An Assistant Diagnosis Method Based on Association Paths Representation Learning for Diseases

Ying Zhong*, Xiaojun Luo

***Abstract*—The introduction of Electronic Health Records (EHRs) has created a new set of challenges and opportunities for clinical research, stimulating the transition of medical health-related studies to a data-driven approach and opening up new prospects for more personalised health care. In light of the intricate correlation between medical data, this paper presents a learning approach to represent medical information connection pathways. First, the method uses extracted medical pathways to enhance potential associations between diseases and learns highly representative patient embedding vectors. Second, by combining the medical information association path, the representative medical concept representation vector is learned. Meanwhile, the self-attention reasoning network learns the patient feature embedding vector using the patient's previous admission records in order to properly anticipate the patient's future health state. Finally, the experiment demonstrates that the proposed method can effectively use the medical path and the patient's historical disease record for disease prediction, and achieve excellent prediction results.**

***Index Terms*—Deep Learning, Electronic Health Records, Healthcare Informatics, Association Path**

I. Introduction

The increased use of electronic health records [1] has encouraged the use of medical and health data to drive research, thus providing new possibilities for more personalized health care. Intelligent assisted diagnosis is a key field of research that focus on predicting illnesses from medical and health records [2], [3] information. The goal is to analyze and anticipate individuals' well-being based on their individual data and historical digital medical records, to assist medical organizations with diagnosis, to increase efficiency and quality of service, and to reduce the cost of medical treatment for patients.

How to model continuous and high-dimensional EHR data and understand the prediction findings is a significant issue in this undertaking. Because observed medical record data is frequently complicated and non-linear, machine learning can be used to solve these constraints. Traditional supervised learning approaches (such as logistic regression, random forest, and naive Bayes) are inadequate for modeling longitudinal processes because they cannot account for relationships between outcomes or characteristics.

Manuscript received March 21, 2023; revised September 11, 2023.

This work was supported in part by Shenzhen Fundamental Research Program under Grant No. 20210317191843003 and Shaanxi Provincial Key R&D Program under Grant No. 2021ZDLGY05-01.

Ying Zhong is a lecturer of Research and Development Institute of Northwestern Polytechnical University, Shenzhen, Guangdong 518057 China (e-mail: yxy_0713@qq.com).

Xiaojun Luo is a postgraduate student of Research and Development Institute of Northwestern Polytechnical University, Shenzhen, Guangdong 518057 China. (e-mail:luo_xj@mail.nwpu.edu.cn)

Recently, some researchers use patient representation learning for clinical assist diagnosis, such as mortality prediction [1], [4], patient subtypes [5], length-of-stay prediction [6], [7], 24-hour decompensation [8], [9] and multi-task learning [1].

However, current prediction methods generally rely on the temporal characteristics of patients receiving medical treatments [10], and these methods rarely take into account some potential associations in medical information, such as the relationship between diseases and other complications, and the association between diseases and symptoms. The associations embedded in medical data are critical for medical data analysis research and diagnosis, such as searching for relevant disease information from a massive amount of data with key symptoms, or predicting the future health of patients by exploring the correlation between diseases in previous diagnosis. Recursive neural network (RNN) [11], [12] is a type of neural network that is created for sequential data, and can adapt well to high-dimensional feature data. It ia commonly used in longitudinal data prediction. However, However, in order to obtain steady performance, the neural network normally requires a huge amount of training labels, which may be highly expensive. RNN models, whether displaying patients' physiological conditions in a continuous or discrete manner, are unable of identifying changes in patients' physiological conditions. Our contributions are listed as follows.

- We propose a disease-assisted diagnosis model based on medical information association path representation learning, called MiaPRL. The model can assist doctors in making diagnostic decisions and reduce patient medical costs.
- We construct a medical knowledge network with a tree-like hierarchical structure using the classification information of medical concepts, and extract medical information association paths from the medical knowledge network, which are used to learn the implicit association relationships between different diseases.
- Finally, we conduct a large number of experiments. Compared with the state-of-art models, our designed model can produce more accurate results.

II. Related Work

Medical researchers are trying to predict the health of patients in the future by analysing their past electronic medical records [2], [3]. A key problem in these activities is finding a way to interpret the predictions generated from modelling large, multi-dimensional medical record data. Since the observed medical record data is typically complex

and non-linear, these limitations can be overcome by using machine learning to predict the timing of events. Traditional supervised learning approaches, such as logistic regression [13], random forest [14], and naive bayes [15], are inadequate for modelling longitudinal processes because they do not account for intertemporal linkages in either outcomes or characteristics. Furthermore, with the exception of retention, these models are not easily interpretable.

Deep learning algorithms are used to perform related tasks that involve learning patient representations, such as predicting length-of-stay [6], [7], predicting mortality [1], [4], [16], phenotyping patients and categorising international classification of diseases (ICD) codes [17]. In this paradigm, deep learning models are used to represent a patient's physical state, and a specific network is employed to provide specific predictions or classifications. The performance of the deep learning model for patient representation depends on the construction of the model using the EHRs.

Due to the temporal nature of EHR data, recurrent neural networks (RNNs) and long short-term memory (LSTM) are widely used. RETIN [18], [19] proposes an interpretable reverse temporal attention model. Dipole [20] combined with bidirectional RNNs to mine EHR data. Camp [21] uses demographic information from the Common Concern model to make diagnostic predictions. StageNet [22] integrates time intervals between visits into LSTM to simulate the health status of patients at different stages. INPREM [23] applies Bayesian neural networks to attention mechanisms to improve the interpretability of the model. HiTANet [24] proposes a hierarchical time-aware attention network for health risk prediction. LSAN combines long-term and short-term information in the EHR for prediction.

In addition, numerous investigations have applied various techniques to the basic RNN architecture to address the lack of density, extensive dimensionality, and diversity of clinical time series information. Lei et al. [25] applied recurrent neural network and auto-encoder to encode patient hospital records as low-dimensional dense vectors, taking into account the temporal variation of EHR, the non-negative tensor factorization models the input sequence as a time tensor as the input of LSTM [26]. In addition to RNN, convolutional neural network (CNN) is also used for learning patient represents. Unlike RNN, CNN cannot process variable length sequence data, so sequences must be pre-processed before being input to CNN. In the literature [3], patient visit sequence data is used as input to the CNN after being filled to a fixed length. Recent works [25], [27] also consider multi-source data to provide some prior knowledge for more accurate prediction. The above studies attempt to improve basic RNNs to handle time intervals in sequences. However, none of them has been designed to measure changes in physiological states.

In this paper, we use medical concepts to classify information, construct a medical knowledge network with a tree hierarchy, and extract the medical information association path from the medical knowledge network to learn the hidden association between different diseases. In addition, we design a path encoder using the self-attention mechanism to extract medical paths for different diseases in each admission record. By merging the medical paths, the hierarchical information of diseases and medical concepts can be effectively captured, and the medical concept representation vectors can be learned. Meanwhile, the self-attention reasoning network can learn the patient's characteristic embedding vector from the historical admission record, and accurately predict the patient's future health risk status.

III. Association paths extraction in medical information

This section mainly introduces an association path extraction method in the heterogeneous medical information network to learn the relationship between diseases and support disease prediction. There are a variety of international standards for classification of medical concepts, which classify medical information such as diseases, symptoms and abnormalities. The hierarchical organisation of classifications such as ICD codes and CCS classification are often used to improve medical concepts.

We denote the medical codes that appear in electronic medical records as $c_1, c_2, ..., c_{|\mathcal{G}|} \in \mathcal{G}$, where $\mathcal{G}$ is a set of medical codes. $\mid \mathcal{G} \mid$ represents the number of medical codes in the dataset. Medical codes only contain the disease codes that appear in the dataset. $\mathcal{P}$ represents all patients in the electronic medical record dataset. For a patient $u \in \mathcal{P}$ in the medical record, its medical record can be regarded as a sequence, which is recorded as $T^{(u)} = \left[V_1^{(u)}, V_2^{(u)}, ..., V_{|T^{(u)}|}^{(u)}\right]$. $\mid T^{(u)} \mid$ is the number of medical records recorded by the patient *u*. Each visit of the patient V_t corresponds to an muti-hot vector $\mid \mathcal{G} \mid$ with dimension $\boldsymbol{x}_t \in \{0,1\}^{|\mathcal{G}|}$. If the value of the *i*-th element is 1, it means that the patient has been diagnosed with the disease c_i. Meanwhile, We use the information related to the classification criteria of medical concepts to construct a medical knowledge network $\mathcal{K}_D$ with tree hierarchy. The medical knowledge network contains categorical information for medical concepts. The leaf nodes consist of medical codes in $\mathcal{G}$. The set of non-leaf nodes $\mathcal{K}_D$ is recorded as $\mathcal{G}'$, and $\mathcal{G}\prime = \left\{c_{|\mathcal{G}|+1}, c_{|\mathcal{G}|+2}, ..., c_{|\mathcal{G}|+|\mathcal{G}'|}\right\}$ represents coarse-grained classification information. $\mathcal{K}_D$ is a directed acyclic graph with the set of nodes $D = \mathcal{G} + \mathcal{G}'$.

Inspired by the meta-path [28], a path ρ can be expressed as $\mathcal{A}_1 \xrightarrow{\mathcal{R}_1} \mathcal{A}_2 \xrightarrow{\mathcal{R}_2} ... \xrightarrow{\mathcal{R}_l} \mathcal{A}_{l+1}$, which is used to represent the relationship $\mathcal{R}$ between objects $\mathcal{A}$. All medical codes associated with this path, including those representing diseases and categories, form a particular medical path. In this paper, the model selects the medical information association path with a high number of co-occurrences of the starting disease and the ending disease for learning.

Fig. 1 shows the process of medical pathway extraction. The left side of the figure is an example of the medical knowledge network $\mathcal{K}_D$, where solid circles represent medical codes for different diseases (C1, C2, C3, C5), and dotted circles represent coarse-grained disease classifications (C4, C6, C7, C8, C9, C10, ROOT). The right side of the figure shows three randomly selected medical information association paths, starting from disease nodes C1, C2 and C5, and randomly walking in the network $\mathcal{K}_D$ to reach the nodes C5, C3 and C3 respectively.

For a patient admission record V_t, a random sample of diseases in the record is firstly carried out, and medical information association paths are constructed for the sampled

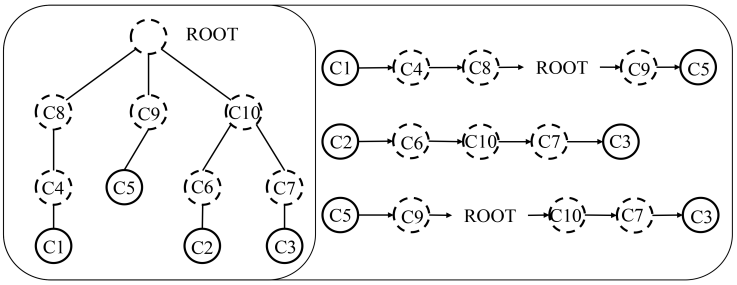

Fig. 1: An example of the medical knowledge network $\mathcal{K}_D$ (The left figure represents the disease classification, and the right figure represents the medical path).

diseases. Each disease can construct multiple medical paths. The collection of medical information association paths sampled and constructed for diseases in the admission record V_t is denoted as $\rho_t^{(set)}$.

IV. Association paths representation learning

This section mainly introduces the medical information association path representation learning method, including medical code pre-training, the self-attention encoder module, medical information association path representation learning module, and patient history medical information representation learning module. Based on these modules, we propose an association path representation learning approach for assisted diagnosis of diseases. Fig. 2 shows the disease prediction model framework based on medical information association paths representation learning.

A. Pre-training of medical codes

To enhance the expressiveness ability of medical concepts, pre-training processing of medical codes is required. The medical code representation learning method GRAM [29], which relies on the attention mechanism, first initialises the medical code embedding e_i with the co-occurrence information of the medical code as the initial feature of the nodes in the medical knowledge network $\mathcal{K}_D$. The representation learning method based on hyperbolic geometry space is [30] suitable for the representation learning of hierarchical information, and can significantly improve the generalisation performance, which is suitable for the medical knowledge network with tree hierarchy. Therefore, we use hyperbolic graph neural network to perform unsupervised training on medical knowledge network $\mathcal{K}_D$.

Firstly, the medical code embedding vector e_i initialized with co-occurrence information is used as the input node feature. The hyperbolic neural network is used to train the medical knowledge network $\mathcal{K}_D$ unsupervised, and a representative feature embedding vector is generated for medical codes. The feature embedding vector m_i learned from the *i*-th medical code is recorded as m_i. By merging the vector representation $\boldsymbol{m_1}, \boldsymbol{m_2}, ..., \boldsymbol{m}_{|\mathcal{G}|+|\mathcal{G}'|}$ of all medical codes, a medical code embedding matrix $\boldsymbol{M} \in \mathbb{R}^{d \times \left(|\mathcal{G}|+|\mathcal{G}'|\right)}$ can be generated, where d is the vector dimension size and m_i is the *i*-th column of $\mathbf{M}$.

For a patient medical record V_t, multiply its corresponding multi-hot vector x_t by the medical code embedding matrix $\mathbf{M}$ to obtain the representation vector $\boldsymbol{v_t} \in \mathbb{R}^d$ of the patient current medical record. The calculation process is as follows.

$$\boldsymbol{v_t} = \tan h\left(\boldsymbol{M x_t}\right) \tag{1}$$

B. Self-attention encode module

The attention mechanism enables the model to capture relationships between elements in the sequence, regardless of their relative position. By employing the self-attention encoding module as a basis, we are able to construct multiple self-attention encoding modules to encode the medical information association path and to learn the representation of the patient's medical history.

The self-attention encoding module consists of two parts, which are the multiple self-attention encoding layer and the position feedforward network. The use of multiple self-attention encoding as opposed to a single attention layer allows the model to focus on information from different positions in different quantum spaces simultaneously. The position feedforward network performs a non-linear transformation for each position element in the input sequence. The multiple self-attention encoding layer is defined as follows.

$$MultiHead\left(\boldsymbol{X}^l\right) = \left[head_1, head_2, ...\right]\boldsymbol{W}^o \tag{2}$$

$$head_i = Attention\left(\boldsymbol{X}^l\boldsymbol{W}_i^Q, \boldsymbol{X}^l\boldsymbol{W}_i^K, \boldsymbol{X}^l\boldsymbol{W}_i^V\right) \tag{3}$$

where X^l is the input matrix of the multiple self-attention encode layer of the *l*-th layer, and $\mathbf{X}^0$ is the initial input of the first layer. $\boldsymbol{W}^O \in \mathbb{R}^{d \times d}$, $\boldsymbol{W}_i^Q \in \mathbb{R}^{d \times \frac{d}{h}}$, $\boldsymbol{W}_i^K \in \mathbb{R}^{d \times \frac{d}{h}}$, $\boldsymbol{W}_i^V \in \mathbb{R}^{d \times \frac{d}{h}}$ is the trainable parameter matrix in each attention encode layer. The self-attention operation is defined as follows.

$$Attention\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{\frac{d}{h}}}\right)\boldsymbol{V} \tag{4}$$

where d refers to the dimension of the input vector, and h represents the number of multiple self-attention encode layers.

For each position vector in the input matrix, the position feedforward network independently performs the same transformation. A feedforward network with a ReLU activation function consists of two connected linear transformations. The specific definition is as follows.

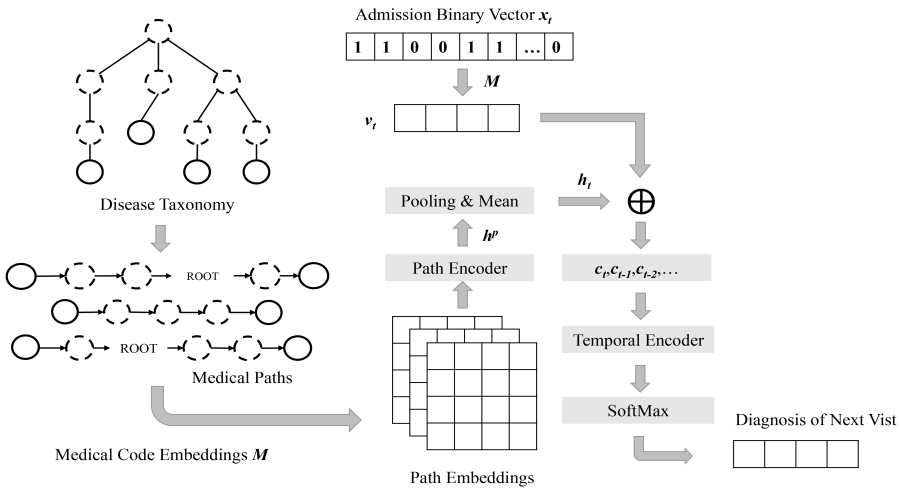

Fig. 2: An framework of disease prediction model based on medical information association paths representation learning

$$FFN(\boldsymbol{z}) = (ReLU(\boldsymbol{z}\boldsymbol{W_1} + \boldsymbol{b_1}))\boldsymbol{W_2} + \boldsymbol{b_2} \quad (5)$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d \times d_a}$, $\boldsymbol{b}_1 \in \mathbb{R}^{d_a}$, $\boldsymbol{W}_2 \in \mathbb{R}^{d \times d_a}$, $\boldsymbol{b}_2 \in \mathbb{R}^{d \times \frac{d}{h}}$ is the trainable parameter matrix of the position feedforward network, and d_a is the transition vector dimension of the matrix transformation in the position feedforward network. Note that a residual operation and normalization step are appended to each component in the self-attention encode module, which consists of the attention encode layer and the position feedforward network.

C. Medical information association path representation learning

A variable length sequence is manipulated by combining multiple self-attention encode modules in order to incorporate the medical code sequence into a low dimensional vector space. Given a medical path set $\rho_t^{(set)}$, for one medical path instance ρ, the feature embedding matrix of its corresponding medical path instance is obtained by merging the feature embedding vectors of the medical code in the medical path instances, which is recorded as $\boldsymbol{X}^{\rho} \in \mathbb{R}^{L \times d}$. The feature embedding vector dimension of each medical code is denoted by d, and L represents the number of medical codes in medical path instances. We introduce a trainable position matrix $\boldsymbol{P}^{(path)} \in \mathbb{R}^{L \times d}$ to enhance the position information of the medical code in the path. Finally, the feature embedding matrix $\boldsymbol{X'}^{\rho}$ can be obtained by adding the two matrices. For a path instance ρ, the encoding process can be summarized as below,

$$\boldsymbol{H}^{\rho} = pathEncoder(\boldsymbol{X'}^{\rho}) \quad (6)$$

where H^{ρ} is the embedding matrix in the medical path instance ρ, each line represents the embedding vector after the medical code is encoded at the corresponding position in the path. The feature embedding vector encoded by the medical information association path instance ρ is recorded as $\boldsymbol{h}_L^{\rho}$.

Finally, the above encoding process is performed for each path instance in the medical information association path set $\rho_t^{(set)}$, and the corresponding feature embedding vector is obtained. By applying the average pooling operation, the vector representation of the medical information association path set corresponding to the *t*-th admission can be obtained.

$$\boldsymbol{h}_t = \frac{1}{|\ \rho_t^{(set)}\ |} \sum_{\rho \in \rho_t^{(set)}} \boldsymbol{h}_{L_\rho}^{\rho} \quad (7)$$

D. Historical medical information representation learning of the patient

The feature vector of the patient medical visit is created by combining the feature embedding vector of the patient medical record and the corresponding medical information association path feature embedding vector.

This method is adopted to learn and encode the patient historical medical records, so as to obtain the embedded feature vector sequence of the medical records is obtained, which is denoted as $[\boldsymbol{c}_1; \boldsymbol{c}_2; ...; \boldsymbol{c}_t]$. The representation learning module of patient history medical information uses self-attention encode module to represent and learn patient history medical treatment information according to embedded characteristics of patient history medical treatment records. The feature vector encoded by the last element in the sequence is taken as the representation vector of the current patient. We adopt a trainable position matrix $\boldsymbol{P}^{(adm)} \in \mathbb{R}^{t \times d}$ to reinforce the time sequence of each patient admission. It is calculated as follows.

$$\boldsymbol{c}_i = \boldsymbol{v}_i + \boldsymbol{h}_i, i \in [1, t] \quad (8)$$

$$[\boldsymbol{c}'_1; \boldsymbol{c}'_2; ...; \boldsymbol{c}'_t]^T = [\boldsymbol{c}_1; \boldsymbol{c}_2; ...; \boldsymbol{c}_t]^T + \boldsymbol{P}^{(adm)} \quad (9)$$

$$[\boldsymbol{o}_1; \boldsymbol{o}_2; ...; \boldsymbol{o}_t] = TemporalEncoder([\boldsymbol{c}_1; \boldsymbol{c}_2; ...; \boldsymbol{c}_t]) \quad (10)$$

The temporal information encoder *TemporalEncoder* is implemented through the self-attention encode module. $\mathbf{c}_i$ is the feature vector of the patient *i*-th medical record, and $\mathbf{o}_t$ is the final feature representation vector of the patient.

V. Algorithm description

The learned patient feature representation vector is used to predict the health status of patients in a certain period of time in the future. Given the final feature representation vector $\boldsymbol{o}_t$ of a patient. It is fed input into a full connection layer for linear transformation, through the sigmoid activation function, the model obtains the probability distribution of the patient's next illness $\hat{\boldsymbol{y}}_{t+1}$. $\boldsymbol{y}_{t+1}$ is a multi-dimensional multi-hot vector used to represent the patient's condition at the time of the *t*+1 medical treatment. If the value of the *i*-th element in $\boldsymbol{y}_{t+1}$ is 1, it means that the patient actually suffers from the disease referred to by the *i*-th element. $\hat{\boldsymbol{y}}_{t+1}$ is an approximation of $\boldsymbol{y}_{t+1}$ generated by the model, representing the prediction result of the patient's illness. Each value in the vector $\hat{\boldsymbol{y}}_{t+1}$ indicates that the patient may suffer from the *i*-th disease. It is calculated as follows.

$$\hat{\boldsymbol{y}}_{t+1} = \sigma\left(\boldsymbol{W}_p \boldsymbol{o}_t + \boldsymbol{b}_p\right) \tag{11}$$

where $\boldsymbol{W}_\rho \in \mathbb{R}^{|\mathcal{G}| \times d}$ and $\boldsymbol{b}_p \in \mathbb{R}^{|\mathcal{G}|}$ are trainable model parameters. σ is the sigmoid activation function. Calculate the binary cross entropy loss function to optimize the model. The loss function is calculated as follows.

$$\mathcal{L} = - \frac{1}{\mid T^{(u)} \mid -1} \sum_{|T^{(u)}|-1}^{t=1} \left(\boldsymbol{y}_t^T \log\left(\hat{\boldsymbol{y}}_t\right) + \left(1 - \boldsymbol{y}_t^T\right) \log\left(1 - \hat{\boldsymbol{y}}_t\right)\right) \tag{12}$$

Algorithm 1 describes an auxiliary diagnosis method for diseases based on medical information association path representation learning. The model is optimized using the Adam optimizer [31].

Algorithm 1 A disease-assistant diagnosis method based on medical information association path representation learning

Input: Medical Knowledge Network $\mathcal{K}_D$; Medical code embedding matrix $\boldsymbol{M} \in \mathbb{R}^{d \times \left(|\mathcal{G}| + |\mathcal{G}'|\right)}$; Patient admission records $T^{(u)} = \left[V_1^{(u)}, V_2^{(u)}, ...V_{|T^{(u)}|}^{(u)}\right]$; Vector representation of patient admission records $\left[\boldsymbol{x}_1^{(u)}, \boldsymbol{x}_2^{(u)}, ..., \boldsymbol{x}_{|T^{(u)}|}^{(u)}\right]$

Output: Results of diseases-auxiliary diagnosis

1: **for** *i*=1...| $T^{(u)}$ | **do**
2: Sampling and constructing medical path sets $\rho_i^{(set)}$ for diseases in V_i
3: $\boldsymbol{v}_t = \tan h\left(\boldsymbol{M}\boldsymbol{x}_t\right)$
4: The representation $\boldsymbol{h}_i$ of medical path set $\rho_i^{(set)}$ is calculated by Formula (7).
5: Calculate the current patient representation vector $\boldsymbol{o}_i$ by Formula (10)
6: $\hat{\boldsymbol{y}}_{t+1} = \sigma\left(\boldsymbol{W}_p \boldsymbol{o}_t + \boldsymbol{b}_p\right)$
7: Calculate the loss function $\mathcal{L}$ by Formula (12)
8: **end for**
9: Optimize loss function $\mathcal{L}$ using Adam optimizer
10: Convergence of the model
11: **return** Prediction results

VI. Experiment

In this section, we first describe the dataset used. Then we list the available state-of-the-art algorithms and compare them with our model. Finally, the ablation experiment and the parameter analysis experiment performed to prove the effectiveness and superiority of the proposed model.

A. datasets

1) MIMIC datasets: We conducted experiments to validate the medical path representation learning-based diagnosis of diseases by utilizing two public datasets, namely MIMIC-III [32] and MIMIC-IV [33]. MIMIC-III is an medical information database of intensive care, which contains inpatient records in the intensive care unit of Beth Israel Deaconess Medical Center from 2001 to 2012, with more than 40,000 patient-related electronic medical record data.

The main statistics of the above two datasets are shown in TABLE I. For the MIMIC-III dataset, the average number of patients admitted to hospital is small. We use the MIMIC-III dataset to evaluate the performance of the model when the training data is insufficient. However, the admission records in the MIMIC-IV dataset have a wide time range, To assess the model's performance in handling long series data, we specifically choose patients with more than 10 admissions.

The first step is to extract patient diagnosis records from the MIMIC dataset. For the MIMIC-III dataset, the fields and their meanings are shown in TABLE II. "SUBJECT_ID" and "HADM_ID" can uniquely determine the patient ID and the medical ID. The "ICD9_CODE" field records the ICD9 code of the patient's disease. We traverse the diagnostic records in the MIMIC-III dataset, using "SUBJECT_ID" as a keyword to build a dictionary containing multiple lists, each of which stores a record of a patient's medical record in the form of an ICD-encoded list. When the traversal is complete, the dictionary is stored for quick access by the model.

For the MIMIC-IV dataset, the fields and their meanings are shown in TABLE III. The MIMIC-IV dataset removes the field "ROW_ID" from the MIMIC-III dataset, which has no practical meaning. As the version of the MIMIC-IV dataset is relatively new, the field "ICD_VERSION" has been added to distinguish the version of the ICD code. The value is 9 or 10. Traverse the diagnostic records in the MIMIC-IV dataset and build a dictionary with "SUBJECT_ID" as a keyword. Due to the inconsistent version of the ICD code in MIMIC-IV, an additional open source tool is required to map the ICD-10 version to the ICD-9 code. As the ICD-10 code is richer and more complex than the ICD-9 code, some information is missing in the mapping process.

2) CCS datasets: We use the CCS multi-level diagnostic hierarchy to build a medical knowledge network with external medical expertise. ICD-9-CM coded clinical information classification software is a classification scheme for medical diagnosis and treatment methods, which can be used for data mining and analysis. The medical knowledge network is constructed by using the CCS multi-level diagnosis hierarchy.

In the CCS classification data, each row represents the CCS classification corresponding to an ICD code, and the category to which each ICD code belongs is marked according to the hierarchical results, which can reach up to 4 levels of classification. By traversing the CCS classification data

TABLE I: Dataset information

Dataset	MIMIC-III	MIMIC-IV
Number of patients	7,499	73,181
Admissions	19,911	294,235
Average number of patients admitted	2.66	4.02
Number of ICD-9 codes	5,549	6,820
Number of ICD-9l categories	939	982

TABLE II: MIMIC-III Dataset fields and their meanings

fields	meaning
ROW_ID	Line Number
SUBJECT_ID	Patient ID
HADM_ID	Admission Record ID
SEQ_NUM	Serial Number
ICD9_CODE	Disease Code

TABLE III: MIMIC-IV Dataset fields and their meanings

fields	meaning
ICD_VERSION	Disease Code Version (ICD9 or ICD10)
SUBJECT_ID	Patient ID
HADM_ID	Admission Record ID
SEQ_NUM	Serial Number
ICD_CODE	Disease Code

row by row, the category information of each ICD code is recorded, and each category information and the predecessor and successor categories of the ICD code are simultaneously recorded, which are used to construct a tree-like medical knowledge network.

B. Experimental setup

For the proposed disease-assisted diagnosis method based on medical information association path representation learning, it is recorded as ProAID in the experiment. We set the number of self-attention encoder modules to 2 and the number of multiple self-attention layers to 4. The hidden vector dimension is set to 128 for both the MIMIC-III and MIMIC-IV datasets. The model learning rate is 0.0006 for MIMIC-III and 0.0002 for MIMIC-IV. For each disease, 200 co-occurring diseases were selected as candidates, and 10 related diseases were randomly selected each time as the destination to construct the medical path. 50 medical pathways associated with each visit of the patient were selected on the MIMIC-III dataset and 20 paths are used on the MIMIC-IV dataset. The dataset is randomly divided into training, validation, and test sets in a ratio of 7:1:2. The output of the model is calculated by taking the K diseases with the highest probability as the prediction results, and calculating the precision and recall values. The values of K are 1,5,10,20.

C. Baseline Models

For the purpose of evaluating the overall accuracy of diagnostic prediction, we apply our technique to the eight baselines (LSTM [34], StageNet [22], RETAIN [35], Dipole [20], Concare [36], GRU [37], GRAM [29], CAMP [21], CHARACTER-BERT [38], LuPIET [35] and GRU-TV [37]). All models are applied to our task, which requires only historical diagnoses and treatments, with adaptation for a fair comparison. Performances may differ from given in the original articles, as side information such as ontology and temporal intervals are not merged.

D. Performances comparison

The experimental results are shown in TABLE IV. It can be seen from the results that the prediction results of the model proposed in this paper outperform all the comparison methods, which verifies that MiaPRL can effectively capture the time information in the patient's medical record and effectively learn the patient's embedding vector. All models performed better on the MIMIC-IV dataset than on the MIMIC-III dataset because the experiments on the MIMIC-III dataset used patient information with more than 10 admissions.

The GRU model exhibited comparable prediction performance to the MiaPRL method proposed in this paper when K=2, 5, and 10 on the two datasets. The main reason is that GRU also used co-occurrence information to pre-train medical codes, and used attention mechanisms to learn the association between medical codes in medical knowledge networks. However, GRU only focuses on the inclusion relationship of connected nodes, or omits the connection between medical codes at a greater distance. MiaPRL uses hyperbolic graph neural networks to capture medical knowledge networks and make better predictions. When constructing the medical pathway, GRAM selects only the most common diseases as starting and ending points, but ignores some less common but more critical concurrent diseases, resulting in missing some important potential information.

In contrast, the CAMP model did not achieve the expected performance and is better suited to training and prediction using the 3-bit ICD-9 class code than the sparse 5-bit ICD-9 coding.

E. Ablation experiment

To verify the effectiveness of each component in the model, we performed ablation experiments on two datasets. First, the pre-training process of the medical code and the learning module of the medical path representation are removed to evaluate the contribution of these components to the prediction task, which are recorded as MiaPRL$_a$ and

TABLE IV: Performances comparison on predictions of MIMIC-III and MIMIC-IV datasets

Dataset	Model	K=2		K=5		K=10		K=20	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
MIMIC-III	LSTM	0.615	0.105	0.492	0.246	0.387	0.397	0.315	0.547
	StageNet	0.621	0.116	0.523	0.257	0.394	0.418	0.327	0.618
	RETAIN	0.574	0.114	0.469	0.189	0.379	0.324	0.276	0.461
	Dipole	0.647	0.128	0.508	0.239	0.385	0.328	0.274	0.439
	Concare	0.681	0.129	0.564	0.267	0.426	0.419	0.324	0.561
	GRU	0.589	0.121	0.487	0.218	0.385	0.332	0.284	0.473
	GRAM	0.651	0.124	0.537	0.241	0.435	0.371	0.297	0.524
	CAMP	0.612	0.121	0.481	0.215	0.371	0.312	0.264	0.425
	GRU-TV	0.608	0.126	0.495	0.226	0.396	0.335	0.289	0.486
	CHARACTER-BERT	0.628	0.122	0.543	0.263	0.416	0.428	0.337	0.581
	LuPIET	0.621	0.119	0.537	0.259	0.407	0.419	0.328	0.571
	MiaPRL	**0.692**	**0.131**	**0.571**	**0.276**	**0.456**	**0.437**	**0.359**	**0.607**
MIMIC-IV	LSTM	0.727	0.129	0.617	0.246	0.514	0.473	0.371	0.485
	StageNet	0.731	0.134	0.635	0.253	0.527	0.489	0.378	0.514
	RETAIN	0.719	0.124	0.607	0.258	0.476	0.403	0.368	0.569
	Dipole	0.726	0.132	0.634	0.258	0.517	0.382	0.358	0.526
	Concare	0.758	0.132	0.618	0.243	0.519	0.427	0.359	0.508
	GRU	0.724	0.131	0.624	0.265	0.508	0.415	0.371	0.576
	GRAM	0.735	0.132	0.651	0.274	0.539	0.417	0.385	0.602
	CAMP	0.712	0.124	0.624	0.249	0.493	0.374	0.342	0.513
	GRU-TV	0.731	0.138	0.629	0.273	0.536	0.438	0.374	0.591
	CHARACTER-BERT	0.756	0.129	0.648	0.265	0.537	0.458	0.367	0.598
	LuPIET	0.742	0.124	0.623	0.261	0.529	0.447	0.361	0.587
	MiaPRL	**0.785**	**0.148**	**0.675**	**0.287**	**0.582**	**0.495**	**0.392**	**0.627**

MiaPRL$_b$ respectively. In addition, the Glove method in the GRAM model is used instead of the hyperbolic graph neural network to train the medical code in the medical knowledge network. The prediction performance of the model is tested, and the result is recorded as MiaPRL$_p$.

The experimental results, showcased in TABLE V, highlight the crucial role played by both the medical code pre-training and the medical path encoder in the model's disease prediction. Removing any component of the model leads to a degradation in performance, underscoring their significance. However, the importance of these components varies. Notably, removing the medical code pre-training processing results in a relatively large decrease in prediction performance. This is because the hyperbolic graph neural network pre-training process effectively captures the hierarchical structure information within the medical knowledge network. By doing so, it enhances the correlation between diseases and significantly improves prediction accuracy. These findings emphasize the necessity of incorporating both pre-training processes to achieve optimal disease prediction performance.

In addition, since the medical path construction and sampling process is random, the medical path with low correlation is likely to affect the model prediction. The model's performance is similar when using the glove method for pre-training medical paths and when using a hyperbolic neural network. This similarity indicates the important role of medical path pre training in learning effective medical path feature embedding vectors.

F. Model sensitivity analysis

In this section, we focus on the sensitivity of the model to hyper-parameters. These key parameters include the number of sampled medical paths, the number of self-coding modules and the number of multi-self-coding layers. According to the experimental setup and evaluation method, the performance of the model with different parameters is tested on the MIMIC-III dataset. The results are shown in Fig. 3.

Start by setting the multi self-attention coding layers to 2, then set the number of self-attention encoder modules to 1, 2, 4, and 8 in succession. It can be observed that the performance of the model is stable with different numbers of multi self-attention coding layers. The evidence suggests that the proposed method is robust and unaffected by changes in the hyper-parameters.

The number of self-attention encoder modules is then set to 4 and the number of multi self-attention coding layers to 1, 2, 4 and 8 respectively. The performance of the model did not change much. Both tests show that the auto-attention encoder module learns the most relevant features from the patient's historical medical records, helping the model to achieve good predictive performance and strong stability.

G. Functional analysis

We used the PyTorch platform to build a program that used the medical information association path representation learning model, and then we refined and tested the model using the pre-processed dataset. Once training was complete, we packaged the model and built a calling interface using Flask to provide diseases diagnosis support based on medical path learning.

First, the user selects a patient from the medical record list and extracts the patient's historical medical records. The ICD coding recognition module transforms the ICD coding in medical records into a one-hot format. Since each medical record contains multiple ICD codes, each patient has at least one medical record. Therefore, the patient's medical record can be represented as a 3-dimensional tensor, with any all sections filled with '0'. The trained model takes the patient's past medical treatment data as input via the disease-aided diagnosis module, which utilizes medical information link representation learning. The model produces representative feature embedding vectors that reflect the patient's past medical treatment history. Feed the patient's feature embedding vector into the trained classifier to obtain the probability distribution of the patient's potential disease risk. Select the top-k ailments with highest probability as the predicted results for the output.

VII. Conclusion

In this paper, we propose a representation learning method based on the association path of medical information. We construct a medical knowledge network with a tree structure based on the categorisation of medical concepts, and use the extracted association paths of medical data from the network to discover the underlying connection between different diseases. The model learns a representative medical concept representation vector by effectively capturing medical

TABLE V: Performances comparison of methods on predictions of MIMIC-III and MIMIC-IV datasets

Dataset	Model	K=2		K=5		K=10		K=20	
		Precision	Reacll	Precision	Recall	Precision	Recall	Precision	Recall
MIMIC-III	MiaPRL$_\alpha$	0.642	0.124	0.536	0.249	0.419	0.378	0.314	0.517
	MiaPRL$_b$	0.678	0.126	0.559	0.264	0.442	0.387	0.316	0.534
	MiaPRL$_p$	0.681	1.128	0.557	0.259	0.448	0.416	0.324	0.586
	MiaPRL	0.692	0.131	0.571	0.276	0.456	0.437	0.359	0.607
MIMIC-IV	MiaPRL$_\alpha$	0.776	0.135	0.628	0.254	0.496	0.387	0.387	0.603
	MiaPRL$_b$	0.773	0.125	0.672	0.271	0.542	0.436	0.373	0.618
	MiaPRL$_p$	0.768	0.135	0.671	0.274	0.567	0.463	0.381	0.583
	MiaPRL	0.785	0.148	0.675	0.287	0.582	0.495	0.392	0.627

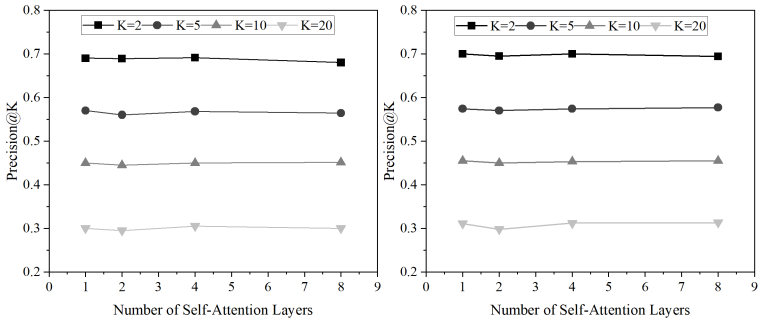

Fig. 3: Hyperparameter settings for comparative experiments

concepts and the hierarchical information of diseases. By taking into account the patient's past medical records and their timeliness, an accurate prognosis of potential diseases can be made.

This paper is not yet a definitive assessment of all the diseases in the dataset. Future studies could focus on certain diseases that are more relevant for diagnosis and treatment, such as certain long-term diseases and rare diseases. Medical facilities that complement traditional ones offer more precise and tailored treatments to patients, ultimately leading to an increase in the efficiency and effectiveness of healthcare.

Data Availability

The data that support the findings of this study are available from the corresponding author.

References

[1] T. Kuno, Y. Sahashi, S. Kawahito, M. Takahashi, M. Iwagami, and N. N. Egorova, "Prediction of in-hospital mortality with machine learning for covid-19 patients treated with steroid and remdesivir," *Journal of Medical Virology*, vol. 94, no. 3, pp. 958–964, 2022.

[2] T. Li, Z. Wang, W. Lu, Q. Zhang, and D. Li, "Electronic health records based reinforcement learning for treatment optimizing," *Information Systems*, vol. 104, p. 101878, 2022.

[3] Z. Liu, J. Zhang, Y. Hou, X. Zhang, G. Li, and Y. Xiang, "Machine learning for multimodal electronic health records-based research: Challenges and perspectives," in *Health Information Processing: 8th China Conference, CHIP 2022, Hangzhou, China, October 21–23, 2022, Revised Selected Papers*. Springer, 2023, pp. 135–155.

[4] S.-C. Lu, C. Xu, C. H. Nguyen, Y. Geng, A. Pfob, and C. Sidey-Gibbons, "Machine learning–based short-term mortality prediction models for patients with cancer using electronic health record data: systematic review and critical appraisal," *JMIR medical informatics*, vol. 10, no. 3, p. e33182, 2022.

[5] C. Thongprayoon, M. A. Mao, A. G. Kattah, M. T. Keddis, P. Pattharanitima, S. B. Erickson, J. J. Dillon, V. D. Garovic, and W. Cheungpasitporn, "Subtyping hospitalized patients with hypokalemia by machine learning consensus clustering and associated mortality risks," *Clinical Kidney Journal*, vol. 15, no. 2, pp. 253–261, 2022.

[6] S. Bacchi, Y. Tan, L. Oakden-Rayner, J. Jannes, T. Kleinig, and S. Koblar, "Machine learning in the prediction of medical inpatient length of stay," *Internal medicine journal*, vol. 52, no. 2, pp. 176–185, 2022.

[7] B. Alsinglawi, O. Alshari, M. Alorjani, O. Mubin, F. Alnajjar, M. Novoa, and O. Darwish, "An explainable machine learning framework for lung cancer hospital length of stay prediction," *Scientific reports*, vol. 12, no. 1, pp. 1–10, 2022.

[8] M. Tek, Y. Çavuşoğlu, C. Demirüstü, A. Birdane, A. Ünalır, B. Görenek, Ö. Göktekin, and N. Ata, "Levosimendan and dobutamine have a similar profile for potential risk for cardiac arrhythmias during 24-hour infusion in patients with acute decompensated heart failure," *Turk Kardiyoloji Dernegi Arsivi: Turk Kardiyoloji Derneginin Yayin Organidir*, vol. 38, no. 5, pp. 334–340, 2010.

[9] C. Mandel, K. Stich, S. Autexier, C. Lüth, A. Ziehn, K. Hochbaum, R. Dembinski, and C. Int-Veen, "Using gated recurrent unit networks for the prediction of hemodynamic and pulmonary decompensation," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 4584–4589.

[10] J. A. Burger, P. M. Barr, T. Robak, C. Owen, P. Ghia, A. Tedeschi, O. Bairey, P. Hillmen, S. E. Coutre, S. Devereux *et al.*, "Long-term efficacy and safety of first-line ibrutinib treatment for patients with cll/sll: 5 years of follow-up from the phase 3 resonate-2 study," *Leukemia*, vol. 34, no. 3, pp. 787–798, 2020.

[11] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, p. 6085, 2018.

[12] J. Zhu, Q. Jiang, Y. Shen, C. Qian, F. Xu, and Q. Zhu, "Application of recurrent neural network to mechanical fault diagnosis: A review," *Journal of Mechanical Science and Technology*, vol. 36, no. 2, pp. 527–542, 2022.

[13] Z. Alhassan, D. Budgen, R. Alshammari, N. Al Moubayed *et al.*, "Predicting current glycated hemoglobin levels in adults from electronic health records: validation of multiple logistic regression algorithm," *JMIR medical informatics*, vol. 8, no. 7, p. e18963, 2020.

[14] L. Yang, H. Wu, X. Jin, P. Zheng, S. Hu, X. Xu, W. Yu, and J. Yan, "Study of cardiovascular disease prediction model based on random forest in eastern china," *Scientific reports*, vol. 10, no. 1, p. 5245, 2020.

[15] J. F. Andry, F. M. Silaen, H. Tannady, and K. H. Saputra, "Electronic health record to predict a heart attack used data mining with naïve bayes method," *Int J Inf & Commun Technol ISSN*, vol. 2252, no. 8776, p. 8776, 2021.

[16] J. Stojanovic, D. Gligorijevic, V. Radosavljevic, N. Djuric, M. Grbovic, and Z. Obradovic, "Modeling healthcare quality via compact representations of electronic health records," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 3, pp. 545–554, 2016.

[17] B. A. Bates, E. Akhabue, M. M. Nahass, A. Mukherjee, E. Hiltner, J. Rock, B. Wilton, G. Mittal, A. Visaria, M. Rua *et al.*, "Validity of international classification of diseases (icd)-10 diagnosis codes for identification of acute heart failure hospitalization and heart failure with reduced versus preserved ejection fraction in a national medicare sample," *Circulation: Cardiovascular Quality and Outcomes*, vol. 16, no. 2, p. e009078, 2023.

[18] K. SLIMANI, Y. RUICHEK, and R. MESSOUSSI, "Compound facial

emotional expression recognition using cnn deep features," *Engineering Letters*, vol. 30, no. 4, pp. 1402–1416, 2022.

[19] S. K. Swee, L. C. Chen, T. S. Chiang, and T. C. Khim, "Deep convolutional neural network for sem image noise variance classification." *Engineering Letters*, vol. 31, no. 1, pp. 328–337, 2023.

[20] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1903–1911.

[21] J. Gao, X. Wang, Y. Wang, Z. Yang, J. Gao, J. Wang, W. Tang, and X. Xie, "Camp: Co-attention memory networks for diagnosis prediction in healthcare," in *2019 IEEE international conference on data mining (ICDM)*. IEEE, 2019, pp. 1036–1041.

[22] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun, "Stagenet: Stage-aware neural networks for health risk prediction," in *Proceedings of The Web Conference 2020*, 2020, pp. 530–540.

[23] X. Zhang, B. Qian, S. Cao, Y. Li, H. Chen, Y. Zheng, and I. Davidson, "Inprem: An interpretable and trustworthy predictive model for healthcare," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 450–460.

[24] M. Ye, J. Luo, C. Xiao, and F. Ma, "Lsan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1753–1762.

[25] E. Manzini, B. Vlacho, J. Franch-Nadal, J. Escudero, A. Génova, E. Reixach, E. Andrés, I. Pizarro, J.-L. Portero, D. Mauricio *et al.*, "Longitudinal deep learning clustering of type 2 diabetes mellitus trajectories using routinely collected health records," *Journal of biomedical informatics*, vol. 135, p. 104218, 2022.

[26] K. Yin, D. Qian, W. K. Cheung, B. C. Fung, and J. Poon, "Learning phenotypes and dynamic patient representations via rnn regularized collective non-negative tensor factorization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 1246–1253.

[27] C. Wu, T. Zhou, Y. Tian, J. Wu, J. Li, and Z. Liu, "A method for the early prediction of chronic diseases based on short sequential medical data," *Artificial Intelligence in Medicine*, vol. 127, p. 102262, 2022.

[28] Y. Chang, C. Chen, W. Hu, Z. Zheng, X. Zhou, and S. Chen, "Megnn: Meta-path extracted graph neural network for heterogeneous graph representation learning," *Knowledge-Based Systems*, vol. 235, p. 107611, 2022.

[29] M. M. Li, K. Huang, and M. Zitnik, "Graph representation learning in biomedicine and healthcare," *Nature Biomedical Engineering*, pp. 1–17, 2022.

[30] Y. Zhang, "Hyperbolic graph neural networks," in *Advances in Graph Neural Networks*. Springer, 2022, pp. 109–130.

[31] H. Kohli, J. Agarwal, and M. Kumar, "An improved method for text detection using adam optimization algorithm," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 230–234, 2022.

[32] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[33] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on mimic-iv dataset," *Scientific Reports*, vol. 12, no. 1, p. 7166, 2022.

[34] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[35] J. Liu, D. Capurro, A. Nguyen, and K. Verspoor, "Improving text-based early prediction by distillation from privileged time-series text," *arXiv preprint arXiv:2301.10887*, 2023.

[36] ——, "note bloat impacts deep learning-based nlp models for clinical prediction tasks," *Journal of biomedical informatics*, vol. 133, p. 104149, 2022.

[37] N. Liu, R. Gao, J. Yuan, C. Park, S. Xing, and S. Gou, "Gru-tv: Time-and velocity-aware gru for patient representation on multivariate clinical time-series data," *arXiv preprint arXiv:2205.04892*, 2022.

[38] E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Heart disease risk factors detection from electronic health records using advanced nlp and deep learning techniques," *Scientific Reports*, vol. 13, no. 1, p. 7173, 2023.