches such as joint word and label embeddings [27], multitask classification [18], and separate machine learning models for different EHR modalities [34]. Our work distinguishes itself from prior research in two ways. Firstly, we frame automatic EHR coding task as a labeled graph generation problem, a novel approach not explored in previous studies. Secondly, our proposed framework incorporates various types of relationships between entities, allowing for more comprehensive modeling of EHR data.

2.2 Graph Representation Learning

The domain of knowledge graphs has witnessed the emergence of various solutions for graph representation learning, regarded as a pivotal technology in this field. These solutions can be broadly categorized into four primary classifications: translation distance models [6], semantic matching models [36], random walk models [31], and subgraph aggregation models [30]. Knowledge graph representation learning models grounded in translation distance predominantly encompass the Trans family of models, exemplified by the TransE model [2].

3 METHODOLOGY

As illustrated in Figure 2, the `LGFat-RGCN` encompasses two principal components: the labeled graph generator G_θ and the labeled graph discriminator D_ζ. In the following sections, we expound upon the architecture of `LGFat-RGCN`.

3.1 Labeled Graph Generator G_θ

The labeled graph generation process is denoted by $< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} >$. Within this formulation, $\mathcal{S}$ represents the state space, while $\mathcal{A}$ constitutes the set of all feasible actions. For example, the subset of $\mathcal{A}$ corresponding to a specific label comprises its neighbors in the global graph. The transition function, denoted by $\mathcal{T}$, facilitates the progression of state transitions, whereas $\mathcal{R}$ signifies the reward function associated with each (state, action) pair. To encourage G_θ to generate labels akin to ground truth, we propose maximizing the expected rewards via the reinforce algorithm. Given a trajectory $\tau = s_1, a_1, s_2, a_2, ..., s_T, a_T$, where a denotes an action, the expected payoff can be computed using Equation 1, 2 and 3. Furthermore, $\bar{R}_\theta$ yields the average expected value for the rewards across trajectories.

$$\begin{aligned} R(\theta) &= E_{\tau \sim P_\theta(\tau)} [R(\tau)] \\ &= E_{a \sim \pi(a)} \end{aligned} \tag{1}$$

$$\begin{aligned} \bar{R}(\theta) &= E_{a \sim \pi(a|S=s, X=x; \theta)} \left[\Sigma_i R\left(s = s_i, X = x, a_i\right) \right] \\ &= \Sigma_t \Sigma_{a_i \in \mathcal{A}} \pi\left(a_i \mid s = s_i, X = x; \theta\right) R_i \end{aligned} \tag{2}$$

$$R_i = R\left(s = s_i, X = x, a_i\right) \tag{3}$$

In this context, $R(\theta)$ denotes the expected reward derived from a single trajectory, while $\bar{R}(\theta)$ signifies the anticipated aggregate reward obtained from one episode, and τ represents the trajectory itself. The labeled graph generator is expressed as G_θ, with its hybrid policy network given by $\pi(a_i|s = s_i, X = x; \theta)$. Here, a_i refers to the label generated based on the current states s_i and x, and $R(s = s_i, X = x, a_i)$ constitutes the reward for producing a_i contingent upon s_i and x. The label a_i can be incorporated into the module D_ζ. Subsequently, we elucidate how the policy gradient can be employed to adjust θ, (notably, $R(s = s_i, X = x, a_i)$ is independent of θ):

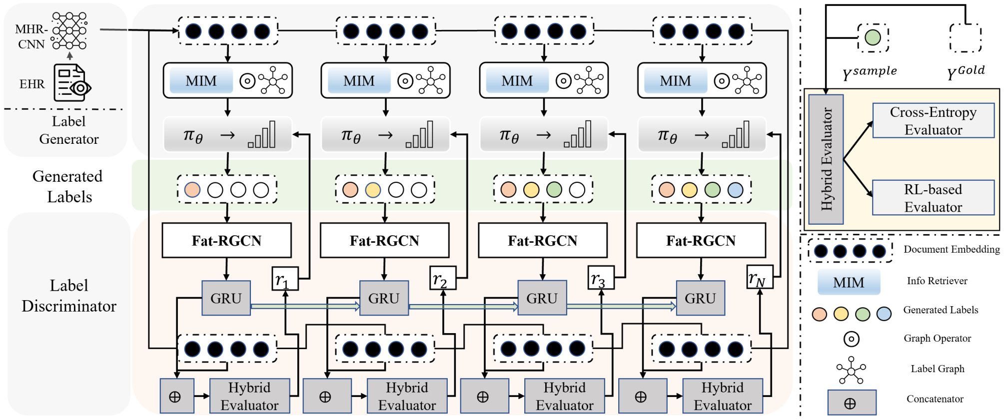


Figure 2: As delineated in the LGFat-RGCN framework, there are two pivotal components: the Label Generator G_θ and the Label Discriminator D_ζ. A thorough elucidation of MIM, MHR-CNN, and Fat-RGCN will be provided in the ensuing sections.

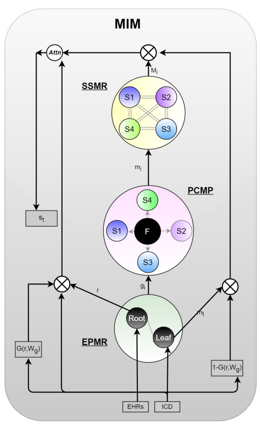


Figure 3: Message Integration Module (MIM).

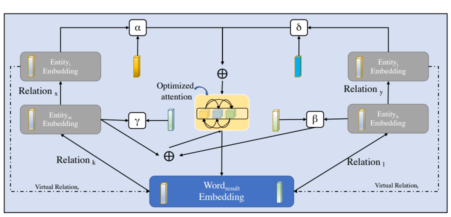


Figure 4: MHMI.

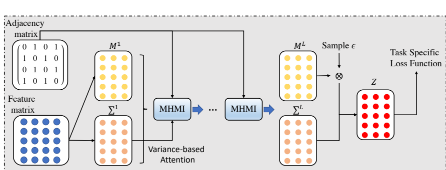


Figure 5: Fat-RCGN.

$$\nabla \bar{R}(\theta) = \sum_{t} \sum_{a_i \in \mathcal{A}} \pi(a_i | s = s_i, X = x; \theta) \\ \nabla \log \pi(a_i | s = s_i, X = x; \theta) \tag{4}$$

The expression $\pi(a_i|s = s_i, X = x; \theta)$ can be articulated as Equation 5:

$$\pi(a_i|s = s_i, X = x; \theta) = \sigma(W(s_i) + b_i) \tag{5}$$

In this representation, W corresponds to a matrix and b denotes a bias term, while the sigmoid activation function is symbolized by σ.

3.2 Labeled Graph Discriminator D_ζ

We devise the trajectory discriminator module D_ζ to procure the reward m_t for each code within the generated path $(c_1, c_2, ..., c_i)$ up to time step i. More precisely, we model h_i as the discrimination probability, as elaborated below:

$$\begin{aligned} h_i &= R_{(s=s_i, X=x, a=a_i)} \\ &= p_s((c_1, c_2, c_3, ..., c_i), x) \\ &= \sigma(M_h(LSTM(h_{k-1}, c_k) \oplus x)) \end{aligned} \tag{6}$$

In this formulation, $\oplus$ symbolizes the concatenation operation, while M_h denotes the weight matrix; c_i refers to the current generated trajectory obtained through iterative application of an LSTM to the ICD code path. To ascertain and gauge the accuracy of D_ζ, we employ a cross-entropy function, which is defined as:

$$Loss_s = -\sum_{(y_i, x) \in S^+} \log p_s(y_i, x) - \\ \sum_{(y_i, x) \in S^-} \log(1 - p_s(y_i, x)) \tag{7}$$

In this expression, S^+ and S^- correspond to positive and negative samples, respectively, while $p_s(y_t, x)$ designates the probability that the sample (y_i, x) is categorized as a positive instance.

3.3 Message Integration Module (MIM)

Our principal encoder for clinical representations is the RPGNet, which encompasses three stages: EHR-to-Path Message Release

(**EPMR**), Parent-to-child message passing (**PCMP**), and Sibling-to-Sibling Message Release (**SSMR**). Consequently, the state s_t can be encoded as depicted by MIM in Figure 2:

$$s_t = (1 - G(r, W_g))r + G(r, W_g)m_t \tag{8}$$

where W_g is a weight matrix and G is a control gate for information transformation based on the r and m_t representations of EHR, respectively.

3.3.1 ***EPMR***. The symbolic representation of the relationship between an EHR and an ICD trajectory, denoted by g_i, can be generated as elaborated below:

$$\begin{aligned} r_t &= (x_i \cdot p_t)^\frown(x_i + p_t)^\frown(x_i - p_t)^\frown(p_t - x_i) \\ g_i &= \tanh(W_p(x_i^\frown p_t^\frown r_t)) \end{aligned} \tag{9}$$

In this context, W_p represents the weight matrix, and $\frown$ symbolizes concatenation. The parameter within W_p is derived from distinct transformations of the EHR representation r_t and the path representation p_t.

3.3.2 ***PCMP***. PCMP is employed to capture the relationship between parent and child ICD codes of ICD code r_i. The association between an EHR and an ICD trajectory is characterized as p_t. Subsequently, this relational representation is propagated from the parent code to all its child codes, generating the relation representation m_i: $n_i = r_i \cdot s_t^p$. Here, $\cdot$ signifies the element-wise product operation, and s_t^p represents the vector representation of each child ICD code.

3.3.3 ***SSMR***. SSMR is employed to encode the associations among sibling ICD codes by facilitating the exchange of information between them. The corresponding formulation is presented below:

$$M_i = \sum_{n \in S_{bi}} C_{attn}(b_i, b_i^n) + b_i \tag{10}$$

In this representation, C_{attn} refers to the attention function, S_{bi} corresponds to all ICD siblings of code b_i, and b_i^n designates the n-th ICD sibling of code b_i.

3.4 MHR-CNN for G_θ's Embedding

Multi-Header Convolutional Filter (**MCF**): Let us assume there are m filters, $f_1, f_2, ..., f_n$, with kernel sizes represented by $k_1, k_2, ..., k_n$. Consequently, m 1-dimensional convolutions can be applied to the input matrix X. The formalization of the convolutional approach is presented below:

$$\begin{aligned} F_1 = f_1(X) &= \bigwedge_{j=1}^{l} \tanh\left(W_1^T X^{j:j+k_1-1}\right) \\ F_n = f_n(X) &= \bigwedge_{j=1}^{l} \tanh\left(W_n^T X^{j:j+k_n-1}\right) \end{aligned} \tag{11}$$

In this representation, $\bigwedge_{j=1}^{l}$ denotes the left-to-right convolutional operations. The sub-matrices of X are indicated by $X^{j:j+k_1-1} \in \mathbb{R}^{k_1 \times d^x}$ and $X^{j:j+k_n-1} \in \mathbb{R}^{k_n \times d^x}$. The weight matrices of the corresponding filters are represented by $W_1 \in \mathbb{R}^{(k_1 \times d^x) \times d^f}$ and $W_n \in \mathbb{R}^{(k_n \times d^x) \times d^f}$.

$$H_m = f_m(E) = \bigwedge_{j=1}^{n} \tanh\left(W_m^T E^{j:j+k_m-1}\right) \tag{12}$$

Multi-Residual Convolutional Block (**MCB**): In the multi-filter convolutional layer, a residual convolutional layer consisting of p residual blocks is positioned above each filter. Comprising the residual block c_{ni} are three convolutional filters: c_{n1}, c_{n2}, and c_{n3}. The computational process is denoted as follows:

$$\begin{gathered} I_1 = c_{ni_1}(I) = \bigwedge_{j=1}^{l} \tanh\left(W_{ni_1}^T I^{j:j+k_n-1}\right), \\ I_2 = c_{ni_2}(I_1)\,; I_3 = c_{ni_3}(I_1); F_{ni} = \tanh(I_2 + I_3)\,, \end{gathered} \tag{13}$$

The calculation process of a Multi-Residual Convolutional Block (**MCB**) is represented by the symbol $\bigwedge_{j=1}^{l}$, which denotes a sequence of convolutional operations. I is the input matrix of the block, and $I^{j:j+k_n-1} \in \mathbb{R}^{k_n \times d^{i-1}}$ represents its submatrices. The weight matrices of the three convolutional filters, namely c_{ni_1}, c_{ni_2} and c_{ni_3}, are represented by $W_{ni_1} \in \mathbb{R}^{(k_n \times d^{i-1}) \times d^i}$ and $W_{ni_3} \in \mathbb{R}^{(1 \times d^{i-1}) \times d^i}$. The kernel sizes of r_{mi_1} and r_{mi_2} are the same as the corresponding filter f_m in the multi-filter convolutional layer, denoted by k_m, but they have different in-channel sizes. On the other hand, r_{mi_3} is a convolutional filter with a kernel size of 1, which is special compared to the other filters.

3.5 Fat-RGCN for D_ζ's Embedding

3.5.1 ***Attention Mechanism Optimization (AMO)***. Three different one-hop neighbor-level-based models are currently in use: Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and Relational Graph Convolutional Networks (RGCN). The GAT model's attention formula consists of two components, namely s_{ij} and n_{ij}.

$$\begin{aligned} \beta^T\,[A_m \| A_n] &= [\beta_{\mathrm{s}} + \beta_{\mathrm{n}}]^T\,[A_m \| A_n] \\ &= \beta_{\mathrm{s}}\,A_m + \beta_{\mathrm{n}}\,A_n \end{aligned} \tag{14}$$

In practice, the original GAT model's parameters are separated into those of s_{ij} and n_{ij}. The Attention parameter α represents the overall GAT.

In other words, the attention mechanism of the GAT model comprises both s_{ij} and n_{ij}, resulting in a more comprehensive approach to attention.

$$\begin{bmatrix} s_{11} & n_{12} & \cdots & n_{1j} \\ n_{21} & s_{22} & \cdots & n_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ n_{i1} & n_{i2} & \cdots & n_{ij} \end{bmatrix} \tag{15}$$

3.5.2 ***One-hop Neighborhood Graph Representation (ONGR)***. This section presents a novel model for **ONGR** that simultaneously accounts for the influence of nodes, relations, and weights. The proposed **ONGR** model employs three attention optimization techniques, including Node Attention in RGCN Convergence (**NARC**), Faster Attention Mechanism in Convergence (**FAMC**), and Faster Attention in Nodes and Relations (**FANR**). This model represents a significant advancement over previous approaches and addresses several deficiencies identified in the literature. Extensive experimentation confirms the effectiveness of the proposed model, with empirical results supporting its efficacy.

NARC: The **NARC** is to directly include GAT's Attention during the RGCN model convergence process.

$$C_u = F\left(\sum_{(n,r)\in\mathcal{P}(u)} \gamma\left(N_r, \beta_G * X_n\right) * R_t\right) \tag{16}$$

FAMC: The **FAMC** strategy adds the Attention weights to the neighbor nodes. As shown below, the formula for central node aggregation.

$$C_u = F\left(\sum_{(n,r)\in\mathcal{P}(u)} \gamma\left(N_r, \beta_O * X_n\right) * R_t\right) \tag{17}$$

The remaining processing techniques are the same as in the first scheme, where β_O stands for the modified GAT's Attention aggregation approach.

FANR: The **FANR** strategy adds Attention weights to nearby nodes and relations.

$$C_u = F\left(\sum_{(n,r)\in\mathcal{P}(u)} \gamma\left(N_r, X_n\right) * R_t * \beta_O\right) \tag{18}$$

As a key step in the proposed methodology, the node representation X_n and relationship representation N_r are first combined using the γ function. Subsequently, we introduce the use of β_O to determine the weights of the combined representations. This weighting process serves to selectively focus on the most relevant features, thereby improving the accuracy of graph neural networks in capturing complex relationships.

3.5.3 Multi-hop Neighborhood Graph Representation (MNGR).

We suggest a gate mechanism be used to filter nodes, given that the inclusion of a significant number of two-hop neighbor nodes results in noise, alongside accurate information. To depict the node aggregation process in **MNGR**, we present the following equation.

$$C_{ui} = F\left(\sum_{(n,r)\in\mathcal{P}(u)} \gamma\left(Z_{ri}, X_{ni}\right) W_r\right) \tag{19}$$

$$C_{uj} = F\left(\sum_{(n,r)\in\mathcal{P}(u)} \gamma\left(Z_{rj}, X_{nj}\right) W_r\right) \tag{20}$$

$$C_u = \left(1 - D\left(C_{uj}\right)\right) \cdot C_{uj} + D\left(C_{uj}\right) \cdot C_{ui} \tag{21}$$

We propose that the gate mechanism $D(C_{uj})$ be applied to filter C_{ui} and C_{uj}, following the aggregation of one-hop neighbors and two-hop neighbors. The letters C_{ui} and C_{uj} are utilized to represent $D(C_{uj})$ after these aggregations.

3.5.4 Multi-hop Model Integration (MHMI).

The revised algorithm model is extensively detailed at the one-hop and multi-hop neighbor levels. Subsequently, we introduce **MHMI** - a novel, multi-relational deep graph representation constructed by integrating multiple-level enhancement techniques. Figure 4 depicts the architecture of this model.

The convergence equation that leverages the Attention mechanism of the modified GAT to calculate β_O is presented below.

$$C_u = \left(1 - D\left(C_{uj}\right)\right) \cdot C_{uj} + D\left(C_{uj}\right) \cdot C_{ui} \tag{22}$$

$$D\left(C_{uj}\right) = \sigma\left(X + A_{uj}\right) \tag{23}$$

The aforementioned formula is evidently based on the multi-hop scheme convergence of C_{ui} and C_{uj}.

4 EXPERIMENTAL SETUP

In this section, we conduct comprehensive experiments aimed at addressing the following research questions:

- **RQ-1:** *What is the performance of* `LGFat-RGCN`*?*
- **RQ-2:** *What is the impact of the key design choices on the performance of* `LGFat-RGCN`*?*
- **RQ-3:** *To what extent is* `LGFat-RGCN` *effective on multi-relational medical graph data?*

4.1 Dataset

MIMIC-III*[10].* `LGFat-RGCN` validation utilized the public MIMIC-III dataset (50,000 records, 2000-2012); distinguished as MIMIC-III full and MIMIC-III top 50.

Cora*[13].* The Cora graph dataset encodes nodes using 1433-dimensional vectors, representing features tied to dictionary terms; 1433 features correspond to the lexicon in 2708 papers.

FB15k-237 *[19].* FB15k-237, a subset of Freebase knowledge base [5] and FB15k [2], comprises 14,541 nodes with 237 edge types, resembling Wikipedia's metadata [22] in a graph database format.

4.2 Metrics

In the experimental section, the evaluation metrics for the `LGFat-RGCN` model include Accuracy, MR, MRR, Hit@1, Hit@3, and Hit@10, as described in [29].

4.3 Baselines

Hierarchy-SVM *&* ***Flat-SVMs*** *[16].* This study proposes two encoding strategies for ICD9 codes: an independent treatment of each code (Flat-SVMs) and a hierarchical consideration of ICD9 codes (Hierarchy-SVM).

C-MemNN *[17] &* ***C-LSTM-Att*** *[20].* C-MemNN employs iterative memory condensation, while C-LSTM-Att utilizes character-aware neural language models for hidden representations.

BI-GRU *[37] &* ***HA-GRU*** *[1].* BI-GRU employs bidirectional gated recurrent units for EHRs integrated embedding, while HA-GRU, an enhanced version, improves the architecture's effectiveness.

***CAML* & *DR-CAML* [15].** CAML utilizes convolutional attention networks for ICD embeddings, while DR-CAML enhances this method for improved performance.

***LAAT* & *JointLAAT* [26].** LAAT introduces ICD code-encoded hidden state attention learning in LSTM, while JointLAAT expands it with a hierarchical joint learning approach.

***ISD* [38] & *MSMN* [35] & *FUSION* [12].** ISD presents a model linking related diagnoses; MSMN uses synonym matching for ICD classification; FUSION tackles redundant diagnosis vocabulary.

5 RESULT AND ANALYSIS

5.1 [RQ-1] Overall Performance and Comparison)

To address RQ1, we present the experimental results from the MIMIC-III dataset for both fundamental core assessment metrics and personalized metrics in Table 1. Upon careful examination of the data presented in Table 1, we draw the following conclusions.

Firstly, the `LGFat-RGCN` model yields the best results across both fundamental core assessment metrics and personalized metrics, demonstrating its efficacy and superiority. The relatively small and varying standard deviation values of the evaluation metrics for the `LGFat-RGCN` model attest to the model's stability.

Secondly, compared to `LGFat-RGCN`, the relatively low AUC and F1 scores for CAML and JointLAAT suggest that these models have limited coverage of rare codes.

Lastly, an analysis and comparison of recursive models based on the GRU class in Table 1 reveal their relatively poor performance compared to other models. The issue of gradient disappearance can be addressed by incorporating a carefully designed CNN residual connection structure.

✍ **Answer to RQ-1:** ► *To sum up, our study on the MIMIC-III dataset (Table 1) demonstrates the superior performance of* `LGFat-RGCN` *in fundamental and personalized metrics. Small standard deviations suggest its stability. Limited coverage of rare codes is implied by low AUC and F1 scores for CAML and JointLAAT, and recursive models based on the GRU class require a CNN residual connection structure to address gradient disappearance.* ◄

The ablation study conducted on the `LGFat-RGCN` model, as detailed in Table 2, demonstrates the importance of individual components to the model's overall performance. Removing ARCL, MIM, or MHR-CNN resulted in substantial declines in the performance metrics across both the MIMIC-III Full and Top50 datasets. The most significant performance deterioration was observed in the absence of the ARCL module, followed by MHR-CNN and MIM. These results emphasize the necessity of each component in the `LGFat-RGCN` model for achieving optimal performance in multi-relational medical graph data analysis.

5.2 [RQ-2] `LGFat-RGCN` Ablation

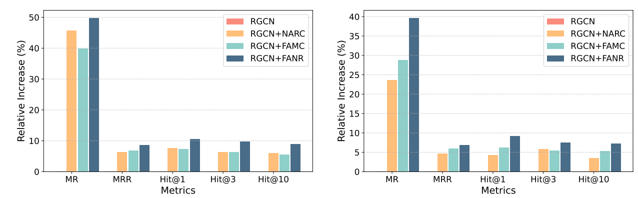
As delineated in Table 2, several ablation scenarios were assessed for the `LGFat-RGCN` model:

1) No ARCL: The absence of ARCL resulted in a substantial performance deterioration of the `LGFat-RGCN` model. Notably, the macro AUC and micro AUC measures for the MIMIC-III Full dataset declined by 15.16% and 13.13%, respectively. A similar trend was observed in the MIMIC-III Top50 dataset.

2) No MIM: Excluding the MIM component led to a comparable performance reduction for the `LGFat-RGCN` model. For instance, in the MIMIC-III Top50 dataset, the macro AUC and micro AUC metrics decreased by 9.07% and 5.97%, respectively.

3) No MHR-CNN: Evaluating the MIMIC-III Full dataset without the MHR-CNN module demonstrated an average decline of 11.02% in both macro AUC and micro AUC measures. An examination of the comparative experimental outcomes revealed that the MHR-CNN module in the `LGFat-RGCN` model enabled a more precise representation of the MHR-CNN text information.

✍ **Answer to RQ-2:** ► *The ablation study in Table 2 highlights the importance of the* `LGFat-RGCN` *model's components. Removing ARCL, MIM, or MHR-CNN led to considerable performance declines across both MIMIC-III datasets. The results emphasize the critical role of each component in the* `LGFat-RGCN` *model for optimal performance in medical graph data analysis.* ◄



(a) FB15K-237 (b) CORA

Figure 6: Experimental results of attentional optimization mechanisms in one-hop neighborhood graph representation schemes on the FB15k-237 and Cora dataset.

5.3 [RQ-3] Representation Experiment

5.3.1 Attention Optimization Comparison. Figure 6 depicts the experimental outcomes derived from an array of investigations, encompassing RGCN replication, RGCN+**NARC**, RGCN+**FAMC**, and RGCN+**FANR**. The two bar plots displaying experimental results feature relative boosting metrics on the vertical axis. As indicated by the results in Figure 6, the integration of attention mechanisms into the heterogeneous graph representation model RGCN, whether through RGCN+**NARC**, RGCN+**FAMC**, or RGCN+**FANR**, results in marked improvements across the five core metrics. These findings substantiate the efficacy of the three attention mechanism optimization algorithms proposed in this study. Ultimately, due to the exceptional performance of **FANR**, this mechanism is incorporated into the final `LGFat-RGCN` model.

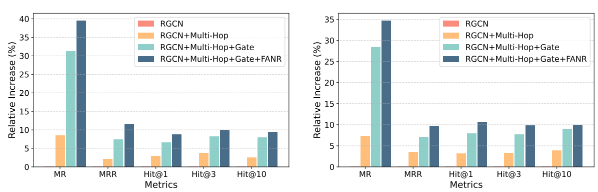
5.3.2 Experiments on Gate Mechanism. The experimental framework encompasses three distinct investigations. The initial experiment aims to reproduce the RGCN baseline model and evaluate its performance. Subsequently, the second experiment, designated as RGCN+Multi-Hop, extends the RGCN model by incorporating two-hop node information into the convergence process. The final experiment, RGCN+Multi-Hop+Gate, integrates a gate mechanism, as outlined in the AliNet study [23], into the RGCN+Multi-Hop

Table 1: Experiment results on MIMIC-III Top50 and MIMIC-III Full. The results of `LGFat-RGCN` are shown in means ± standard deviations

Model	MIMIC-III Full					MIMIC-III Top50				
	AUC		F1		P@8	AUC		F1		P@5
	Macro	Micro	Macro	Micro		Macro	Micro	Macro	Micro	
Hierarchy-SVM	0.456	0.438	0.009	0.001	0.202	0.376	0.368	0.041	0.079	0.144
Flat-SVMs	0.482	0.467	0.011	0.002	0.242	0.439	0.401	0.048	0.093	0.179
C-MemNN	0.833	0.913	0.082	0.514	0.695	0.824	0.896	0.509	0.588	0.596
C-LSTM-Att	0.831	0.908	0.079	0.511	0.687	0.816	0.892	0.501	0.575	0.574
BI-GRU	0.500	0.547	0.002	0.140	0.317	0.501	0.594	0.035	0.268	0.228
HA-GRU	0.501	0.509	0.017	0.004	0.296	0.500	0.436	0.072	0.124	0.205
CAML	0.895	0.959	0.088	0.539	0.709	0.875	0.909	0.532	0.614	0.609
DR-CAML	0.897	0.961	0.086	0.529	0.609	0.884	0.916	0.576	0.633	0.618
LAAT	0.919	0.963	0.099	0.575	0.738	0.925	0.946	0.666	0.715	0.675
JointLAAT	0.941	0.965	0.107	0.577	0.735	0.925	0.946	0.661	0.716	0.671
ISD	0.938	0.967	0.119	0.559	0.745	0.935	0.949	0.679	0.717	0.682
MSMN	0.943	0.965	0.103	0.584	0.752	0.928	0.947	0.683	0.725	0.680
FUSION	0.915	0.964	0.088	0.636	0.736	0.909	0.933	0.619	0.674	0.647
`LGFat-RGCN`	**0.989** (+4.88%) ± 0.002	**0.998** (+3.21%) ± 0.001	**0.134** (+12.61%) ± 0.001	**0.789** (+19.39%) ± 0.002	**0.798** (+6.12%) ± 0.001	**0.981** (+4.91%) ± 0.001	**0.989** (+4.21%) ± 0.002	**0.754** (+7.10%) ± 0.001	**0.787** (+8.55%) ± 0.002	**0.763** (+11.88%) ± 0.001

Table 2: Ablation experiment results on MIMIC-III Top50 and MIMIC-III Full datasets. The standard deviation of `LGFat-RGCN` results is consistent with the previous table, so it is omitted in this table.

Model	MIMIC-III Full					MIMIC-III Top50				
	AUC		F1		P@8	AUC		F1		P@5
	Macro	Micro	Macro	Micro		Macro	Micro	Macro	Micro	
`LGFat-RGCN`	**0.983**	**0.998**	**0.134**	**0.622**	**0.798**	**0.981**	**0.989**	**0.754**	**0.787**	**0.763**
No ARCL	0.834	0.867	0.098	0.509	0.645	0.813	0.852	0.594	0.619	0.521
No MIM	0.901	0.923	0.095	0.547	0.732	0.892	0.930	0.674	0.718	0.626
No MHR-CNN	0.862	0.901	0.099	0.515	0.659	0.833	0.889	0.637	0.629	0.573



(a) FB15K-237 (b) CORA

Figure 7: Comparison of core metrics results of graph characterization methods based on multi-hop neighbor aggregation as well as gate mechanism on FB15k-237 and Cora dataset.

model. Figure 7 illustrates the percentage magnitude of improvement achieved by the optimized model relative to the baseline RGCN model, as represented on the vertical axis for each metric assessed. The outcomes depicted in Figure 7 underscore the efficacy of the gate mechanism introduced in this study, which proficiently filters out noise information from neighboring nodes while retaining salient feature information of key adjacent nodes.

✍ **Answer to RQ-3:** ► *In conclusion, the integration of attention mechanisms and the addition of a gate mechanism into the RGCN model led to significant improvements in performance. The final `LGFat-RGCN` model, incorporating FANR and the gate mechanism, demonstrated improved accuracy at the top-k recommendation.* ◄

6 CONCLUSION

In the present investigation, the encoding and classification of EHR are reconceptualized as the construction of adversarial hierarchical labeled graphs. This study introduces the adversarial migration-based labeled graph generation network (`LGFat-RGCN`), which incorporates MHR-CNN and Fat-RGCN modules to capture diverse medical text patterns, as well as a message integration module (MIM) to encode EHR connections. Experimental results on the MIMIC-III benchmark dataset reveal that the `LGFat-RGCN` model notably surpasses multiple comparable baseline models, achieving the highest performance reported thus far. Future research endeavors will focus on augmenting the `LGFat-RGCN` model's performance through the exploration of prior knowledge incorporation, automated hyper-parameter tuning, an enhanced loss function, and optimized graph representation in subsequent phases.

7 ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China under Grant 61973069, 62106003, 2022A1515011474, and 62102265. This research was also supported by Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant: GML-KF-22-29. We would also thank Tianyue Chang (Tsinghua University), Baiqi Li (East China Normal University), Junze Liu (Changchun University of Science and Technology) for their contributions to ideas in this paper.

REFERENCES

[1] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2018. Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence*.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

[3] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, (July 2020), 3105–3114. DOI: 10.18653/v1/2020.acl-main.282.

[4] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. Hypercore: hyperbolic and co-graph representation for automatic icd coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3105–3114.

[5] Niel Chah. 2017. Freebase-triples: a methodology for processing the freebase data dumps. *arXiv preprint arXiv:1712.08707*.

[6] Yuanfei Dai, Shiping Wang, Neal N Xiong, and Wenzhong Guo. 2020. A survey on knowledge graph embedding: approaches, applications and benchmarks. *Electronics*, 9, 5, 750.

[7] Luciano RS De Lima, Alberto HF Laender, and Berthier A Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the seventh international conference on Information and knowledge management*, 132–139.

[8] Matúš Falis, Maciej Pajak, Aneta Lisowska, Patrick Schrempf, Lucas Deckers, Shadia Mikhael, Sotirios Tsaftaris, and Alison O'Neil. 2019. Ontological attention ensembles for capturing semantic concepts in icd code prediction from clinical text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 168–177.

[9] Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2020. Dilated convolutional attention network for medical code assignment from clinical text. *arXiv preprint arXiv:2009.14578*.

[10] Alistair EW Johnson et al. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 1, 1–9.

[11] Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. In *proceedings of the AAAI conference on artificial intelligence* number 05. Vol. 34, 8180–8187.

[12] Junyu Luo, Cao Xiao, Lucas Glass, Jimeng Sun, and Fenglong Ma. 2021. Fusion: towards automated icd coding via feature compression. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2096–2101.

[13] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3, 127–163.

[14] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, (June 2018), 1101–1111. DOI: 10.18653/v1/N18-1100.

[15] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

[16] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21, 2, 231–237.

[17] Aaditya Prakash, Siyuan Zhao, Sadid A Hasan, Vivek Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2017. Condensed memory networks for clinical diagnostic inferencing. In *Thirty-first AAAI conference on artificial intelligence*.

[18] Najmeh Sadoughi, Greg P Finley, James Fone, Vignesh Murali, Maxim Korenevski, Slava Baryshnikov, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. Medical code prediction with multi-view convolution and description-regularized label-dependent attention. *arXiv preprint arXiv:1811.01468*.

[19] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*. Springer, 593–607.

[20] Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*.

[21] Aaron Sonabend, Winston Cai, Yuri Ahuja, Ashwin Ananthakrishnan, Zongqi Xia, Sheng Yu, and Chuan Hong. 2020. Automated icd coding via unsupervised knowledge integration (unite). *International journal of medical informatics*, 139, 104135.

[22] Fabian Stephany and Fabian Braesemann. 2017. An exploration of wikipedia data as a measure of regional knowledge distribution. In *International Conference on Social Informatics*. Springer, 31–40.

[23] Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 01. Vol. 34, 222–229.

[24] Qiuling Suo, Fenglong Ma, Ye Yuan, Mengdi Huai, Weida Zhong, Jing Gao, and Aidong Zhang. 2018. Deep patient similarity learning for personalized healthcare. *IEEE transactions on nanobioscience*, 17, 3, 219–227.

[25] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3, 1, 17.

[26] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for icd coding from clinical text. *arXiv preprint arXiv:2007.06351*.

[27] Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2321–2331.

[28] Shanshan Wang, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jian-Yun Nie, Jun Ma, and Maarten de Rijke. 2020. Coding electronic health records with adversarial reinforcement path generation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 801–810.

[29] Zhihao Wang and Xin Li. 2019. Hybrid-te: hybrid translation-based temporal knowledge graph embedding. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 1446–1451.

[30] Lingfei Wu, Jiliang Tang, Yinglong Xia, Jian Pei, and Xiaojie Guo. 2021. The sixth international workshop on deep learning on graphs-methods and applications (dlg-kdd'21). In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 4167–4168.

[31] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32, 1, 4–24.

[32] Pengtao Xie and Eric Xing. 2018. A neural architecture for automated icd coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1066–1076.

[33] Xiancheng Xie, Yun Xiong, Philip S Yu, and Yangyong Zhu. 2019. Ehr coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 649–658.

[34] Keyang Xu et al. 2019. Multimodal machine learning for automated icd coding. In *Machine learning for healthcare conference*. PMLR, 197–215.

[35] Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: multiple synonyms matching network for automatic icd coding. *arXiv preprint arXiv:2203.01515*.

[36] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: a review of methods and applications. *AI Open*, 1, 57–81.

[37] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 207–212.

[38] Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic icd coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5948–5957.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009