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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

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.
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 $G_0$ and the labeled graph discriminator $D_C$. In the following sections, we expound upon the architecture of LGFat-RGCN.

3.1 Labeled Graph Generator $G_0$

The labeled graph generation process is denoted by $<S,A,T,R>$. Within this formulation, $S$ represents the state space, while $A$ constitutes the set of all feasible actions. For example, the subset of $A$ corresponding to a specific label comprises its neighbors in the global graph. The transition function, denoted by $T$, facilitates the progression of state transitions, whereas $R$ signifies the reward function associated with each (state, action) pair. To encourage $G_0$ 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.

$$R(\theta) = E_{\tau \sim \rho(\tau)} [R(\tau)]$$

$$\bar{R}(\theta) = E_{a \sim \pi(a)} \left[ \sum_i R(s = s_i, X = x, a_i) \right]$$

$$\bar{R}_i = R(s = s_i, X = x, a_i)$$

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 $\tau$ represents the trajectory itself. The labeled graph generator is expressed as $G_0$, 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 $D_C$. Subsequently, we elucidate how the policy gradient can be employed to adjust $\theta$, (notably, $R(s = s_i, X = x, a_i)$ is independent of $\theta$):
Figure 2: As delineated in the LGFat-RGCN framework, there are two pivotal components: the Label Generator $G_{θ}$ and the Label Discriminator $D_{ξ}$. A thorough elucidation of MIM, MHR-CNN, and Fat-RGCN will be provided in the ensuing sections.

\[
\nabla \hat{R}(θ) = \sum \sum \pi(a_i | s = s_i, X = x; θ) \nabla \log \pi(a_i | s = s_i, X = x; θ)
\]

(4)

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

\[
\pi(a_i | s = s_i, X = x; θ) = \sigma(W(s_i) + b_i)
\]

(5)

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

3.2 Labeled Graph Discriminator $D_{ξ}$

We devise the trajectory discriminator module $D_{ξ}$ to procure the reward $m_i$ 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:

\[
h_{i} = R_{(s=s_{i}, X=x, a=a_{i})} = p_{s}(\{c_{1}, c_{2}, c_{3}, \ldots, c_{i}\}, x) = \sigma(M_{h}(LSTM(h_{k-1}, c_{i}) \oplus x))
\]

(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 $D_{ξ}$, we employ a cross-entropy function, which is defined as:

\[
Loss_{ξ} = - \sum_{(y_{i}, x) \in S^+} \log p_{s}(y_{i}, x) - \sum_{(y_{i}, x) \in S^-} \log(1 - p_{s}(y_{i}, x))
\]

(7)

In this expression, $S^+$ and $S^-$ correspond to positive and negative samples, respectively, while $p_s(y_i, x)$ 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:

$$s_t = (1 - G(r, W_g)) r + G(r, W_g) m_t$$

(8)

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

3.3.1 EPMR. The symbolic representation of the relationship be-
tween an EHR and an ICD trajectory, denoted by $g_t$, can be generated
as elaborated below:

$$r_t = (x_t \cdot p_t)^- (x_t + p_t)^- (p_t - x_t)$$

$$g_t = \tanh(W_p(x_t^r p_t^r r_t))$$

(9)

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

3.3.2 PCMP. PCMP is employed to capture the relationship be-
tween parent and child ICD codes of ICD code $r_t$. The association
between an EHR and an ICD trajectory is characterized as $p_t$. Subse-
quently, this relational representation is propagated from the parent
code to all its child codes, generating the relation representation $m_t$:

$$n_i = r_i \cdot s_i^p$$. Here, $\cdot$ signifies the element-wise product operation, and
$s_i^p$ 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:

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

(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 $b_i$.

4 MHR-CNN for $G_\theta$’s Embedding

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

$$F_t = f_t(X) = \sum_{j=1}^{l} \tanh(W_t^j X^{j+k_1-1})$$

$$F_m = f_m(X) = \sum_{j=1}^{l} \tanh(W_t^j X^{j+k_m-1})$$

(11)

In this representation, $\sum_{j=1}^{l}$ denotes the left-to-right convolutional
operations. The sub-matrices of $X$ are indicated by $X^{j+k_1-1} \in \mathbb{R}^{k_1 \times d_f}$ and $X^{j+k_m-1} \in \mathbb{R}^{k_m \times d_f}$. The weight matrices of the cor-
responding filters are represented by $W_1 \in \mathbb{R}^{(k_1 \times d_f)\times d_f}$ and $W_m \in \mathbb{R}^{(k_m \times d_f)\times d_f}$.

$$H_m = f_m(E) = \sum_{j=1}^{n} \tanh(W_t^j X^{j+k_m-1})$$

(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:

$$I_1 = c_{ni}(I) = \sum_{j=1}^{l} \tanh(W_t^j X^{j+k_m-1})$$

$$I_2 = c_{ni}(I_1)$$

(13)

The calculation process of a Multi-Residual Convolutional Block
(MCB) is represented by the symbol $\sum_{j=1}^{l}$, which denotes a se-
quence of convolutional operations. $I$ is the input matrix of the block,
and $X^{j+k_m-1} \in \mathbb{R}^{k_m \times d_f}$ represents its submatrices. The weight
matrices of the three convolutional filters, namely $c_{n1i}$, $c_{n2i}$, and $c_{n3i}$,
are represented by $W_{ni1} \in \mathbb{R}^{(k_1 \times d_f)\times d_f}$ and $W_{ni2} \in \mathbb{R}^{(k_2 \times d_f)\times d_f}$.

3.5 Fat-RGCN for $D_c$’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 $n_{ij}$.

$$\beta^T [A_m||A_n] = [\beta_a + \beta_n]^T [A_m||A_n]$$

$$= \beta_n A_m + \beta_a A_n$$

(14)

In practice, the original GAT model’s parameters are separated
into those of $s_{ij}$ and $n_{ij}$; the Attention parameter $\alpha$ represents the
overall GAT.

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

$$\begin{bmatrix}
    s_{i1} & s_{i2} & \cdots & s_{ij} \\
    n_{i1} & n_{i2} & \cdots & n_{ij}
\end{bmatrix}$$

(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 tech-
niques, 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.

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

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

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

The remaining processing techniques are the same as in the first scheme, where \( \beta_O \) stands for the modified GAT’s Attention aggregation approach.

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

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

As a key step in the proposed methodology, the node representation \( X_n \) and relationship representation \( N_r \) are first combined using the \( \gamma \) function. Subsequently, we introduce the use of \( \beta_O \) 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.

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

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

\[
C_u = (1 - D (C_{uj})) \cdot C_{uj} + D (C_{uj}) \cdot C_{ui}
\]  
(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 \( \beta_O \) is presented below.

\[
C_u = (1 - D (C_{uj})) \cdot C_{uj} + D (C_{uj}) \cdot C_{ui}
\]  
(22)

\[
D (C_{uj}) = \sigma (X + A_{ui})
\]  
(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.

廋 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.

廋 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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>MIMIC-III Full</th>
<th>MIMIC-III Top50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>F1</td>
</tr>
<tr>
<td></td>
<td>Macro Micro</td>
<td>Macro Micro</td>
</tr>
<tr>
<td>Hierarchy-SVM</td>
<td>0.456</td>
<td>0.438</td>
</tr>
<tr>
<td>Flat-SVMs</td>
<td>0.482</td>
<td>0.467</td>
</tr>
<tr>
<td>C-MemNN</td>
<td>0.833</td>
<td>0.913</td>
</tr>
<tr>
<td>C-LSTM-Att</td>
<td>0.831</td>
<td>0.908</td>
</tr>
<tr>
<td>BI-GRU</td>
<td>0.500</td>
<td>0.547</td>
</tr>
<tr>
<td>HA-GRU</td>
<td>0.501</td>
<td>0.509</td>
</tr>
<tr>
<td>CAML</td>
<td>0.897</td>
<td>0.959</td>
</tr>
<tr>
<td>DR-CAML</td>
<td>0.897</td>
<td>0.961</td>
</tr>
<tr>
<td>LAAT</td>
<td>0.919</td>
<td>0.963</td>
</tr>
<tr>
<td>JointLAAT</td>
<td>0.941</td>
<td>0.965</td>
</tr>
<tr>
<td>DD</td>
<td>0.938</td>
<td>0.969</td>
</tr>
<tr>
<td>MSMN</td>
<td>0.943</td>
<td>0.965</td>
</tr>
<tr>
<td>FUSION</td>
<td>0.915</td>
<td>0.964</td>
</tr>
<tr>
<td>LGFat-RGCN</td>
<td>0.989</td>
<td>0.998</td>
</tr>
</tbody>
</table>

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.

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>MIMIC-III Full</th>
<th>MIMIC-III Top50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>F1</td>
</tr>
<tr>
<td></td>
<td>Macro Micro</td>
<td>Macro Micro</td>
</tr>
<tr>
<td>LGFat-RGCN</td>
<td>0.983</td>
<td>0.998</td>
</tr>
<tr>
<td>No ARCL</td>
<td>0.834</td>
<td>0.867</td>
</tr>
<tr>
<td>No MIM</td>
<td>0.901</td>
<td>0.923</td>
</tr>
<tr>
<td>No MHR-CNN</td>
<td>0.862</td>
<td>0.901</td>
</tr>
</tbody>
</table>

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.

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 hyperparameter 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


