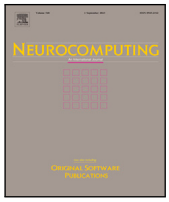
Neurocomputing 592 (2024) 127736

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Graph learning with label attention and hyperbolic embedding for temporal event prediction in healthcare

Usman Naseem [a,*], Surendrabikram Thapa [b], Qi Zhang [c], Shoujin Wang [d], Junaid Rashid [e], Liang Hu [c], Amir Hussain [f]

[a] School of Computing, Macquarie University, Sydney, Australia
[b] Virginia Tech, Blacksburg, 24060, VA, USA
[c] Tongji University, Shanghai, 200092, Shanghai, China
[d] University of Technology Sydney, Sydney, 2007, NSW, Australia
[e] Department of Data Science, Sejong University, Seoul, 05006, Seoul, Republic of Korea
[f] School of Computing, Edinburgh Napier University, Edinburgh, EH11 4BN, Scotland, United Kingdom

ARTICLE INFO

Communicated by N. Zeng

Keywords:
Temporal event prediction
Hierarchical embeddings
Graph neural networks
Clinical notes

ABSTRACT

The digitization of healthcare systems has led to the proliferation of electronic health records (EHRs), serving as comprehensive repositories of patient information. However, the vast volume and complexity of EHR data present challenges in extracting meaningful insights. This paper addresses the need for automated analysis of EHRs by proposing a novel graph learning model with label attention (GLLA) for temporal event prediction. GLLA utilizes graph neural networks to capture intricate relationships between medical codes and patients, incorporating hierarchical structures and shared risk factors. Furthermore, it introduces the Label Attention and Attention-based Transformer (LAAT) algorithm to analyze unstructured clinical notes as a multi-label classification problem. Evaluation on the widely-used MIMIC III dataset demonstrates the efficacy of GLLA in enhancing diagnostic prediction performance. The contributions of this research include a comprehensive analysis of existing models, the identification of limitations, and the development of innovative approaches to improve the accuracy and effectiveness of EHR analysis. Ultimately, GLLA aims to advance healthcare decision-making, disease management strategies, and patient outcomes.

1. Introduction

Artificial intelligence and machine learning have transformed healthcare systems worldwide. These technologies have revolutionized the analysis and interpretation of vast amounts of healthcare data, offering unprecedented opportunities to improve patient outcomes, optimize resource allocation, and enhance decision-making processes [1,2]. In recent years, the digitization of healthcare systems has led to the widespread use of electronic health records (EHRs), transforming how patient data is stored and accessed [3,4]. EHRs contain comprehensive information such as medical histories, clinical notes, diagnostic tests, and treatment plans. However, the sheer volume and complexity of EHR data make it challenging for healthcare professionals to extract meaningful insights and make informed decisions promptly.

To address this challenge, researchers and practitioners are turning to machine learning and artificial intelligence to automate the analysis of EHRs [5]. By leveraging these technologies, valuable patterns, correlations, and predictive signals can be discovered within the vast amount of data. This, in turn, can improve healthcare outcomes, enhance disease prevention strategies, and enable personalized treatment plans.

One crucial area of focus in EHR analysis is the prediction of temporal events, such as future diseases and heart failure. Traditional approaches based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown promise in event prediction. However, they often overlook important aspects of the data that can significantly impact prediction accuracy.

The International Classification of Diseases, 9th Revision (ICD-9) coding system provides a standardized representation of diseases and surgeries, forming a hierarchical structure with unique codes assigned to each condition. Effectively utilizing this structured data is essential for developing accurate prediction models. Graph learning and graph embedding techniques have emerged as powerful tools for representing and understanding complex relationships within hierarchical data. Traditional approaches based on Euclidean space embeddings have limitations in capturing the hierarchical structures inherent in medical

* Corresponding author.
E-mail address: usman.naseem@mq.edu.au (U. Naseem).

https://doi.org/10.1016/j.neucom.2024.127736
Received 7 January 2024; Received in revised form 28 March 2024; Accepted 20 April 2024
Available online 24 April 2024

code data [6]. However, the introduction of hyperbolic embeddings, which represent the hierarchy in hyperbolic space, shows promise in better capturing hierarchical relationships [7].

In the field of healthcare prediction using EHR, two main directions have been pursued. The first direction focuses on uncovering connections between diseases using network analysis to identify shared risk factors and comorbidities [8–10]. These methods leverage disease relationships to improve disease representation and prediction accuracy [11]. The second direction involves applying natural language processing (NLP) techniques to clinical notes [12–14]. Medical diagnosis tasks can be treated as multi-label classification problems, and various models based on CNNs and LSTM networks have outperformed traditional machine-learning approaches in this domain [5,15,16]. However, CNN and LSTM models have their limitations. CNN models often require multiple layers to capture all the text information, while LSTM models face challenges in parallelization during training and inference. The BERT model, based on transformer architecture, has revolutionized NLP but faces challenges when applied to medical code prediction due to unstructured clinical notes, large output code spaces, and long-tail sparsity issues.

Addressing such issues, this paper proposes a **G**raph **L**earning model with **L**abel **A**ttention (GLLA) for temporal event prediction in healthcare to address these challenges and improve diagnostic prediction performance. The research focuses on two main aspects. Firstly, it uses graph neural networks to capture the intricate relationships between medical codes and patients, incorporating hierarchical structures and considering shared risk factors. Secondly, it analyzes unstructured clinical notes by transforming the task into a multi-label classification problem and introduces the use of Label Attention and Attention-based Transformer (LAAT) algorithm to extract informative features from clinical text.

To evaluate the proposed model, the freely-available MIMIC III dataset is utilized, providing diverse patient information. The contributions of this research lie in the analysis of existing models, identification of limitations, and the development of novel approaches to enhance prediction accuracy and effectiveness in diagnosing patients' health events. By leveraging graph learning techniques and addressing the challenges in medical code prediction, this work aims to contribute to the automation of EHR analysis, ultimately leading to improved healthcare decision-making, more effective disease management, and better patient outcomes.

2. Related works

In the field of healthcare prediction and analysis, several studies have been conducted to leverage various techniques and models for accurate diagnosis and temporal event prediction. Understanding the existing research in this domain is crucial for building upon prior knowledge and identifying the gaps that the current study aims to address. In this section, we provide a comprehensive review of the relevant literature, highlighting the contributions and limitations of existing models and techniques in the context of healthcare prediction.

Previous studies have explored the hierarchical structure of ICD-9 codes and the significance of proper data representation for improved model performance. For instance, Li and Yu [15] investigated the hierarchical embedding method for ICD-9 codes, demonstrating the importance of capturing the hierarchical relationships between diseases. They employed graph neural networks to model patient-disease and disease ontology graphs, improving the prediction accuracy of common diseases and heart failure. Graph neural networks have shown promising results in temporal event analysis in various healthcare tasks [17–19].

However, traditional embedding algorithms based on Euclidean space have shown limitations in effectively representing hierarchical information [6]. To address this, Nickel and Kiela [7] introduced an approach that learns low-dimensional representations of hierarchical structure graphs in hyperbolic space. By leveraging hyperbolic embedding, researchers have been able to capture the hierarchical relationships between medical codes more effectively. This approach has shown promise in improving the accuracy and interpretability of healthcare prediction models.

In the domain of clinical text analysis, natural language processing techniques have been applied to extract meaningful information from unstructured clinical notes. Multi-label classification models based on Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have demonstrated superior performance in medical diagnosis tasks compared to conventional machine learning approaches [5,12,13,15,16,20–22]. These models leverage the inherent sequential and contextual information present in clinical notes, enabling accurate prediction of disease labels.

However, CNN models often require multiple layers to capture the complete text information, resulting in increased computational complexity. On the other hand, LSTM models face challenges in parallelization during training and inference. To address these limitations, the BERT model [23] and transformer-based models have emerged as powerful tools for NLP tasks. Nevertheless, applying transformer models to medical code prediction remains challenging due to the unstructured nature of clinical notes, the large output code spaces, and the long-tail sparsity issue [15,24].

Lu et al. [11] proposed a collaborative graph learning model, CGL, for temporal event prediction in healthcare. This model sought to improve upon existing algorithms based on convolutional neural networks (CNNs) and attention mechanisms. The researchers employed a hierarchical embedding method for ICD-9 codes, specifically focusing on the hierarchical relationships within the dendritic structure diseases code. They implemented a collaborative graph neural network to capture the relationships between patients and diseases, as well as the disease ontology graph. Additionally, they designed a TF–IDF approach to extract information from unstructured data, such as clinical notes. The incorporation of domain knowledge, graph neural networks, and TF–IDF analysis in the CGL model represented a notable advancement in healthcare prediction, particularly in predicting future diseases and heart failure. By leveraging these techniques, Lu et al. [11] aimed to enhance the accuracy and effectiveness of temporal event prediction in healthcare.

In light of these research directions and their respective limitations, the current study proposes a novel **G**raph **L**earning model with **L**abel **A**ttention (GLLA) for temporal event prediction in healthcare. Our method, GLLA, builds upon the collaborative graph learning model (CGL) [11] and advances in the hyperbolic embedding to enhance the accuracy and interpretability of patient health event predictions. This model aims to address the shortcomings of existing approaches by effectively utilizing domain knowledge through hyperbolic embedding, incorporating hierarchical structures, and introducing the Label Attention and Attention-based Transformer (LAAT) algorithm [14] for clinical text analysis. By leveraging these advancements, the GLLA model aims to improve the accuracy, interpretability, and efficiency of patient health event predictions.

Through an analysis of related works, we have identified the gaps and challenges in existing models and techniques. The subsequent sections of this paper will present the proposed GLLA model in detail, outlining its architecture, training process, and evaluation results.

3. Methodology

The GLLA model consists of three main modules, as depicted in Fig. 1. The first module includes a hyperbolic embedding layer, collaborative graph learning layer, and bi-LSTM layer. These components extract hidden features from medical codes, patients' diagnoses, and admission durations. The second module focuses on extracting information from clinical notes using a bi-LSTM layer and label attention

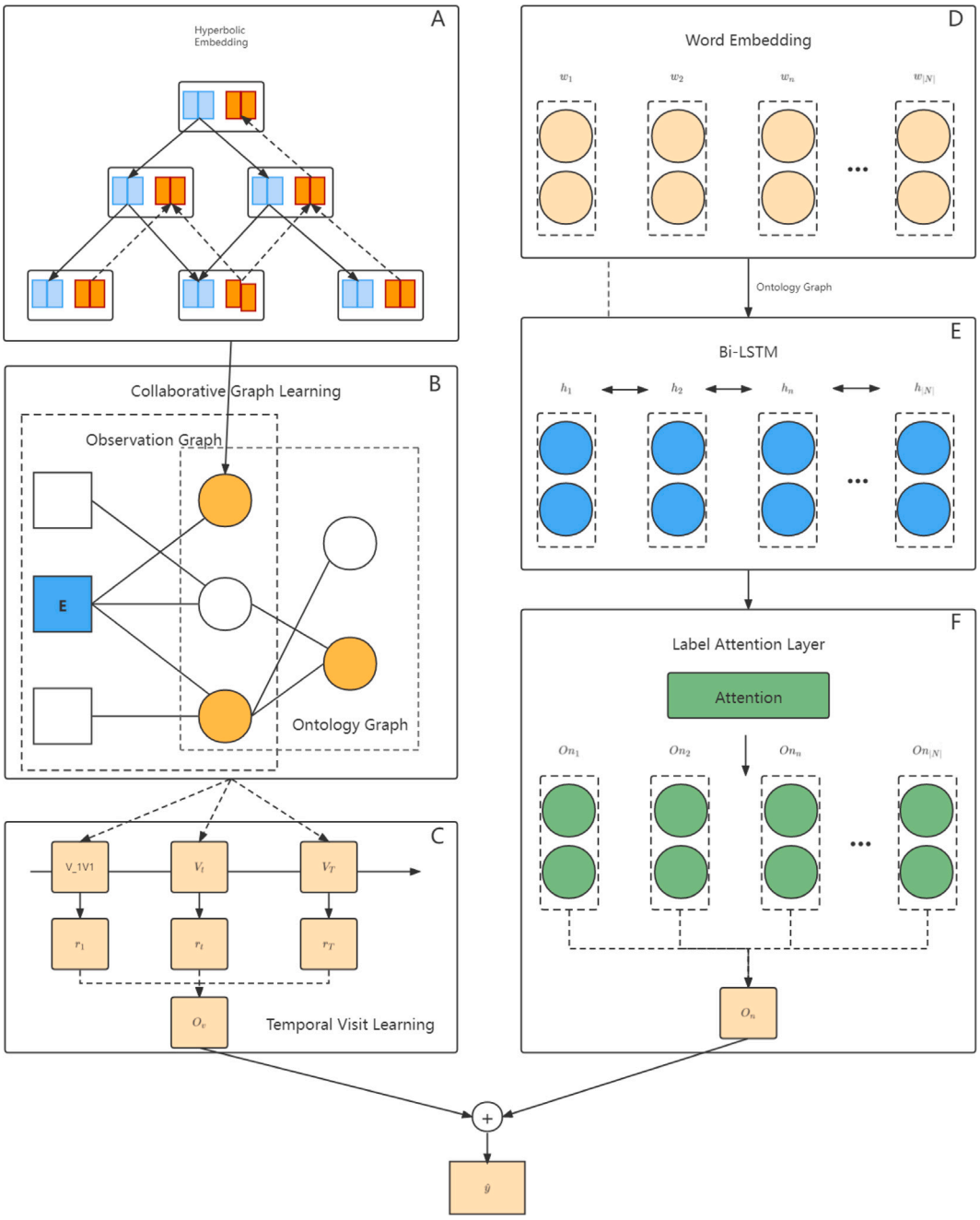


Fig. 1. Proposed architecture with label attention and hyperbolic embedding.

layer. The third module performs classification based on the extracted features. Our model enhances the CGL approach [11] by effectively utilizing diverse information sources, leading to improved performance in classification tasks.

3.1. Hyberbolic embedding layer

The ICD-9 system provides a domain knowledge base that categorizes diseases using medical codes at different levels. However, the original CGL model only utilizes a hierarchical embedding approach that considers the inheritance relationship between codes. To uncover more latent features, we sought a new method.

Drawing inspiration from the work of Nickel and Kiela [7], we turned to the Poincaré ball model, which effectively captures the hierarchical structure inherent in the medical codes system. Assuming a hierarchical structure denoted as Hie with h levels, each node in the structure is encoded into a hyperbolic space using the ball model. Consequently, medical codes, represented as vectors c_i and c_j are embedded as vectors e_i and e_j in the hyperbolic space. The distance between these embedded codes is determined by an Eq. (1), which captures their proximity or dissimilarity.

$$d(e_i, e_j) = cosh^{-1}\left(1 + 2\frac{\|e_i - e_j\|^2}{(1 - \|e_i\|^2)(1 - \|e_j\|^2)}\right) \tag{1}$$

In the dataset, most diseases observed in patients tend to be located at lower levels of the hierarchical structure, corresponding to leaf nodes. However, if there is a high-level disease (non-leaf node) present, virtual child nodes are created to fill the lower levels. This ensures that in the last level of the structure, there are both actual leaf nodes and virtual leaf nodes. Let us assume the set of medical codes as C,

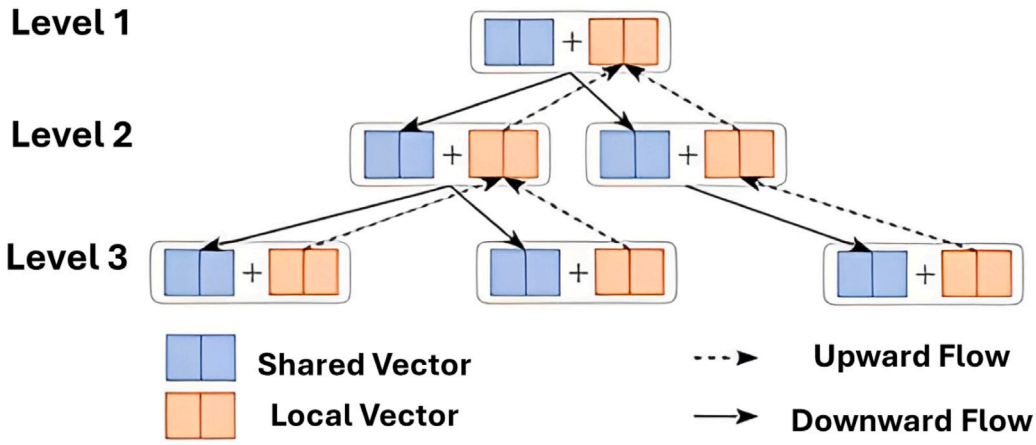


Fig. 2. Hyperbolic embedding with information flow.

and denote the number of diseases as $\|Hie\|$ in Hie, the hierarchical structure.

The hierarchical structure, Hie, follows a pattern where higher-level diseases contain more general information, while lower-level diseases provide more specific descriptions of the diseases themselves. In other words, nodes at higher levels act as ancestors, while nodes at lower levels serve as their children. To capture both the similarities and hierarchical information among diseases, an information flow strategy [11] is employed.

As shown in Fig. 2, the strategy involves two directions of information flow: upward and downward. The upward flow summarizes information from ancestor nodes to their parents, while the downward flow carries information from parents to their children. To implement the information flow strategy, two trainable embedding vectors, $s_i \in \mathbb{R}^d$ and $t_i \in \mathbb{R}^d$, are initialized to form a medical code representation c_i. In our implementation, s_i represents similarity, while t_i represents distinction. Additionally, e'_i is set as the public embedding vector of c_i whereas λ_i is used as a trainable coefficient to combine s_i and t_i, yielding the calculation of e'_i as shown in Eq. (2).

$$e'_i = \lambda_i \times s_i + t_i \times (1 - \lambda_i) \in \mathbb{R}^d \tag{2}$$

Setting a node $c_j \in Hie$ as the parent of another node c_j, and $c_k \in Hie$ as a child of c_i, and n_i as the number of c_i's children, the downward flow is calculated as Eq. (3), while the upward flow is calculated as Eq. (4).

$$\text{downward flow} : s'_i = e'_j \tag{3}$$

$$\text{upward flow} : t'_i = \frac{1}{n_i} \sum_{k=1}^{n_i} t_k \tag{4}$$

According to Lu et al. [11], the assumption is made that the distance between two nodes should be large if they are not connected, and small when they are connected. Using the equation of hyperbolic distance (X), the loss function, L_{rec} for reconstructing the hierarchical structure is defined. The set of connected node pairs in Hie is denoted as $A = \{(i,j)|c_i, c_j \in Hie\}$, and nodes not connected to c_i are in the set $N(i) = \{j'|(c_i, c'_j) \notin A\}$. This loss function in Eq. (5) helps pre-train the medical code embeddings through backpropagation.

$$\mathcal{L}_{rec} = -\sum_{(i,j)\in A} log \frac{e^{-d(e_i,e_j)}}{\sum_{j' \in N(i)\cup\{v\}} e^{-d(e_i,e_{j'})}} \tag{5}$$

Ultimately, the collection of medical code embeddings $E \in \mathbb{R}^{|c|\times d}$, generated through this process, serves as input for the subsequent collaborative graph learning layer.

3.2. Collaborative graph learning layer

The collaborative graph learning layer (CGL) serves as a crucial component in the model proposed by Lu et al. [11]. It was founded on two reasonable and logical hypotheses:

a. Patients who share similar medical records in their history are more likely to develop the same disease in the future.
b. Diseases belonging to the same higher-level category tend to share common causes, symptoms, and complications.

Based on these hypotheses, we construct a collaborative graph, denoted as $G = \{G_{PC}, G_{CC}\}$, consisting of an observation graph G_{PC} and an ontology graph G_{CC}. The observation graph captures the relationship between patients and diseases, represented by the patient-disease adjacency matrix $A_{PC} \in \{0,1\}^{|P|\times|C|}$. An edge (p, c_i) is added to the graph when a patient is diagnosed with a specific disease c_i, and the corresponding entry in $A_{PC}[p][i]$ is set to 1.

The ontology graph, denoted as G_{CC}, represents the horizontal relationship between medical codes. Nodes in G_{CC} are pure medical codes. Another adjacency matrix, $B_{CC} \in \{0,1\}^{|C|\times|C|}$, is created for simulating the horizontal relationship between every two codes. In the ICD-9 system, diseases that share a common ancestor are likely to exhibit similarities. To simulate this relationship, an edge, $(c_i, c_j)_t$, is added to the graph when two codes have the same ancestor in the lowest level, l. Also, $B_{CC}[i][j]$ will be set as l. For any code, c_i, $B_{CC}[i][i] = 0$. So, the B_{CC} matrix here is a dense matrix which will cause the calculation to become more complex. Therefore, we make a B'_{CC} matrix initialized with zeros. $B'_{CC}[i][j]$ and $B'_{CC}[j][i]$ will be set as 1 when two disease c_i and c_j show in one patient's medical history. Let $A_{CC} = B_{CC} \odot B'_{CC}$, thus we can ignore the pairs of codes that never happen in the dataset.

Unlike the method used by Miotto et al. [25], to calculate patient embeddings and extract hidden features, we assign initial embeddings to each patient. The patient embedding matrix is represented as $P \in \mathbb{R}^{|P|\times d_p}$, where d_p represents the dimension of each patient's embedding. Similarly, we have $H_p^{(k)} \in \mathbb{R}^{|P|\times d_p^{(k)}}$ to represent hidden features of patient, p and $H_c^{(k)} \in \mathbb{R}^{|C|\times d_c^{(k)}}$ to indicate hidden features of code, c. Here, k represents the level of the layer the patient input is given into and we set $H_p^{(0)} = P$ and $H_c^{(0)} = E$ in our implementation.

To retrieve the hidden features of codes and patients in the next layer, we map each patient's embedding, $H_c^{(k)}$ into the patient dimension and aggregate it with the corresponding code embeddings, A_{PC}.

$$X_p^{(k)} = H_p^{(k)} + A_{PC} H_c^{(k)} W_{CP}^{(k)} \in \mathbb{R}^{|P|\times d_p^{(k)}} \tag{6}$$

We use $W_{CP}^{(k)}$ in Eq. (6) to map code embeddings, H_c^k to patient embeddings. To aggregate code c_j into code c_i, an ontology weight,

ϕ_j will be assigned to c_j, if c_i and c_j are connected in level l. Thus, we now have:

$$\phi_j(l) = \sigma(\mu_j \times l + \theta_j) \tag{7}$$

In the monotonic equation (7), $\phi_j(l)$, μ_j and θ_j are the trainable variables and σ is a sigmoid activation function. Through increasing and decreasing weights based on levels, the model can reinforce the influence of horizontal relationships between medical codes. Since A_{CC} stores the connected nodes, we have:

$$\Phi = \sigma(M \times A_{CC} + \Theta) \in \mathbb{R}^{|C|\times|C|} \tag{8}$$

where, $\Phi \in \mathbb{R}^{|C|\times|C|}$ is the collection of ϕ_j weights, $\Theta \in \mathbb{R}^{|C|}$ as the collection of θ_j and M is the collection of μ_j.

To map $H_p^{(k)}$ into the codes' dimension and then aggregate it with patient adjacency, we have:

$$X_c^{(k)} = H_c^{(k)} + A_{PC}^T H_p^{(k)} W_{PC}^{(k)} + \Phi H_c^{(k)} \in \mathbb{R}^{|C|\times d_c^{(k)}} \tag{9}$$

Here, $W_{PC}^{(k)} \in \mathbb{R}^{d_p^{(k)}\times d_c^{(k)}}$ is a matrix that maps patient embeddings $H_p^{(k)}$ into the code embedding dimensions. Afterwards, the next layer's $H_c^{(k)}$ and $H_p^{(k)}$ are generated based on the equation below:

$$H_{\{p,c\}}^{(k+1)} = ReLU\left(BatchNorm\left(X_{\{p,c\}}^{(k)} W_{\{p,c\}}^{(k)}\right)\right) \tag{10}$$

The $W_{\{p,c\}}^{(k)}$ here helps map $X_{\{p,c\}}^{(k)}$ into the next layer. The $BatchNorm$ is for normalizing features, thus, stabilizing the network during training. Additionally, only $H_c^{(k)}$ is calculated into K graph layers, due to only medical codes being required for further calculation. We use $H_c^{(k)} \in \mathbb{R}^{|C|\times d_c^{(k)}}$ as the final code embeddings. In summary, the collaborative graph learning layer incorporates the relationships between patients and diseases, as well as the hierarchical structure of medical codes, to extract meaningful hidden features and facilitate temporal event prediction.

3.3. Feature extraction for visits

To capture the temporal features between each visit of a patient, we employ an LSTM layer that treats visits as sequences. For each patient, p, an embedding, v_t is computed for each visit, t as follows:

$$v_t = \frac{1}{|C_t|}\sum_{c_i \in C_t} H_c^i \in \mathbb{R}^{d_c^{(k)}} \tag{11}$$

Information of neighboring codes of code c_i is stored in H_c^i through the connection of patient nodes. Through this function, the model becomes capable of predicting a disease that never occurred in a patient's medical history. LSTM is employed on each v_t to extract the temporal features of a visit. We use $R = \{r_1, r_2, \ldots, r_T\}$ to indicate these temporal features and use q as the size of the LSTM cell as follows:

$$R = r_1, r_2, \ldots, r_T = LSTM(v_1, v_2, \ldots, v_T) \in \mathbb{R}^{T\times q} \tag{12}$$

Additionally, a strategy of location-based attention [26] is employed to get the hidden representations of O_v for all visits:

$$\alpha = softmax(Rw_\alpha) \in \mathbb{R}^T \tag{13}$$

$$O_v = \alpha R \in \mathbb{R}^q \tag{14}$$

Each visit is assigned attention weight α and α has a context vector w_a for attention.

3.4. Bi-LSTM layer

In the original CGL model, only TF–IDF rectified attention was used to guide note embedding [11], which had a limited impact on the final results. Recognizing that clinical notes contain more effective features, we incorporate the LAAT (Language-Aware Attention for Text Classification) module proposed by Vu et al. [14]. LAAT precisely handles unstructured clinical notes as it utilizes attention mechanisms to focus on relevant parts of the clinical notes, enabling the model to attend to important words or phrases that contribute to the prediction of multiple labels. LAAT employs a transformer architecture, which is well-suited for capturing long-range dependencies and contextual information in sequential data like clinical notes. The transformer model processes the clinical text at the token level, allowing it to capture intricate linguistic nuances and relationships between words. This module enables accurate diagnosis prediction based solely on clinical notes.

Before inputting the notes into the bi-LSTM layer, we employ word2vec to pre-train word embeddings. Assuming there are n words in the notes corpus, we represent the word tokens as $w_1, w_2, w_3, \ldots, w_i, \ldots, w_n$. The pre-trained embedding for each word, w_i is denoted as e_{w_i}, and the embedding size is d_e. A bi-LSTM layer is then utilized to extract contextual information from these words. The bi-LSTM layer learns latent feature vectors for input words within a sequence. The hidden states of the ith word in the LSTM is calculated as:

$$\begin{aligned}\overrightarrow{h_i} &= \overrightarrow{LSTM}(e_{w_1:w_i})\\ \overleftarrow{h_i} &= \overleftarrow{LSTM}(e_{w_1:w_i})\end{aligned} \tag{15}$$

The size of h_i is $2m$ when we set the dimension of hidden states of LSTM as m. Then, we concatenate all the words' hidden vectors to build a matrix $H_{all} = [h_1, h_2, \ldots, h_n] \in \mathbb{R}^{2m\times n}$.

3.5. Attention layer

To transform H_{all} into label-specific vectors, we employ the label attention mechanism [14] since clinical notes often have multiple labels. Using H_all as the input, and $|L|$ as the output, the label-specific weight vectors are calculated as:

$$\begin{aligned}Z &= tanh(W H_{all})\\ Weight_L &= softmax(UZ)\end{aligned} \tag{16}$$

In the above Eq. (16), in the matrix $W \in \mathbb{R}^{d_a\times 2m}$, d_a is the hyperparameter which will be tuned, generating a matrix $Z \in \mathbb{R}^{d_a\times n}$. Then to calculate the weight matrix, $Weight_L \in \mathbb{R}^{|L|\times n}$, Z is multiplied with another matrix $U \in \mathbb{R}^{d_a\times|L|}$. Each ith row in $Weight_L$ representing a weight vector corresponding to the ith label in L. The $softmax$ here is to guarantee the sum of weights equals to 1 in each row of $Weight_L$. Label-specific vectors then is computed by multiplying with H_{all} as: $O_n = H_{all} Weight_L^T$. Here each of the matrix $O_n \in \mathbb{R}^{|L|\times wm}$ representing the note input corresponding to the ith label in L. After all the feature extraction layers mentioned above, each patient now has the final embedding $O = O_v \bigoplus O_n$.

3.6. Output layer

In our modified layer based on CGL, we remove the TF–IDF rectified penalty loss as it is not utilized in this model. Both the diagnosis prediction and heart failure prediction tasks now utilize a fully connected layer with a sigmoid function on the output layer. Thus, the final loss function to train the model is defined as:

$$\mathcal{L} = CrossEntropy(\hat{y}, y) \tag{17}$$

The ground truth for heart failure or medical codes, depending on the task, is denoted as y in the loss function.

3.7. Incorporation of hierarchical structures and shared risk factors by GLLA

GLLA model architecture is designed to effectively capture relationships between medical codes and patients by leveraging graph neural networks (GNNs) and incorporating hierarchical structures and shared risk factors. A brief explanation of the GLLA model architecture and how it achieves these objectives is given below:

I. **Graph Neural Networks (GNNs)**: GLLA utilizes GNNs to represent and analyze the relationships between medical codes and patients. GNNs are well-suited for modeling complex graph structures, making them ideal for capturing the intricate relationships present in healthcare data. In the context of GLLA, GNNs enable the model to learn representations of patients and medical codes based on their interactions within the graph.

II. **Incorporating Hierarchical Structures**: GLLA incorporates hierarchical structures by considering the hierarchical relationships between medical codes, such as those defined in the International Classification of Diseases (ICD) coding system. By encoding the hierarchical relationships between codes, GLLA can capture the inherent structure of disease ontology and leverage this information to improve prediction accuracy. For example, GLLA may assign higher importance to codes that are more closely related within the hierarchical structure, reflecting their shared attributes or characteristics.

III. **Shared Risk Factors**: GLLA incorporates shared risk factors by analyzing the connections between patients and medical codes within the graph. Shared risk factors refer to common attributes or conditions that may increase the likelihood of certain medical outcomes. By examining the co-occurrence patterns of medical codes across patient records, GLLA can identify shared risk factors and leverage this information to improve predictive performance. For instance, if multiple patients with similar demographic characteristics exhibit a particular set of medical conditions, GLLA may learn to associate those conditions with the shared risk factors present in the patient population.

To illustrate these mechanisms, consider the following examples:

- **Hierarchical Structures**: Suppose the ICD coding system is used to represent medical conditions, with codes organized hierarchically based on their clinical characteristics. In this case, GLLA may learn to prioritize codes that are more specific or detailed within the hierarchy when making predictions. For instance, if a patient is diagnosed with a specific subtype of cancer (e.g., malignant neoplasm of the breast), GLLA may assign higher importance to codes representing that subtype compared to more general codes (e.g., neoplasm).
- **Shared Risk Factors**: Imagine a scenario where multiple patients from a similar demographic group (e.g., elderly individuals with a history of cardiovascular disease) are diagnosed with heart failure. In this case, GLLA may learn to identify the shared risk factors (e.g., age, pre-existing conditions) associated with heart failure and use this information to improve prediction accuracy. By analyzing the co-occurrence patterns of medical codes and patient attributes, GLLA can uncover latent relationships and leverage them to make more accurate predictions.

Overall, the GLLA model architecture leverages graph neural networks to capture complex relationships between medical codes and patients, incorporating hierarchical structures and shared risk factors to improve predictive performance in healthcare applications.

3.8. Computation and parallelization in GLLA

Transformers rely heavily on self-attention mechanisms that enable parallel computations across different sequence positions. Unlike LSTMs which have sequential dependencies, self-attention can be computed in parallel, allowing better utilization of modern hardware parallelism. Thus, our GLLA algorithm is more parallelizable than existing LSTM-based techniques. Furthermore, LSTMs have recurrent connections that make parallelization across timesteps difficult. Transformers avoid this recurrence by using self-attention over the entire sequence in parallel. This parallelizability could help speed up training and inference for the clinical notes analysis task compared to using regular LSTM layers.

Table 1
Statistics of the MIMIC-III dataset used in our implementation.

Patient number	7125
Avg. visit number per participant	2.66
Patient number with heart failures	2604
Medical code (disease) number	4795
Avg. code number per visit	13.27
Dictionary size in notes	67,913
Avg. word number per note	4732.38

4. Experiments and discussion

4.1. Dataset

In this paper, we used the freely-available MIMIC-III Clinical Database. The dataset includes various types of data such as laboratory test results, medications, mortality, and more. However, for our purposes, we focused on four specific data files: ADMISSION, DIAGNOSES_ICD, NOTEEVENTS, and PATIENT.

The ADMISSION file contains information about the admission ID and admission time of patients. The DIAGNOSES_ICD file contains the diagnosed medical codes of patients during their different admissions. The NOTEEVENTS file stores the clinical notes of patients during their different admissions. Finally, the PATIENT file includes information such as the birthday and gender of the patients.

4.2. Data analysis

In our implementation, a total of 7125 data points were used. However, it should be noted that some patients had missing notes or admissions, resulting in smaller dataset size. On average, each patient had 2.66 visits, as shown in Table 1.

4.3. Model configuration

In the GLLA model, the embeddings for medical codes and patients are initialized randomly. The embedding sizes are set as follows: 32 for medical codes, 16 for patients, and 100 for notes. The GNN layer is applied twice ($L = 2$) to capture hierarchical and collaborative information. The hidden layer dimensions are set to 64, 128, and 32 for different layers. For the BiLSTM layer, the unit size (q) is set to 200, allowing the extraction of contextual information from the clinical notes. The Adam optimizer is used for both the model and hyperbolic embedding generation. The model is trained for 200 epochs, while the hyperbolic embedding generation is performed for 500 epochs. The learning rate for the model is set to 0.001, determining the step size during optimization. These hyperparameter settings provide the configuration for training the GLLA model and generating the hyperbolic embeddings.

4.4. Baselines for comparison

In our experimental setting, we compare seven state-of-the-art baseline models with the proposed GLLA. We have described each one of them briefly below:

- **RETAIN**: The REverse Time AttentIoN model (RETAIN) [27] uses a two-level neural attention mechanism and representation learning to analyze electronic health record (EHR) data to make predictions in healthcare applications. It first learns a general patient representation from EHR data, then applies a double-layered attention mechanism. These visit-level attention and variable-level attention prioritize recent visits and key medical codes within those visits respectively giving both accuracy and interoperability. The implementation of this model can be found at https://github.com/mp2893/retain.

- **Deepr**: Deepr [28] transforms a patient's electronic medical record (EMR) into a sequence of elements separated by time gaps and hospital transfers. It then applies a convolutional neural network (CNN) to this sequence to identify and combine predictive patterns (motifs) within medical data, without the need for manual feature engineering. This allows Deepr to analyze medical records, even with their irregular timing, to predict future risks and potentially improve healthcare outcomes. The Deepr implementation was done using pyhealth package.[1]
- **GRAM**: GRAM [29] is a graph-based attention model that tackles challenges in applying deep learning to healthcare data, especially with limited samples. It incorporates medical knowledge from ontologies and uses an attention mechanism to make the most of limited patient data. This allows GRAM to learn better representations of medical concepts and achieve good results in predicting future diagnoses, even with less training data compared to traditional methods. We use the official implementation[2] of GRAM for our experimentation.
- **Dipole**: Dipole [30] focuses on the prediction of future illnesses using patients' electronic health records. It addresses the challenges of modeling the sequence of visits and the wealth of medical codes within the EHR data. Dipole's approach utilizes bidirectional neural networks to consider a patient's entire medical history and incorporates attention mechanisms to pinpoint important connections between visits. This not only improves the accuracy of the prediction but also allows healthcare professionals to understand and interpret the reasoning behind the predictions. The implementation of Dipole can be found at https://anonymous.4open.science/r/Dipole-FF7C.
- **Timeline**: Bai et al. [31] propose an RNN-based deep learning model called Timeline to analyze Electronic Health Records (EHR) data, particularly focusing on the challenge of capturing how diseases progress over time. Firstly, it utilizes time decay factors to assign weights to medical codes. These weights dynamically change based on how relevant a code is at different points in the patient's history. Secondly, a disease progression function is implemented to model how specific diseases influence these time decay factors. Finally, an attention mechanism focuses on the most critical medical codes for each patient visit, considering both the codes themselves and the time elapsed since their initial appearance. The implementation of Timeline can be found at https://github.com/tiantiantu/Timeline.
- **MedGCN**: MedGCN [32] uses graph convolutional networks (GCNs) to connect patients, medications, lab tests, and other medical details. By analyzing these connections, MedGCN can recommend medications based on a patient's existing information, even if some lab tests are missing. It can also estimate those missing lab test values. We use the official implementation of MedGCN for our comparison. The code can be found at https://github.com/mocherson/MedGCN.
- **CGL**: As explained in Sections 2 and 3, CGL uses a collaborative graph learning model to explore relationships between patients and diseases. This model incorporates both the structured data and the unstructured text data in EHRs, using an attention mechanism to focus on important details [11]. This model has achieved competitive results in predicting health events compared to prior methods. The official implementation[3] provided by the authors was used for the comparison.

[1] https://pyhealth.readthedocs.io/en/latest/.
[2] https://github.com/mp2893/gram.
[3] https://github.com/LuChang-CS/CGL.

Table 2
Diagnosis prediction and heart failure prediction results.

Models	Diagnosis			Heart Failure	
	w-F_1 (%)	R@20 (%)	R@20 (%)	AUC (%)	F_1 (%)
RETAIN [27]	19.66%	33.90%	42.93%	82.73%	71.12%
Deeper [28]	12.38%	28.15%	37.56%	81.29%	68.42%
GRAM [29]	21.06%	36.37%	45.61%	82.82%	71.43%
Dipole [30]	11.24%	26.96%	36.83%	81.66%	70.01%
Timeline [31]	16.83%	32.08%	41.97%	80.75%	69.81%
MedGCN [32]	20.93%	35.69%	43.36%	81.25%	70.86%
CGL [11]	23.75%	37.35%	49.70%	83.93%	70.22%
GLLA	**33.09%**	**45.16%**	**57.41%**	**89.47%**	**76.10%**

Table 3
Diagnosis prediction and heart failure prediction results.

Models	Diagnosis			Heart failure		
	w-F_1 (%)	R@20 (%)	Params.	AUC (%)	F_1 (%)	Params.
$GLLA_{h-}$	32.24%	44.58%	9,375,827	88.75%	**77.30%**	7,453,433
$GLLA_{LA-}$	24.33%	36.56%	7,310,000	82.93%	72.15%	7,310,000
$GLLA_{n-}$	22.52%	36.55%	7,132,000	82.42%	71.28%	7,240,420
GLLA	**33.09%**	**45.16%**	**9,375,827**	**89.47%**	76.10%	**7,453,433**

4.5. Model evaluation

In evaluating the diagnosis predictions, we employ the Macro F1 score and top-k recall metrics. The Macro F1 score provides a comprehensive measure of the model's effectiveness in predicting diagnoses, considering both precision and recall for each individual label. The top-k recall metric assesses the proportion of true positive diagnoses that are captured within the top-k predicted diagnoses. Similarly, for heart failure predictions, we utilize the Macro F1 score and the area under the receiver operating characteristic (ROC) curve, commonly referred to as AUC. Furthermore, the GLLA model is compared against seven state-of-the-art baseline models to conduct a comprehensive analysis of its performance in relation to established approaches.

As shown in Table 2, the proposed GLLA model outperforms all the other models. It is hypothesized that this is mostly because label attention is used on the basis of CGL to make clinic notes more useful.

4.6. Ablation study

Ablation studies are necessary to validate the effectiveness of applied strategies [33,34]. In order to analyze the individual contributions of different components in the GLLA model, three variations of the model were compared through an ablation study: GLLA with hierarchical embedding replaced by hyperbolic embedding ($GLLA_{h-}$), GLLA without the label-attention layer ($GLLA_{LA-}$), and GLLA without clinical notes ($GLLA_{n-}$). The results of the ablation study are presented in Table 3. For the diagnosis prediction task, GLLA achieved the highest w-F1 score of 33.09%, outperforming all ablated versions. However, for heart failure prediction, $GLLA_{h-}$ (with hierarchical embedding replaced with hyperbolic embedding) attained the highest F1 score of 77.30%, surpassing the full GLLA model (76.10%). The hyperbolic embedding, which captures the hierarchical structure of the ICD-9-CM system and extracts bottom-up summary information to identify potential disease relationships, shows improvements in diagnosis prediction. However, it does not yield significant enhancements in heart failure prediction, possibly due to the task's focus on a single disease and the contingent nature of the results. Interestingly, the ablation of the label-attention layer ($GLLA_{LA-}$) and the removal of clinical notes ($GLLA_{n-}$) resulted in substantial performance degradation for both diagnosis and heart

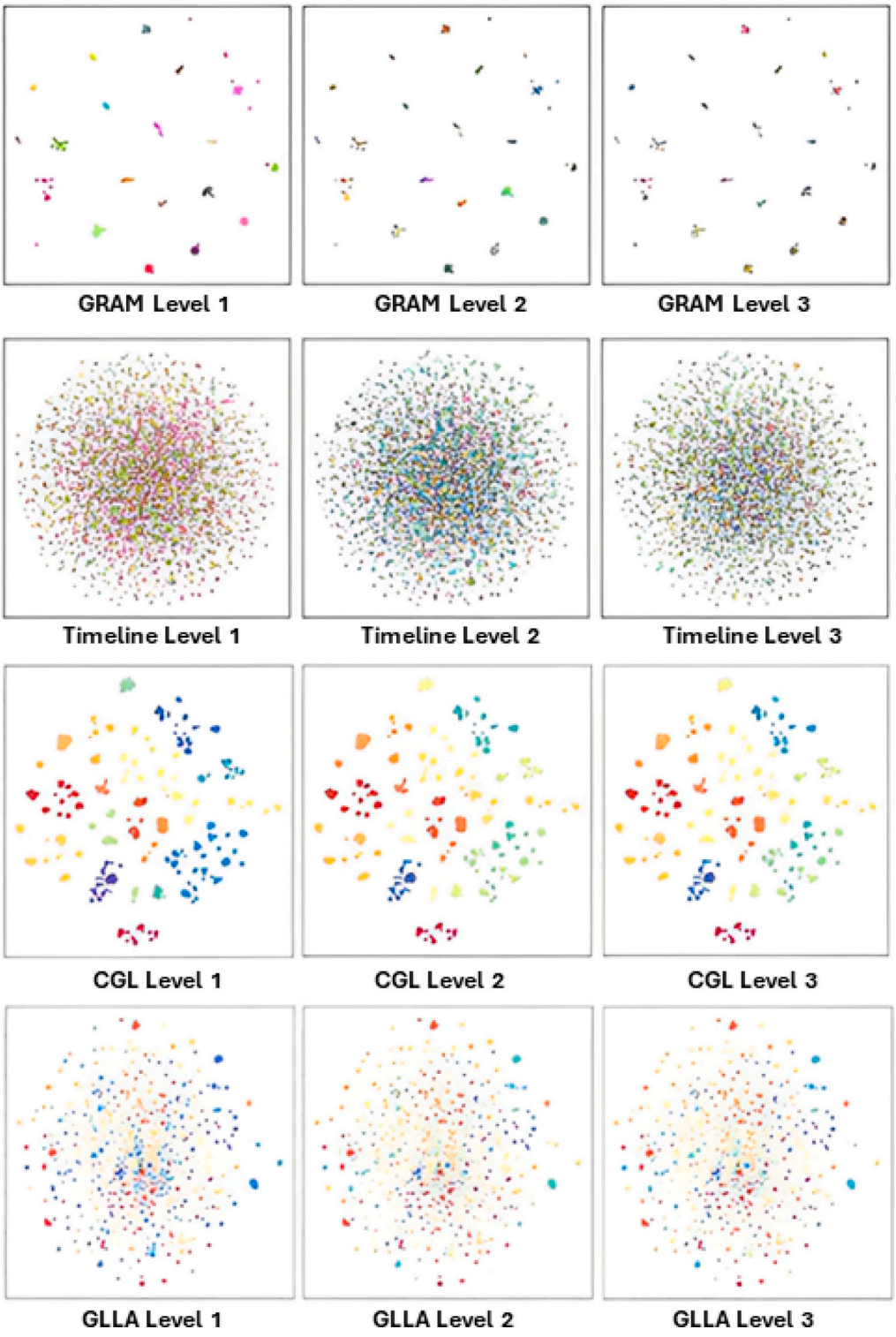


Fig. 3. Code embeddings in 3 levels learned by GRAM, Timeline, CGL and GLLA. Colors correspond to disease types in each level.

failure prediction tasks, highlighting their importance in the overall model's effectiveness.

Overall, the ablation study highlights the complementary strengths of the different components in GLLA. The hyperbolic embedding demonstrates advantages in diagnosis prediction but does not exhibit the same level of effectiveness in heart failure prediction, likely due to the task's characteristics and result variability.

4.7. Prediction analysis

To visualize the relationship between diseases after code embedding, t-SNE (t-Distributed Stochastic Neighbor Embedding) was employed [35]. t-SNE is commonly used to preserve pairwise similarities among neighboring nodes, with relative distances reflecting their similarities. Fig. 3 illustrates the code embeddings learned by GRAM,

Timeline CGL, and GLLA in three levels. Each dot's color represents a different disease type. In CGL and GRAM, dots of the same color are grouped together, but the distances between them are shorter in CGL. This indicates that the code relationships after embedding in CGL closely align with the summarized relationships in the ICD-9 system.

In the GLLA graph, some similar colors are grouped together, while others are distributed randomly. This can be attributed to the hyperbolic embedding in GLLA, which maps code relationships to a hyperbolic space. However, the visualization method, t-SNE, reduces the data dimensionality to a two-dimensional space, which may not accurately reflect the relationships between diseases. Furthermore, compared to the hierarchical structure in CGL, hyperbolic embedding in GLLA can extract bottom-up summary information and uncover potential relationships among diseases, which may have been overlooked by the ICD-9 system.

4.8. Further discussions

While the proposed GLLA model demonstrates significant enhancements in overall diagnostic prediction performance, we acknowledge that there is room for improvement specifically in the heart failure prediction task. One potential challenge we have identified is the imbalanced nature of the dataset, with heart failure cases being underrepresented compared to the diversity of diagnostic codes. This data imbalance could hinder the model's ability to effectively learn the relevant patterns for accurate heart failure prediction. To address this, we plan to explore techniques such as oversampling [36,37], undersampling [37], or specialized class-weighted loss functions to mitigate the impact of imbalanced data.

Moreover, we recognize that heart failure often involves complex temporal dynamics and progressions that may not be fully captured by our current approach of treating patient visits as independent sequences. Integrating alternative architectures that explicitly model such temporal dependencies could potentially enhance the model's ability to predict heart failure accurately. Additionally, we aim to investigate the incorporation of domain-specific knowledge [38,39], such as expert-curated risk factors or comorbidities, which may provide valuable insights for improving heart failure prediction performance. Furthermore, we intend to explore multi-task learning approaches [40–42], where the model is trained simultaneously on diagnosis prediction and heart failure prediction, potentially leveraging shared representations to improve performance on both tasks. Finally, ensemble methods [43,44] that combine predictions from multiple models could lead to more robust and accurate heart failure predictions.

4.9. Ethical considerations

In this research, focused on development of algorithms for healthcare applications, we acknowledge the profound ethical implications and potential impacts on patient outcomes. While the technical evaluation of GLLA demonstrates its effectiveness, we recognize the crucial need to assess and mitigate potential biases and unfairness in our model. The MIMIC-III dataset used for training may itself contain inherent biases or under-representation of certain demographic groups, which could propagate through the model's predictions. To ensure responsible deployment, there is a need to conduct a comprehensive analysis of GLLA's performance across different demographic factors, such as age, gender, and ethnicity, to identify any concerning disparities or biases. Taking the importance of transparency and explainability in consideration, we find a need to develop interpretability methods that provide clear explanations for GLLA's predictions, enabling auditing and building trust among healthcare professionals and patients. Furthermore, there is a need to explore techniques like data augmentation, adversarial debiasing, and incorporating fairness constraints during training to actively mitigate biases present in the data or model. We aim to explore these aspects in future research to make such models more accurate, unbiased, and inclusive.

5. Conclusion

The proposed Graph Learning with Label Attention (GLLA) model offers significant advancements in temporal event prediction in healthcare. By integrating label attention, hyperbolic embeddings, and collaborative graph learning, GLLA addresses several limitations of existing approaches and demonstrates superior performance in diagnostic prediction and heart failure prediction tasks. The comprehensive evaluation on the widely-used MIMIC III dataset showcased GLLA's superior performance compared to state-of-the-art baseline models, demonstrating its efficacy in managing and interpreting the vast and intricate data contained in electronic health records (EHRs). While GLLA exhibited substantial improvements in diagnosis prediction, there is still room for further enhancements in heart failure prediction. Future research could focus on optimizing the model's performance specifically for this task, potentially by exploring additional techniques or incorporating domain-specific knowledge. Overall, this work contributes to the automation of EHR analysis and has the potential to advance healthcare decision-making processes, disease management strategies, and ultimately, improve patient outcomes. By addressing the challenges of temporal event prediction in healthcare, GLLA paves the way for more effective utilization of the vast amounts of data available in modern healthcare systems.

CRediT authorship contribution statement

Usman Naseem: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Surendrabikram Thapa:** Writing – original draft, Methodology, Conceptualization. **Qi Zhang:** Writing – original draft, Methodology. **Shoujin Wang:** Writing – review & editing, Writing – original draft. **Junaid Rashid:** Writing – review & editing, Writing – original draft. **Liang Hu:** Writing – review & editing, Writing – original draft, Supervision. **Amir Hussain:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Usman Naseem reports was provided by James Cook University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We used publicly available data and gave a reference to it in our paper.

References

[1] U. Naseem, S. Thapa, Q. Zhang, L. Hu, J. Rashid, M. Nasim, Incorporating historical information by disentangling hidden representations for mental health surveillance on social media, Soc. Netw. Anal. Min. 14 (1) (2023) 9.
[2] P. Wu, Z. Wang, B. Zheng, H. Li, F.E. Alsaadi, N. Zeng, AGGN: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion, Comput. Biol. Med. 152 (2023) 106457.
[3] M. Hobensack, J. Song, D. Scharp, K.H. Bowles, M. Topaz, Machine learning applied to electronic health record data in home healthcare: a scoping review, Int. J. Med. Inform. 170 (2023) 104978.
[4] E. Hossain, R. Rana, N. Higgins, J. Soar, P.D. Barua, A.R. Pisani, K. Turner, Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review, Comput. Biol. Med. 155 (2023) 106649.
[5] M. Li, Z. Fei, M. Zeng, F.-X. Wu, Y. Li, Y. Pan, J. Wang, Automated ICD-9 coding via a deep learning approach, IEEE/ACM Trans. Comput. Biol. Bioinform. 16 (4) (2018) 1193–1202.
[6] A. Zeb, A.U. Haq, J. Chen, Z. Lei, D. Zhang, Learning hyperbolic attention-based embeddings for link prediction in knowledge graphs, Knowl.-Based Syst. 229 (2021) 107369.

[7] M. Nickel, D. Kiela, Poincaré embeddings for learning hierarchical representations, Adv. Neural Inf. Process. Syst. 30 (2017).
[8] A. Aguado, F. Moratalla-Navarro, F. López-Simarro, V. Moreno, MorbiNet: multimorbidity networks in adult general population. Analysis of type 2 diabetes mellitus comorbidity, Sci. Rep. 10 (1) (2020) 1–12.
[9] B. Fotouhi, N. Momeni, M.A. Riolo, D.L. Buckeridge, Statistical methods for constructing disease comorbidity networks from longitudinal inpatient data, Appl. Netw. Sci. 3 (2018) 1–34.
[10] F. Folino, C. Pizzuti, M. Ventura, A comorbidity network approach to predict disease risk, in: Information Technology in Bio-and Medical Informatics, ITBAM 2010: First International Conference, Bilbao, Spain, September 1-2, 2010. Proceedings, Springer, 2010, pp. 102–109.
[11] C. Lu, C.K. Reddy, P. Chakraborty, S. Kleinberg, Y. Ning, Collaborative graph learning with auxiliary text for temporal event prediction in healthcare, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021.
[12] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, N. Elhadad, Multi-label classification of patient notes: Case study on ICD code assignment, in: 2018 AAAI Joint Workshop on Health Intelligence, W3PHIAI 2018, AAAI Press, 2018, pp. 409–416.
[13] J. Huang, C. Osorio, L.W. Sy, An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes, Comput. Methods Programs Biomed. 177 (2019) 141–153.
[14] T. Vu, D.Q. Nguyen, A. Nguyen, A label attention model for ICD coding from clinical text, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 3335–3341.
[15] F. Li, H. Yu, ICD coding from clinical text using multi-filter residual convolutional neural network, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 8180–8187.
[16] X. Xie, Y. Xiong, P.S. Yu, Y. Zhu, 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, 2019, pp. 649–658.
[17] L.A. Passos, J.P. Papa, J. Del Ser, A. Hussain, A. Adeel, Multimodal audio-visual information fusion using canonical-correlated graph neural network for energy-efficient speech enhancement, Inf. Fusion 90 (2023) 1–11.
[18] S.N. Golmaei, X. Luo, DeepNote-GNN: predicting hospital readmission using clinical notes and patient network, in: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, 2021, pp. 1–9.
[19] A. Sharma, P.K. Singh, P. Nikashina, V. Gavrilenko, A. Tselykh, A. Bozhenyuk, AI and GNN model for predictive analytics on patient data and its usefulness in digital healthcare technologies, in: IoT, Big Data and AI for Improving Quality of Everyday Life: Present and Future Challenges: IOT, Data Science and Artificial Intelligence Technologies, Springer, 2023, pp. 331–345.
[20] J. Chen, F. Teng, Z. Ma, L. Chen, L. Huang, X. Li, A multi-channel convolutional neural network for ICD coding, in: 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering, ISKE, IEEE, 2019, pp. 1178–1184.
[21] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, 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), 2018, pp. 1101–1111.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[23] J.D.M.-W.C. Kenton, L.K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[24] S. Ji, E. Cambria, P. Marttinen, Dilated convolutional attention network for medical code assignment from clinical text, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, 2020, pp. 73–78.
[25] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (1) (2016) 1–10.
[26] M.-T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, 2015, arXiv preprint arXiv:1508.04025.
[27] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, Adv. Neural Inf. Process. Syst. 29 (2016).
[28] P. Nguyen, T. Tran, N. Wickramasinghe, S. Venkatesh, Deepr: a convolutional net for medical records, IEEE J. Biomed. Health Inform. 21 (1) (2016) 22–30.
[29] E. Choi, M.T. Bahadori, L. Song, W.F. Stewart, J. Sun, GRAM: graph-based attention model for healthcare representation learning, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 787–795.
[30] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, J. Gao, Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1903–1911.
[31] T. Bai, S. Zhang, B.L. Egleston, S. Vucetic, Interpretable representation learning for healthcare via capturing disease progression through time, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 43–51.
[32] C. Mao, L. Yao, Y. Luo, MedGCN: Medication recommendation and lab test imputation via graph convolutional networks, J. Biomed. Inform. 127 (2022) 104000.
[33] H. Li, Z. Wang, C. Lan, P. Wu, N. Zeng, A novel dynamic multiobjective optimization algorithm with non-inductive transfer learning based on multi-strategy adaptive selection, IEEE Trans. Neural Netw. Learn. Syst. (2023).
[34] S. Adhikari, S. Thapa, U. Naseem, H.Y. Lu, G. Bharathy, M. Prasad, Explainable hybrid word representations for sentiment analysis of financial news, Neural Netw. 164 (2023) 115–123.
[35] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).
[36] T. Wongvorachan, S. He, O. Bulut, A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining, Information 14 (1) (2023) 54.
[37] R. Mohammed, J. Rawashdeh, M. Abdullah, Machine learning with oversampling and undersampling techniques: overview study and experimental results, in: 2020 11th International Conference on Information and Communication Systems, ICICS, IEEE, 2020, pp. 243–248.
[38] B. Abu-Salih, Domain-specific knowledge graphs: A survey, J. Netw. Comput. Appl. 185 (2021) 103076.
[39] A. Zafar, S.K. Sahoo, H. Bhardawaj, A. Das, A. Ekbal, KI-MAG: A knowledge-infused abstractive question answering system in medical domain, Neurocomputing 571 (2024) 127141.
[40] P. Wu, Z. Wang, H. Li, N. Zeng, KD-PAR: A knowledge distillation-based pedestrian attribute recognition model with multi-label mixed feature learning network, Expert Syst. Appl. 237 (2024) 121305.
[41] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. 34 (12) (2021) 5586–5609.
[42] S. Lee, Y. Son, Multitask learning with single gradient step update for task balancing, Neurocomputing 467 (2022) 442–453.
[43] A. Mohammed, R. Kora, A comprehensive review on ensemble deep learning: Opportunities and challenges, J. King Saud Univ.-Comput. Inf. Sci. 35 (2) (2023) 757–774.
[44] X. Wu, C. Wen, Z. Wang, W. Liu, J. Yang, A novel ensemble-learning-based convolution neural network for handling imbalanced data, Cogn. Comput. 16 (1) (2024) 177–190.

Usman Naseem is a Lecturer with the School of Computing, Macquarie University, Australia. Prior to persuading his research, he worked in leading ICT companies for over ten years. His research interests include natural language processing and its applications incldluing healthcare. He publishes and serves with the Program Committee, including the Area Chair for several top-tier venues, including ACL, EMNLP, NAACL, COLING, WSDM, Webconf, and AAAi. He also delivered several invited talks and a tutorial on recommender systems at Webconf 2023. His work in NLP has attracted attention from the World Health Organization and earned him the Nepean Blue Mountains Local Health District (NBMLHD) Board Chair's Quality Award, in 2021. He also received the prestigious IEEE Best Transactions Paper Award, in 2022.

Surendrabikram Thapa is a Research Faculty with Virginia Tech, where he works primarily on deep learning. He was a Visiting Scholar at the University of Technology Sydney, Australia, in 2020. During graduate (M.S.) study at Virginia Tech, he was funded by the National Science Foundation (NSF) Grant and various government agencies. He has published research papers at various reputed conferences and journals. His research interests include natural language processing, computational social sciences and computer vision applications. Currently, his research is funded by various local and federal government agencies in the USA along with several industry partners. He has served as a program committee member and organizer for numerous conferences and workshops. He has been serving as a reviewer for several reputed journals and conferences.

Qi Zhang received his first Ph.D. from the Beijing Institute of Technology, Beijing, China in 2020, and his second Ph.D. from the University of Technology Sydney, Sydney, NSW, Australia, in 2023. He is currently a Research Fellow at Tongji University, Shanghai, China. He has authored more than 30 high-quality papers in premier conferences and journals, including NeurIPS, AAAI, IJCAI, WWW, SIGIR, ICDM, ECAI, DSAA, TKDE, TOIS, TNNLS, ESWA, and Pattern Recognition et al. His primary research interests include multimodal learning, time series analysis, frequency neural networks in various tasks such as recommender systems, fake news detection, mental health analysis, and neuroscience analysis.

Shoujin Wang received the PhD degree in Data Science from the University of Technology Sydney (UTS), Sydney, NSW, Australia, in 2019. He is currently a Lecturer

in Data Science with UTS. He has authored high-quality papers in premier conferences and journals, including International World Wide Web Conference (TheWebConf), AAAI Conference on Artificial Intelligence (AAAI), International Joint Conferences on Artificial Intelligence (IJCAI), and the ACM Computing Surveys (ACMCSUR). His research interests include data mining, machine learning, recommender systems, and fake news mitigation. He is a recipient of some prestigious awards, including the 2022 DSAA Next-generation Data Scientist Award and the 2022 Club Melbourne Fellowship Award.

Junaid Rashid received the B.S. and M.S. degrees in computer science from the COMSATS Institute of Information Technology, Wah Campus, Pakistan, in 2014 and 2016, respectively, and the Ph.D. degree in computer science from the University of Engineering and Technology, Taxila, Pakistan, in 2020. Since November 2020, he has been an Honorary Senior Postdoctoral Research Fellow with the Center for AI and Data Science, Edinburgh Napier University, U.K., and received the Fellowship, in 2022. He worked as a Research Professor and Postdoctoral Researcher at Kongju National University, Korea. He is currently working as an Assistant Professor at Sejong University, Korea. He has published research papers in prestigious journals and conferences. His research interests include data science, machine learning, natural language processing, topic modeling, text mining, information retrieval, big data, pattern recognition, fuzzy systems, medical informatics, biomedical text analytics, and software engineering. He received the fullyfunded scholarship for the Ph.D. degree. He has served/been serving as a reviewer for various reputed journals and conferences. He has been an Academic Editor of PLOS One.

Liang Hu is a professor at Tongji University. He is also the chief AI scientist with DeepBlue Academic of Sciences. His research interests include recommender systems, data mining, machine learning, representation learning and general artificial intelligence. He has published a number of papers in toprank international conferences and journals in the area of recommender systems, including WWW, IJCAI, AAAI, ICDM, ICWS, TOIS, JWSR. He has delivered several tutorials in top-rank conferences, including IJCAI, AAAI, ICDM and PAKDD.

Amir Hussain received the B.Eng. (Hons.) and Ph.D. degrees from the University of Strathclyde, Glasgow, U.K., in 1992 and 1997, respectively. He is the Founding Director of the Centre of AI and Robotics, Edinburgh Napier University, Edinburgh, U.K. He has led major national and international projects and supervised over 40 Ph.D. students. He has authored several patents and over 600 publications, including around 300 journal articles and 20 books/monographs. His research interests are cross-disciplinary and industry-led and aimed at developing responsible artificial intelligence (AI) and cognitive data science technologies to engineer the smart healthcare and industrial systems of tomorrow. Dr. Hussain is the Founding Chief Editor of Springer's Cognitive Computation journal and an invited editor for various other journals. Among other distinguished roles, he has served as the General Chair for the flagship 2020 IEEE World Congress on Computational Intelligence (WCCI) and the 2023 IEEE Smart World Congress (SWC). He is the Chair of the IEEE U.K. and Ireland Chapter of the IEEE Industry Applications Society.