Information Fusion 102 (2024) 102069

Contents lists available at ScienceDirect

Information Fusion

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

EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation

Shuai Niu [b,a], Jing Ma [b], Liang Bai [d], Zhihua Wang [e], Li Guo [f], Xian Yang [a,c,*]

[a] Alliance Manchester Business School, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[b] The Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China
[c] Data Science Institute, Imperial College London, Exhibition Road, London, SW7 2BX, UK
[d] The Computer and Information Technology School, Shanxi University, Shanxi Road, Tai Yuan, Shan Xi, 030006, China
[e] Shanghai Institute for Advanced Study of Zhejiang University, Dangui Road, Shanghai, China
[f] Department of Computing and Mathematics, Manchester Metropolitan University, Oxford Road, Manchester, M15 6BH, UK

ARTICLE INFO

Dataset link: https://physionet.org/content/mimiciii/1.4/, https://portal.dbmi.hms.harvard.edu/projects/n2c2-2014/, https://github.com/Healthcare-Data-Mining-Laboratory/EHR-KnowGen

Keywords:
Multimodal learning
Multimodal electronic health records
Knowledge enhancement
Generative large language model
Disease diagnosis

ABSTRACT

Electronic health records (EHRs) contain diverse patient information, including medical notes, clinical events, and laboratory test results. Integrating this multimodal data can improve disease diagnoses using deep learning models. However, effectively combining different modalities for diagnosis remains challenging. Previous approaches, such as attention mechanisms and contrastive learning, have attempted to address this but do not fully integrate the modalities into a unified feature space. This paper presents EHR-KnowGen, a multimodal learning model enhanced with external domain knowledge, for improved disease diagnosis generation from diverse patient information in EHRs. Unlike previous approaches, our model integrates different modalities into a unified feature space with soft prompts learning and leverages large language models (LLMs) to generate disease diagnoses. By incorporating external domain knowledge from different levels of granularity, we enhance the extraction and fusion of multimodal information, resulting in more accurate diagnosis generation. Experimental results on real-world EHR datasets demonstrate the superiority of our generative model over comparative methods, providing explainable evidence to enhance the understanding of diagnosis results.

1. Introduction

Given the increasing prevalence of intelligent medical health, extensive collections of electronic health records (EHRs) have been accumulated to serve as valuable datasets for the advancement of deep learning in the healthcare domain. Common EHR modalities encompass diverse data types, which mainly can be concluded to be structured data (e.g., patients' demographics, laboratory testing results, clinical events, medications, etc.) and unstructured data (e.g., medical notes, radiology reports, magnetic resonance imaging (MRI), etc.). Structured and unstructured EHRs can often be used for different healthcare tasks with different deep learning models, such as disease diagnosis [1–3], disease risk prediction [4–6], patients' mortality prediction [7,8], and intensive care unit (ICU) stay time estimation [7]. Nonetheless, the prevailing research on the aforementioned task primarily centers around handling single modalities, with limited regard for accommodating multiple modalities, and a lack of emphasis on addressing unstructured medical notes. Consequently, these studies may overlook the potential for enhanced prediction outcomes by failing to extract the wealth of patient information embedded within medical notes and by neglecting a holistic comprehension of the patient's condition through integrated analysis of diverse modalities. In this study, our primary focus lies in the accurate determination of patients' disease diagnoses. To achieve this, we adopt a comprehensive approach that incorporates multiple EHR modalities, encompassing unstructured medical notes, clinical events, and laboratory testing results.

There has been a growing interest in research on multimodal EHR learning [3,7,9–11]. Current approaches to multimodal learning primarily focus on utilizing different deep learning models for each modality. For instance, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are commonly employed to process textual medical notes and numerical or discrete structured EHR data for disease risk prediction [3,7]. Furthermore, with the advent of large language models (LLMs) and their application in the healthcare research field, more and more LLMs such as ClinicalBERT [12], BioBERT [13], and MedBERT [14] have been applied to process textual EHR data for

* Corresponding author at: Alliance Manchester Business School, University of Manchester, Oxford Road, Manchester, M13 9PL, UK.
E-mail addresses: 20483007@life.hkbu.edu.hk (S. Niu), majing@comp.hkbu.edu.hk (J. Ma), bailiang@sxu.edu.cn (L. Bai), zhihua.wang@zju.edu.cn (Z. Wang), L.Guo@mmu.ac.uk (L. Guo), xian.yang@manchester.ac.uk (X. Yang).

https://doi.org/10.1016/j.inffus.2023.102069
Received 28 June 2023; Received in revised form 31 August 2023; Accepted 6 October 2023
Available online 13 October 2023

various downstream healthcare tasks. Consequently, the combination of LLMs with other deep learning models has gained popularity and demonstrated better prediction performance in recent healthcare research, leveraging the strengths of LLMs in processing natural language data [9,10,15]. However, this approach of using separate encoders to process different modalities and then attempting to fuse these modalities using techniques such as attention mechanisms [7,9] or contrastive learning [10] does not fully integrate the data from these modalities into a unified feature space.

Our research primarily centers around effectively integrating different modalities of EHR data, with a particular emphasis on unstructured textual data. We achieve this by leveraging generative LLMs. Recently, large language models such as LLaMA [16] and GPT4 [16] have emerged as a novel paradigm in various tasks involving diverse input data. These models stand out due to the abundant prompts that support instruction tuning within generative approaches, distinguishing them from most discriminant deep learning models. In the realm of healthcare research, PromptEHR [17] proposed a synthetic multimodal EHR generation model inspired by the generative approach of MedGAN [18]. PromptEHR utilized LLMs with the patient's demographic information as a conditional prompt, along with other modalities such as longitudinal diagnosis events and medication events. However, PromptEHR primarily focuses on aspects other than disease diagnosis. To the best of our knowledge, limited existing research specifically explores the application of generative methods for disease diagnosis. The beauty of introducing generative models for disease diagnosis lies in their ability to generate realistic and diverse samples that represent possible disease outcomes. Unlike classification models that assign a single label to each input, generative models can capture the complex relationships and variations within the data.

One of the challenges in multimodal EHR learning is that the extracted information may not necessarily contain clinically relevant insights, which hinders the potential for further improvement in prediction performance. To overcome this limitation, incorporating external domain knowledge becomes crucial. By incorporating domain knowledge, generative models can benefit from additional guidance during disease name generation from a semantic perspective, leading to improved accuracy in disease generation. Several research works [19–21] have utilized domain knowledge to construct knowledge graphs to generate more accurate disease diagnoses and provide interpretation evidence. Additionally, external knowledge from disease names has been incorporated through attention mechanisms to provide explainable disease risk predictions [1,2,4,9]. However, to the best of our knowledge, limited research has introduced domain knowledge at a semantic level through generative approaches using multimodal EHRs in the healthcare domain.

To address the research gap, we propose a model called Knowledge-enhanced Multimodal Learning for Disease Diagnosis Generation (EHR-KnowGen). Our approach aims to bridge the gap between multimodal EHR data and clinically relevant information by leveraging domain knowledge. We convert diverse types of multimodal data, such as medical notes, discrete clinical event sequences, and numerical laboratory testing results, into a unified textual format. The textual data from different modalities are then encoded using a unified LLM, with soft prompts introduced to mitigate the modality gap. Additionally, we integrate hierarchical fine-grained and coarse-grained domain knowledge to guide the fusion of multimodal EHR data and extract clinically relevant information. By incorporating this knowledge-enhanced data fusion mechanism, our model enhances the accuracy of disease diagnosis generation, empowering generative models to excel in the healthcare domain.

Our main contributions can be summarized as follows:

- To fill the gap of limited research in generative disease diagnosis, we propose a disease diagnosis generative model that encodes different EHR modalities into a shared encoding space based on T5 [22]. Moreover, we employ soft prompts for each modality to enhance the latent representation of the multimodal EHR embedding.
- To enable the extraction of disease-related information from multimodal EHR data and offer interpretive evidence for disease diagnosis results, we integrate external fine-grained and coarse-grained knowledge into the disease diagnosis generation process. This incorporation of knowledge serves as a guiding mechanism to enhance the accuracy and interpretability of disease diagnoses, empowering our model to provide valuable insights in the healthcare domain.
- To validate the effectiveness of our model, we evaluate it on publicly available datasets, including the multimodal EHR dataset MIMIC-III [23] and the single-modality EHR dataset N2C2-2014 [24].

2. Related works

2.1. Mutimodal learning with EHRs

Typical EHR data exhibit heterogeneity and the nature of multimodality. They typically contain textual modalities such as medical notes, numerical modalities such as laboratory testing results and ECG waveforms, and discrete modalities such as clinical events. In the context of textual EHRs, word embeddings with attention mechanisms have been employed for disease classification [1,2], and LLMs equipped with attention mechanisms have also been leveraged for multi-disease diagnosis [4,9]. For numerical and discrete EHR modalities, RNNs are commonly employed for multi-disease diagnosis, as demonstrated in studies such as [5,7,25–28]. Moreover, multimodal learning has been extensively explored and applied in various fields [29–33]. In the domain of healthcare research, multimodal learning has also been introduced to address the heterogeneity of multimodal EHR data, enhancing the predictive power of models through the appropriate fusion of different modalities. For example, RAIM [7] exhibited superior performance compared to single-modality models in predicting patient mortality, physiological decompensation, and ICU stay duration, which achieved this by simultaneously analyzing ECG waveforms, laboratory testing results, and clinical events using a multi-channel Gated Recurrent Unit (GRU) coupled with multiple attention mechanisms. MNN [3] adopted a unified approach to model longitudinal medical notes and multi-disease diagnosis codes of patients for diagnosis prediction. On the other hand, LDAM [9] employed channel-wise RNNs and ClinicalBERT with medical notes and laboratory testing results, establishing a connection between the two modalities using label-dependent attention mechanisms to provide an explainable and accurate disease risk prediction.

2.2. Large language models and prompt learning in healthcare

In the field of Natural Language Processing (NLP), language models have emerged as a powerful tool for various downstream NLP tasks. Prominent models like BERT [34], BART [35], T5 [22], and GPT3 [36] have been developed and widely utilized. Within the healthcare domain, models such as ClinicalBERT [12] and MedBERT [14] have successfully processed textual EHR data for different downstream tasks in healthcare research. Prompt learning, a promising technique, employs natural language prompts as input for generation tasks. It enables fine-tuning of LLMs for specific tasks with limited examples, thereby reducing the reliance on labeled data for training. Additionally, prompt learning helps bridge the gap in modality distribution by constructing suitable prompts for each modality [37]. Despite its potential, prompt learning in healthcare research remains relatively under-explored. Recently, PromptEHR [17] utilized LLMs and prompt learning for generating longitudinal multimodal EHRs. In this study, we incorporate soft prompts [38] into a generative large language

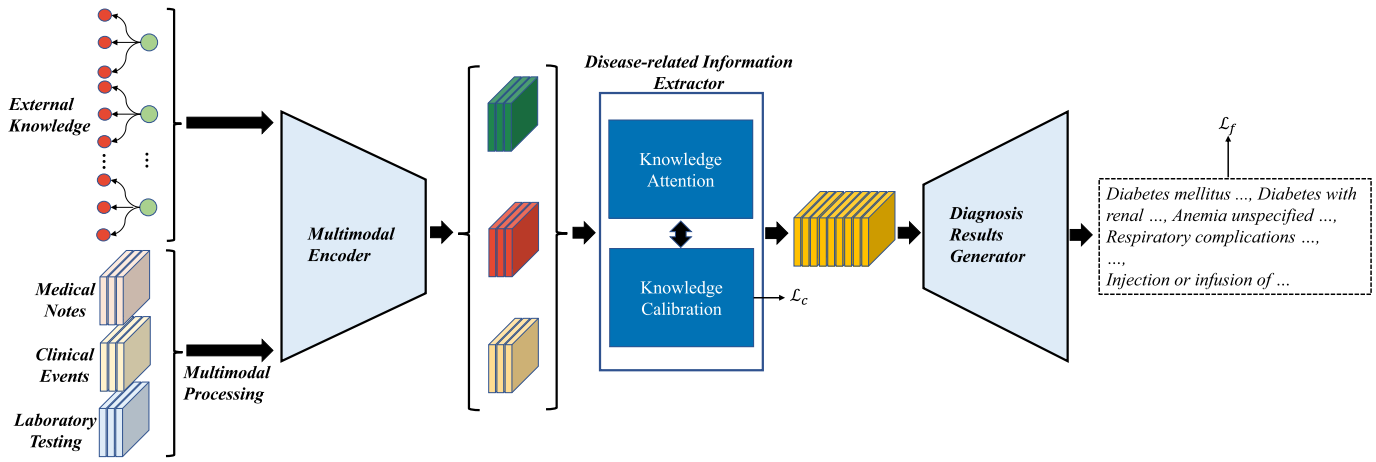

Fig. 1. The overall structure of EHR-KnowGen involves three components, multimodal encoder, disease-related information extractor, and diagnosis results generator. Initially, the multimodal EHRs, comprising medical notes, clinical events, and laboratory test results, are fed into the multimodal encoder. This encoder encodes multimodal EHRs and external knowledge into multimodal representation and external knowledge representations. Then, they are forwarded to the disease-related information extractor. This extractor comprises a knowledge attention module and a knowledge calibration module, which generates the knowledge-enhanced multimodal representation. Finally, the learned multimodal representation is used to generate disease names through a diagnosis results generator.

model to enhance the latent multimodal EHR representation, thereby contributing to the advancement of prompt learning in healthcare research.

2.3. Explicitly incorporating external knowledge

One of the key challenges in precision medicine is to enhance the interpretability of deep learning models while maintaining their predictive capabilities. A common approach to tackle this challenge involves incorporating external medical knowledge. For example, KAME [20] and GRAM [19] have utilized medical knowledge graphs with attention mechanisms on ICD codes for multi-disease diagnosis. The attention weights assigned to the knowledge graph nodes can represent the diseases a patient experiences during each hospital visit. CAML [1], LDAM [9], and LERP [4] have introduced external medical information through the cross-attention mechanism, using target disease risk labels in their models. This approach enables the disease diagnosis model to focus more on medical features related to the target label. PRIME [39] was the pioneering work that incorporated prior medical knowledge for risk prediction. It models prior medical knowledge as posterior regularization and learns the desired posterior distribution using a log-linear model. In our research, we incorporate fine-grained domain knowledge derived from the names of ICD-9 codes[1] and coarse-grained domain knowledge obtained from the names of grouped ICD-9 codes provided by HCUP-CSS codes[2] for disease diagnosis using a generative LLM.

3. Methodology

3.1. Model overview

Our proposed model, EHR-KnowGen, is specifically designed to leverage multimodal EHR data for multi-disease diagnosis. The overall architecture of EHR-KnowGen is represented in Fig. 1. It comprises three crucial modules, namely the multimodal encoder, disease-related information extractor, and diagnosis results generator. The objective of our model is to enable precise and interpretable disease diagnosis, thereby enhancing the quality of healthcare outcomes.

In the multimodal EHR encoder module, we process the multimodal EHR data $\boldsymbol{x}_n$ of each patient using a dedicated encoder. The encoder incorporates medical notes $\boldsymbol{m}_n$, clinical events $\boldsymbol{e}_n$, and laboratory testing results $\boldsymbol{l}_n$ to generate a multimodal representation $\boldsymbol{H}_n$. To improve the quality of this representation, we incorporate soft prompt embeddings $\boldsymbol{P}^{(m)}$, $\boldsymbol{P}^{(e)}$, and $\boldsymbol{P}^{(l)}$. These soft prompts provide additional guidance and context during the encoding process, enhancing the overall representation of the multimodal EHR data.

In the disease-related information extractor module, we introduce ICD codes to facilitate the extraction of disease-related information from the multimodal EHR data. This module employs a cross-attention mechanism to focus on and extract valuable disease-related details from the multimodal EHR embedding. Fig. 2 illustrates the coarse-grained and fine-grained ICD codes. The fine-grained diseases are represented as $\boldsymbol{C}^{(f)}$, while the coarse-grained diseases are represented as $\boldsymbol{C}^{(c)}$. For each patient n, the diseases to be diagnosed are represented as $\boldsymbol{y}_n^{(c)}$ and $\boldsymbol{y}_n^{(f)}$. $\boldsymbol{y}_n^{(f)}$ is the subset of the corresponding ICD codes. The presence or absence of diseases at the coarse-grained level is indicated by the binary vector $\boldsymbol{y}_n^{(c)}$, while $\boldsymbol{y}_n^{(f)}$ contains a list of fine-grained disease names that are related to patient n. This design enables the capturing of diseases at different levels of granularity. To predict the coarse-grained diseases, we employ a multi-class classifier. Determining coarse-grained diseases provides valuable insights that aid in determining fine-grained diseases, supporting the process of disease diagnosis. The disease-related information extractor module outputs the disease-related encoding matrices $\boldsymbol{H}_n^{(f)}$ and $\boldsymbol{H}_n^{(c)}$, which contain the extracted disease-related information at different levels of granularity.

In the disease diagnosis generator module, we employ the latent representations $\boldsymbol{H}_n^{(f)}$, $\boldsymbol{H}_n^{(c)}$, and $\boldsymbol{H}_n$ to generate the predicted fine-grained diseases $\boldsymbol{y}_n^{(f)}$. Instead of using a classifier, we utilize a language model-based generator due to the presence of fine-grained diseases and the potential challenge of limited sample sizes for each disease, which can make classification more difficult. In the subsequent subsections, we will provide a more detailed explanation of each module.

3.2. Multimodal encoder

In this section, we will introduce the multimodal encoder of the EHR-KnowGen model. With the objective of capturing and integrating various modalities within a unified encoding space, our approach employs a large language mode T5. To further enhance the latent multimodal representation, we introduce soft prompts for each EHR modality during the encoding process.

The example input multimodal EHR data is illustrated in Fig. 3. We apply three different textualization methods to convert three modalities

[1] https://www.cdc.gov/nchs/icd/icd9~cm.htm.
[2] https://hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp

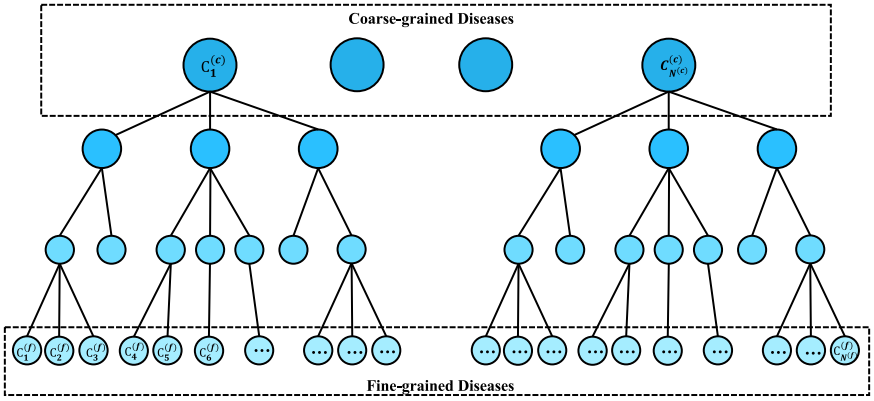

Fig. 2. The illustration of external domain knowledge from coarse-grained and fine-grained ICDs.

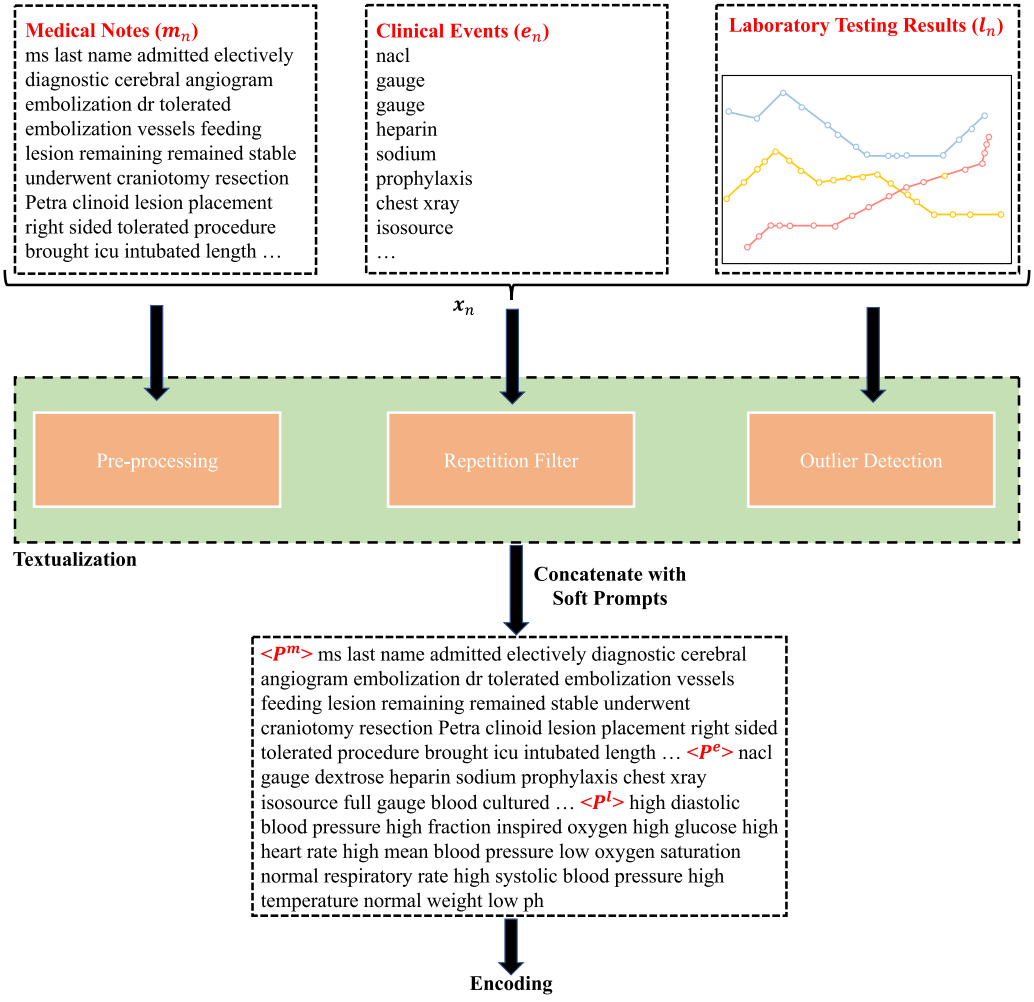

Fig. 3. Illustration of multimodal EHR data and data preprocessing steps.

into textual formats. For unstructured medical notes $\boldsymbol{m}_n$, we deploy data pre-processing methods to remove stopwords and noises. For discrete clinical events, we employ a filtering mechanism to remove repeated events, resulting in a refined set of events that are subsequently described in textual formats. As for the time-series laboratory testing results, we apply the box-plot outlier detection method [40] to identify abnormal signals within the continuous sequence. These abnormal signals are then described in textual formats. In addition, we define three learnable soft prompts $\boldsymbol{P}^{(m)}$, $\boldsymbol{P}^{(e)}$, and $\boldsymbol{P}^{(l)} \in \mathbb{R}^{N_p \times D}$ for medical notes, clinical events, and laboratory testing results, respectively. In this context, N_p denotes the length of the embeddings for the soft prompts, while D signifies the dimensionality of the embeddings. For the sake of convenience, we shall exclude the explicit mention of the length of the soft prompt embeddings (N_p) in the subsequent sections. The soft prompt embeddings are then fed into the encoder along with the medical notes, clinical events, and laboratory testing results to get the

multimodal representation:

$$\boldsymbol{H}_n = f_{enc}(\boldsymbol{P}^{(m)}, \boldsymbol{P}^{(e)}, \boldsymbol{P}^{(l)}, \boldsymbol{x}_n). \tag{1}$$

Here, $\boldsymbol{H}_n \in \mathbb{R}^{N_n \times D}$ represents the fused embeddings of the three modalities for patient n and N_n is the length of $\boldsymbol{x}_n$. $f_{enc}(\cdot)$ denotes the Transformer-based encoder. The incorporation of the soft prompts during the encoding process would facilitate the integration of diverse EHR data sources as discussed in [37].

3.3. Disease-related information extractor

The extraction and representation of valuable information from different multimodal EHRs, as well as capturing the relationships between various modalities, pose a significant challenge for a basic multimodal encoder. To overcome this challenge, we can leverage external medical knowledge to enhance feature extraction and facilitate the fusion of valuable multimodal information. In our approach, we incorporate fine-grained diseases to provide interpretation evidence that aids in the identification of important tokens and phrases from multimodal EHRs. This is accomplished through the use of a fine-grained knowledge attention module, which focuses on capturing disease-related features with a higher level of granularity. Furthermore, we introduce coarse-grained disease information as parent nodes within the knowledge representation. The coarse-grained diseases serve as higher-level categories or groups that include multiple fine-grained diseases. By incorporating coarse-grained disease information, we can refine and contextualize the extracted disease-related features, providing a more comprehensive and accurate representation of the patient's condition.

3.3.1. Fine-grained knowledge attention module

As illustrated in Fig. 4, the fine-grained knowledge attention module utilizes external domain knowledge from fine-grained disease names $\boldsymbol{C}^{(f)}$ to identify important features within the multimodal EHR embedding $\boldsymbol{H}_n$. To incorporate fine-grained knowledge, we introduce a learnable soft prompt $\boldsymbol{P}^{(f)} \in \mathbb{R}^{N_p \times D}$ and combine it with $\boldsymbol{C}^{(f)}$ within the encoder to get:

$$\boldsymbol{E}^{(f)} = f_{enc}(\boldsymbol{P}^{(f)}, \boldsymbol{C}^{(f)}), \tag{2}$$

where $\boldsymbol{E}^{(f)} \in \mathbb{R}^{|\boldsymbol{C}^{(f)}| \times D}$, $|\boldsymbol{C}^{(f)}|$ refers to the number of diseases in $\boldsymbol{C}^{(f)}$. Next, we use the embeddings $\boldsymbol{H}_n$ and $\boldsymbol{E}^{(f)}$ to compute the knowledge attention score vector $\boldsymbol{\alpha}_n$. This involves applying three learnable weights $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v$ to transform $\boldsymbol{E}^{(f)}$ and $\boldsymbol{H}_n$ into the query, key, and value matrices:

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{W}_q \boldsymbol{E}^{(f)}, \\ \boldsymbol{K}_n &= \boldsymbol{W}_k \boldsymbol{H}_n, \\ \boldsymbol{V}_n &= \boldsymbol{W}_v \boldsymbol{H}_n. \end{aligned} \tag{3}$$

The scaled-dot similarity between $\boldsymbol{K}_n$ and $\boldsymbol{Q}$ are obtained from:

$$\boldsymbol{G}_n = \boldsymbol{K}_n \cdot \boldsymbol{Q} = \frac{\boldsymbol{K}_n^T \boldsymbol{Q}}{\sqrt{D}}, \tag{4}$$

where, $\cdot$ is the scale-dot operation, T denotes the transpose operator, $\boldsymbol{G}_n \in \mathbb{R}^{N_n \times |\boldsymbol{C}^{(f)}|}$ is the scaled-dot similarity matrix.

To capture the relative spatial information of consecutive words and improve the extraction of implicit information from multimodal textual EHR data, we utilize a one-dimensional (1D) CNN along with a max-pooling layer applied to the scaled-dot similarity matrix $\boldsymbol{G}_n$. This process results in the following transformation:

$$\boldsymbol{u}_n = f_{mp}(f_{ReLU}(f_{conv}(\boldsymbol{G}_n, k_1, r)), k_2) \tag{5}$$

where $\boldsymbol{u}_n \in \mathbb{R}^{N_n}$ represents the resulting embedding after applying the 1D CNN and max-pooling layers. The function $f_{conv}(\cdot)$ denotes the 1D CNN layer, $f_{mp}(\cdot)$ represents the max-pooling layer, and $f_{ReLU}(\cdot)$ corresponds to the nonlinear activation layer. The parameter k_1 indicates the kernel width of the CNN, r denotes the padding size of the CNN, and k_2 represents the kernel width of the max-pooling operation. After obtaining $\boldsymbol{u}_n$, we employ the softmax function to generate the similarity score vector $\boldsymbol{\alpha}_n \in \mathbb{R}^{N_n}$, where its ith element is obtained from:

$$\alpha_{n,i} = \frac{e^{u_{n,i}}}{\sum_{i=1}^{N_n} e^{u_{n,i}}}. \tag{6}$$

Afterwards, the weighted fusion embedding vector $\boldsymbol{H}_n^{(f)} \in \mathbb{R}^{N_n \times D}$ is generated by combining $\boldsymbol{V}_n$ and $\boldsymbol{\alpha}_n$ through:

$$\boldsymbol{H}_n^{(f)} = \boldsymbol{V}_n \odot (\boldsymbol{\alpha}_n \mathbf{1}^T) + \boldsymbol{V}_n, \tag{7}$$

where, $\mathbf{1} \in \mathbb{R}^D$ represents a column vector of ones, $\odot$ denotes the element-wise production, $(\boldsymbol{\alpha}_n \mathbf{1}^T)$ results in a matrix where each column is a copy of the vector $\boldsymbol{\alpha}_n$.

3.3.2. Coarse-grained knowledge calibration module

In order to further improve the extraction and fusion of multimodal information in EHRs using external fine-grained knowledge, we introduce coarse-grained disease knowledge labels denoted as $\boldsymbol{y}_n^{(c)}$. These labels are utilized to calibrate and constrain the weights used in generating the knowledge attention scores $\boldsymbol{\alpha}_n$ during the multimodal fusion process, which can be achieved through backpropagation. The structure of the knowledge calibration module is depicted in Fig. 5. The multimodal embedding $\boldsymbol{H}_n^{(f)}$ is employed to generate a coarse-grained disease distribution score vector denoted as $\hat{\boldsymbol{y}}_n^{(c)} \in \mathbb{R}^{|\boldsymbol{C}^{(c)}|}$, where $|\boldsymbol{C}^{(c)}|$ refers to the number of diseases in $\boldsymbol{C}^{(c)}$ and each element i is obtained from:

$$\hat{\boldsymbol{y}}_n^{(c)} = \sigma(f_c(\frac{1}{N_n} \sum_{i=1}^{N_n} \boldsymbol{H}_{n,i}^{(f)})). \tag{8}$$

Here, $\boldsymbol{H}_{n,i}^{(f)}$ is the ith row of $\boldsymbol{H}_n^{(f)}$, $\sigma(\cdot)$ is the Sigmoid activation function, $f_c(\cdot)$ is a fully connected layer and used to decrease embedding dimension from D to $|\boldsymbol{C}^{(c)}|$. The loss function for predicting coarse-grained diseases is defined as:

$$\mathcal{L}_c = -\frac{1}{N|\boldsymbol{C}^{(c)}|} \sum_{n=1}^{N} \sum_{j=1}^{|\boldsymbol{C}^{(c)}|} (y_{n,j}^{(c)} \log(\hat{y}_{n,j}^{(c)}) + (1 - y_{n,j}^{(c)}) \log(1 - \hat{y}_{n,j}^{(c)})), \tag{9}$$

where N is the total number of patients.

The predicted coarse-grained disease label $\hat{\boldsymbol{y}}_n^{(c)}$ is utilized to generate an embedding matrix, $\boldsymbol{H}_n^{(c)}$, which contains coarse-grained disease-related information from EHRs. This embedding matrix is generated using the following equation:

$$\boldsymbol{H}_n^{(c)} = \boldsymbol{E}^{(c)} \odot (\hat{\boldsymbol{y}}_n^{(c)} \mathbf{1}^T) + \boldsymbol{E}^{(c)}, \tag{10}$$

where $(\hat{\boldsymbol{y}}_n^{(c)} \mathbf{1}^T)$ yields a matrix in which each column is a copy of the vector $\hat{\boldsymbol{y}}_n^{(c)}$. Here, we have:

$$\boldsymbol{E}^{(c)} = f_{enc}(\boldsymbol{P}^{(c)}, \boldsymbol{C}^{(c)}), \tag{11}$$

where $\boldsymbol{P}^{(c)} \in \mathbb{R}^{N_p \times D}$ is the learnable soft prompt for $\boldsymbol{C}^{(c)}$, $\boldsymbol{E}^{(c)} \in \mathbb{R}^{|\boldsymbol{C}^{(c)}| \times D}$ and $\boldsymbol{H}_n^{(c)} \in \mathbb{R}^{|\boldsymbol{C}^{(c)}| \times D}$.

3.4. Diagnosis results generator

In the diagnosis results generator, the matrices $\boldsymbol{H}_n$, $\boldsymbol{H}_n^{(f)}$, and $\boldsymbol{H}_n^{(c)}$ are first combined via:

$$\boldsymbol{H}_n^{(d)} = f_n(\boldsymbol{H}_n^{(c)} \oplus \boldsymbol{H}_n^{(f)} \oplus \boldsymbol{H}_n), \tag{12}$$

where f_n is a layer normalization layer. Then, a Transformer-based decoder is used to generate the disease diagnosis prediction $p(\boldsymbol{y}_n^{(f)} | \boldsymbol{H}_n^{(d)}; \boldsymbol{\theta})$. Therefore, the second training objective is to minimize the negative log-likelihood given by:

$$\mathcal{L}_f = -\frac{1}{N} \sum_{n=1}^{N} \log p(\boldsymbol{y}_n^{(f)} | \boldsymbol{H}_n^{(d)}; \theta) = -\frac{1}{N} \sum_{n=1}^{N} \log p(\boldsymbol{y}_n^{(f)} | \boldsymbol{x}_n, \boldsymbol{C}^{(c)}, \boldsymbol{C}^{(f)}; \boldsymbol{\theta}), \tag{13}$$

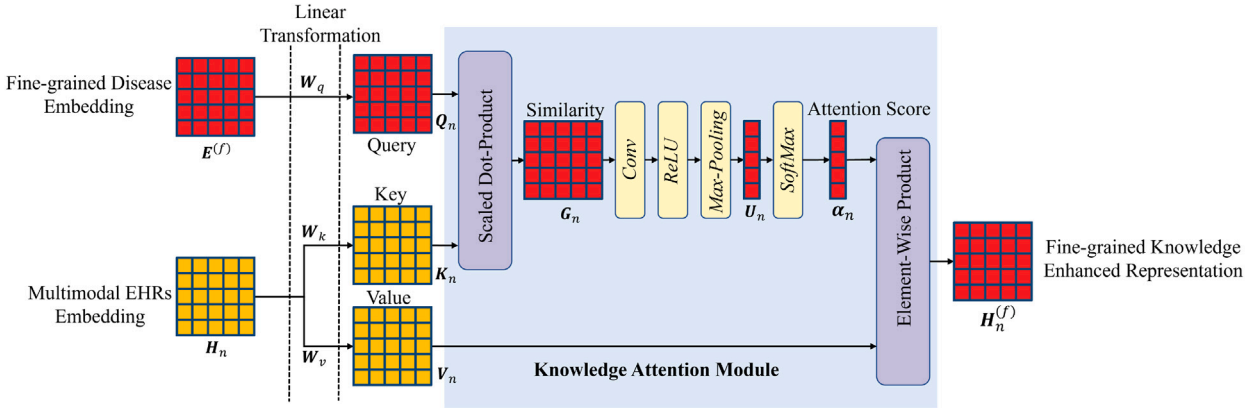

Fig. 4. Fine-grained knowledge attention module.

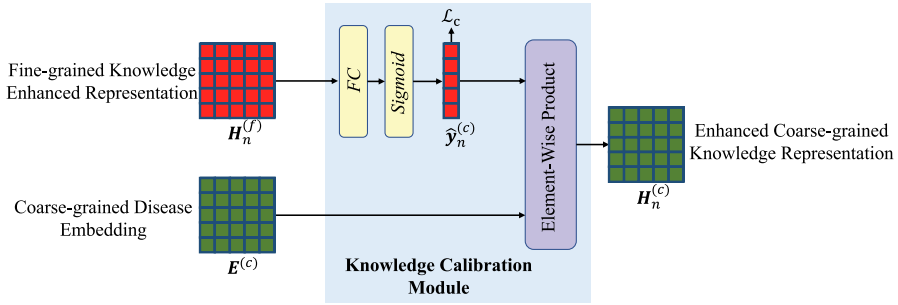

Fig. 5. Coarse-grained knowledge calibration module.

where θ represents the model parameters to be trained, $\boldsymbol{x}_n$ represents the input data for patient n, and $\boldsymbol{C}^{(c)}$ and $\boldsymbol{C}^{(f)}$ refer to the coarse-grained and fine-grained disease representations, respectively. The negative log-likelihood is calculated for each patient and then averaged over the entire dataset with N patients.

In EHR-KnowGen, the training objective for our model is defined as follows:

$$\mathcal{L} = \mathcal{L}_f + \lambda \mathcal{L}_c. \tag{14}$$

Here, $\mathcal{L}_c$ represents the prediction loss for coarse-grained diseases, $\mathcal{L}_f$ represents the negative log-likelihood loss for fine-grained disease prediction, and λ is a trade-off hyperparameter that balances the importance of the two objectives. The training procedure to optimize EHR-KnowGen by minimizing the loss defined in Eq. (14) is outlined in Algorithm 1.

4. Results and discussions

In the experimental section, we commence by introducing a publicly available EHR dataset that encompasses diverse modalities, thereby providing a comprehensive representation of patient information. This enables our proposed model, EHR-KnowGen, to leverage the rich and diverse data within EHRs for enhanced multi-disease diagnosis. To evaluate the performance of our model, we conduct quantitative assessments by comparing it against baseline methods in the disease diagnosis task. Through these experiments, we demonstrate the effectiveness and superiority of our proposed EHR-KnowGen model when compared to existing approaches.

To show the model robustness and effectiveness of different mechanisms of our model, we perform ablation studies and sensitivity analyses. This involves examining the performance of our model across different ablations of the various functions within each module, as well as exploring the impact of different hyperparameter settings. This analysis provides valuable insights into the effects of different modules and configuration options on the predictive capabilities of our model.

Furthermore, we place emphasis on the interpretability aspect of our model. Through thorough analyses, we investigate the knowledge attention mechanism employed by our model with fine-grained knowledge to identify the specific elements within the multimodal EHR data that significantly contribute to the disease diagnosis. This interpretability feature allows us to gain insights and explain the important factors considered by our model.

Through these comprehensive evaluations, our objective is to validate the effectiveness, interpretability, and applicability of our proposed EHR-KnowGen model in disease diagnosis using multimodal EHR datasets. Additionally, we specifically investigate the model's applicability in predicting less frequent or rare diseases, which often present challenges in accurate diagnosis and risk assessment. Furthermore, we also conduct a study using our model on a publicly available single modality dataset, N2C2-2014 [24], to demonstrate the effectiveness and robustness of our model in disease diagnosis.

4.1. Dataset

The MIMIC-III dataset [23] is a publicly available database containing de-identified health data from patients. In our research, we focused on utilizing medical notes, laboratory testing results, and clinical events as the primary input modalities for generating disease codes. Table 1 shows the details of the MIMIC-III dataset.

To preprocess the data, we performed several steps. We first extracted patients' EHRs that contained medical notes. Then, non-English words and stop-words of medical notes were removed from the medical notes and clinical events. We retained only unique clinical events for each hospital visit. A box-plot anomaly detection method [40] was applied to transform continuous laboratory testing results into textual descriptions that identify discrete anomaly signs. Additionally, we fill the missing data of clinical events or laboratory testing results with

Algorithm 1 The EHR-KnowGen model

Input Given patients' EHR data and external knowledge data $\boldsymbol{C}^{(f)}$ and $\boldsymbol{C}^{(c)}$, where each patient's EHR data $\boldsymbol{x}_n$ consists of multimodal information including medical notes $\boldsymbol{m}_n$, clinical events $\boldsymbol{e}_n$, and laboratory testing results $\boldsymbol{l}_n$.
while not converge **do**
  for Each batch **do**
    for Each patient n **do**
      Generate multimodal representation of EHR $\boldsymbol{H}_n$ from $\boldsymbol{x}_n$ using (1).
      Encode fine-grained knowledge data $\boldsymbol{C}^{(f)}$ to $\boldsymbol{E}^{(f)}$ based on (2).
      Calculate the fine-grained knowledge attention score vector $\boldsymbol{\alpha}_n$ using (3), (4), (5), and (6).
      Generate the knowledge-weighted multimodal embedding $\boldsymbol{H}_n^{(f)}$ using (7).
      Generate the coarse-grained disease distribution score vector $\hat{\boldsymbol{y}}_n^{(c)}$ using (8).
      Calculate the binary cross-entropy loss between predicted coarse-grained disease distribution $\hat{\boldsymbol{y}}_n^{(c)}$ and ground truth distribution $\boldsymbol{y}_n^{(c)}$ as in (9).
      Generate the coarse-grained disease embedding $\boldsymbol{H}_n^{(c)}$ which is weighted by $\hat{\boldsymbol{y}}_n^{(c)}$ based on (10) and (11).
      Combine $\boldsymbol{H}_n$, $\boldsymbol{H}_n^{(f)}$, and $\boldsymbol{H}_n^{(c)}$ using (12).
      Generate fine-grained diseases and calculate the negative log-likelihood loss of the disease generation as defined on (13).
    end for
    Update model parameters by minimizing the loss defined in Eq. (14) for patients in each batch.
  end for
end while

"no patient's clinical events" or "no patient's laboratory testing results". To evaluate the performance of our model, we adopted the same data splitting strategy as in [41], dividing the dataset into training and test sets at a ratio of 4:1 for performance evaluation.

4.2. Baselines

We utilized the following baseline methods for comparison in our study.

- **GRU**: GRU [42] is a typical type of RNNs that has been widely used for processing time-series numerical data. It employs gate mechanisms to capture long-range dependencies and model complex temporal dynamics in sequential data.
- **TRANS**: Transformer [42] is an encoder–decoder language model for processing sequential data. The main structure is based on the self-attention mechanism, which allows the model to pay different attention to different parts of the input sequence to calculate the representation of each element.
- **BERT**: BERT [34] is a pre-trained language understanding model that employs a bidirectional transformer encoder architecture for various NLP tasks, including sentiment analysis, named entity recognition, and question-answering systems.
- **CAML**: CAML [1] is an interpretable medical textual classification model that integrates label-embedding and cross-attention mechanisms to provide an interpretable medical text classification model. For a fair comparison, we upgraded the encoder of CAML with BERT.
- **GPT2**: GPT2 [43] is a pre-trained language understanding model that builds upon the transformer decoder architecture, utilizing a unidirectional, self-attention mechanism to generate coherent and contextually relevant text.

Table 1
Summary of the MIMIC-III dataset.

Dataset	MIMIC-III
# EHRs	22,220
# Patients	19,017
# Medical Notes	22,220
# Clinical Events	21,307
# Lab Testing	21,509
Data Samples	**Text**: "the female ... bathroom floor estimated initially elevated glucose admitted icu started insulin anion gap closed sugars transitioned insulin glargine units transferred medical acute renal failure initial creatine decrease wnl lisinopril restarted medical hip started perioperative beta-blocker medically cleared operating underwent or if tolerated procedure transferred recovery floor physical therapy improves strength transfused units packed red blood cells acute... **Event**: "kcl, normal saline, insulin, potassium phosphate" **Lab**: "normal diastolic blood pressure, high fraction inspired oxygen, high glucose, normal heart rate, high mean blood pressure, normal oxygen saturation, low respiratory rate, low systolic blood pressure, normal temperature, normal weight, normal ph"

- **T5**: T5 [22] is a versatile language model with encoder–decoder architecture that frames various NLP tasks as a unified text generation problem, achieving state-of-the-art results across a wide range of language understanding and generation tasks.
- **VSET**: VSET [44] is a transformer-based multimodal learning model for processing video and audio processing, by designing an attention-based multimodal fusion component. In this work, we replace the video with a concatenation of medical notes and clinical events and replace the audio input with laboratory testing results.
- **LDAM**: LDAM [9] is a multimodal learning model for multi-disease diagnosis by incorporating a label-dependent attention mechanism with modalities of medical notes and laboratory testing data, the discrete clinical events will be fed into the model with medical notes together.
- **PromptEHR**: PromptEHR [17] is an EHR generation model using prompt learning and a pretrained language understanding model. In this work, we use PromptEHR to process medical notes, clinical events, and laboratory testing results for multi-disease diagnosis.
- **LLaMA**: LLaMA [16] is a Human Feedback Reinforcement Learning (HFRL)-based instruction large language model. For this research, we use LLaMA on multimodal EHR data to diagnose diseases.

For all comparative models, except LLaMA, the learning rate was set to 1×10^{-5}, and the embedding size was 512. The ADAM optimizer was chosen for the model training. We also applied the dropout strategy with a dropout rate of 0.3. As for LLaMA, we trained it using the DeepSpeed framework,[3] with a learning rate of 2×10^{-5} and gradient accumulation. All models were implemented using PyTorch and trained on two NVIDIA TESLA A100-80G GPUs.

4.3. Disease diagnosis performance

4.3.1. Evaluation metrics

For the following baseline model comparison, we apply both Micro and Macro Precision, Recall, F1 score, and Accuracy to measure the model evaluation performance.

- **True Positives (TP)**: The actual class of the sample is positive and the result recognized by the model is also positive.

[3] https://github.com/microsoft/DeepSpeed

Table 2
Risk prediction results for the MIMIC-III dataset.

Models	MIMIC-III							
	Modality	Micro			Macro			ACC
		Pre	Recall	F1	Pre	Recall	F1	
BERT	M	0.3136	0.2202	0.2588	0.2643	0.1829	0.2075	0.3359
GPT2	M	0.3674	0.2047	0.2629	0.3166	0.1688	0.2091	0.3694
T5	M	0.3707	0.2041	0.2641	0.2880	0.1530	0.1967	0.3545
CAML	M	0.2978	0.2269	0.2618	0.2537	0.1859	0.2134	0.3407
LLaMA	M	0.3530	0.2351	0.2822	0.2910	0.1848	0.2178	0.3798
BERT	E	0.2049	0.0726	0.1072	0.1428	0.0407	0.0596	0.3382
GPT2	E	0.1829	0.0859	0.1169	0.1208	0.0566	0.0750	0.3173
T5	E	0.2007	0.0976	0.1313	0.1222	0.0643	0.0788	0.3543
LLaMA	E	0.2985	0.0447	0.0778	0.1420	0.0240	0.0368	0.3749
GRU	L	0.4286	0.0004	0.0008	0.0347	0.0001	0.0002	**0.4241**
TRANS	L	0.1324	0.0028	0.0050	0.0028	0.0027	0.0026	0.4180
BERT	M,E,L	0.3360	0.2176	0.2641	0.2833	0.1866	0.2188	0.3414
GPT2	M,E,L	0.3662	0.2116	0.2682	0.3313	0.1715	0.2163	0.3657
T5	M,E,L	0.3344	0.2262	0.2689	0.2814	0.1883	0.2153	0.3613
CAML	M,E,L	0.3079	0.2321	0.2658	0.2708	0.1972	0.2210	0.3256
VSET	M,E,L	0.3296	0.2268	0.2688	0.2788	0.1756	0.2049	0.3740
LDAM	M,E,L	0.3128	0.2388	0.2706	0.2776	0.2036	0.2273	0.3523
PromptEHR	M,E,L	0.3050	0.2601	0.2808	0.2499	0.2037	0.2155	0.3678
LLaMA	M,E,L	0.3197	0.2606	0.2871	0.2679	0.2020	0.2232	0.3489
EHR-KnowGen	M,E,L	0.2742	0.3228	**0.2965**	0.2354	0.2537	**0.2376**	0.3813

★ M, E, and L represents medical notes, clinical events, and laboratory testing results, respectively.

- **False Positives (FP)**: The actual class of the sample is negative but the result recognized by the model is positive.
- **True Negative (TN)**: The actual class of the sample is negative and the result recognized by the model is also negative.
- **False Negative (FN)**: The actual class of the sample is positive but the result recognized by the model is negative.

$$
\begin{aligned}
Micro\ Precision &= \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i},\\
Micro\ Recall &= \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i},\\
Micro\ F1 &= \frac{2 * Micro\ Precision * Micro\ Recall}{Micro\ Precision + Micro\ Recall},\\
Macro\ Precision &= \sum_i Precision_i / L,\\
Macro\ Recall &= \sum_i Recall_i / L,\\
Macro\ F1 &= \frac{2 * Macro\ Precision * Macro\ Recall}{Macro\ Precision + Macro\ Recall},
\end{aligned}
\tag{15}
$$

where i denotes the class index, and L represents the total number of classes.

4.3.2. Comparison with baseline methods

Based on the data presented in Table 2, we have made several key findings regarding the evaluation performance of various comparative models and our model, EHR-KnowGen.

Firstly, when only focusing on the textual modality, it is observed that BERT, GPT2, and T5 attain comparable evaluation performance within the encompassing Micro and Macro perspectives. Moreover, the introduction of label-embedding and cross-attention mechanisms in BERT by CAML shows a positive impact on disease diagnosis. Furthermore, LLaMA exhibits the most favorable evaluation performance in both Micro and Macro F1 scores and Accuracy, highlighting the significant effectiveness of the HFRL approach utilized in LLaMA. Secondly, when only considering the clinical events modality, large language understanding models face challenges in effectively leveraging their capabilities due to the limited volume and information available in each clinical event of a patient for making predictions on the disease diagnosis. Thirdly, in the case of purely using the time-series numerical laboratory testing modality, classical time-series processing models such as GRU and Transformer encounter difficulties in predicting a large amount of disease diagnosis problems. Despite achieving high accuracy values, their F1 scores are considerably low, suggesting label imbalance issues.

When incorporating multiple modalities, it is observed that most models achieve improved evaluation performance in terms of Micro F1 score, Macro F1 score, and Accuracy compared to models trained solely on a single modality. This suggests that the utilization of multiple modalities positively affects the task of multi-disease diagnosis. Furthermore, baseline models like VSET, LDAM, and LLaMA, which utilize multiple modalities, outperform competing baseline models that only use a single modality. Notably, LLaMA achieves the highest evaluation performance in both Micro and Macro F1 scores across all baseline models, which can be attributed to the employment of HFRL techniques. Additionally, generative models such as PromptEHR, LLaMA, and our proposed model EHR-KnowGen exhibit superior prediction performance compared to comparative discriminant models, indicating the superiority of generative models in handling complex multi-disease prediction tasks using the MIMIC-III dataset.

Finally, our proposed model, EHR-KnowGen, exhibits the most favorable evaluation performance on the MIMIC-III datasets. It achieves the best Micro and Macro F1 scores and Accuracy among all comparative models. These findings indicate that EHR-KnowGen represents a state-of-the-art generative model for effectively tackling the challenges associated with complex multi-disease prediction tasks. Overall, these findings shed light on the strengths and weaknesses of various models across different modalities and datasets, underscoring the efficacy of our EHR-KnowGen model in tackling the challenges inherent in complex disease prediction.

4.3.3. Model complexity analysis

The graph displayed in Fig. 6 offers an analysis of the computational times required by our proposed model, EHR-KnowGen, compared to several baseline models. These models were all run under identical conditions, utilizing consistent batch sizes and epoch numbers, and leveraging the computational capabilities of the NVIDIA TESLA A100-80G GPU and Xeon Gold 6226 CPU.

A close examination of Fig. 6 reveals that BERT, T5, and CAML have notably lower computational times compared to the other models. However, it is crucial to point out that these models still fall short of achieving optimal Micro and Macro F1 scores. On the opposite end of the spectrum, GPT2 takes an inordinately long time to process, and the

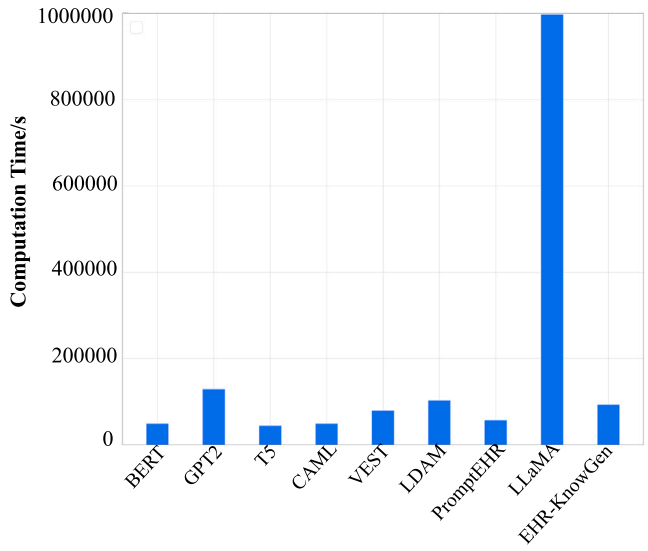

Fig. 6. Computation time of all comparative models during training.

incremental improvements it offers are modest at best when compared with other baseline language models. Among multimodal models like VEST, LDAM, and PromptEHR, the latter emerges as the most promising, albeit with increased computational time. Specifically, PromptEHR outclasses VEST and LDAM in both effectiveness and efficiency. On a different note, LLaMA, an HFRL model, achieves the highest Micro and Macro F1 scores among all baseline models, albeit with considerable computational demands.

Our EHR-KnowGen model's computational time aligns closely with multimodal models like VEST and LDAM. Importantly, it significantly outperforms all other models when evaluated on the designated metrics, underlining its superior efficacy in this context.

4.3.4. Ablation studies

In order to demonstrate the effectiveness of different mechanisms in our model, we included the following ablated versions in our experiment:

- **EHR-KnowGen-I**: EHR-KnowGen-I is an ablated version of our model EHR-KnowGen, where the whole information extractor module is removed.
- **EHR-KnowGen-II**: EHR-KnowGen-II is an ablated version of our model EHR-KnowGen, where the coarse-grained knowledge calibration module is removed while retaining the fine-grained knowledge attention module.
- **EHR-KnowGen-III**: EHR-KnowGen-III is an ablated version of our model EHR-KnowGen, where the soft prompts are removed.
- **EHR-KnowGen-IV**: EHR-KnowGen-IV is an ablated version of our model EHR-KnowGen, where a classifier was used instead of the generation decoder.

Table 3 presents the results of the ablation studies conducted on our EHR-KnowGen model, revealing several noteworthy findings. Firstly, EHR-KnowGen-I exhibited a significant decrease in both the Micro and Macro F1 scores, highlighting the effectiveness of the knowledge attention module and knowledge calibration module. The inclusion of the fine-grained knowledge attention module in EHR-KnowGen-II resulted in a noticeable improvement in its F1 scores. Conversely, EHR-KnowGen-III showed decreased performance across various evaluation metrics, emphasizing the importance of incorporating prompts for multimodal learning. Among the ablation models, EHR-KnowGen-IV also exhibits a notable decline in F1 score and Accuracy compared to our principal model, EHR-KnowGen, reinforcing the notion that generative models are more effective than discriminative models in complex disease diagnosis tasks.

In summary, the ablation studies conducted on EHR-KnowGen validate the efficacy of the knowledge attention module and knowledge calibration module in enhancing performance. They also highlight the significance of prompt learning in multimodal learning and underscore the superiority of generative models over discriminant models in the context of disease diagnosis tasks.

4.3.5. Sensitivity analysis

We conducted several sensitivity analyses to assess the impact of various factors on the performance of our EHR-KnowGen model. Fig. 7 illustrates the F1 scores and Accuracy achieved when varying different factors in our model.

We conduct a comparative analysis of four hyperparameters associated with EHR-KnowGen in both the model training and inference phases. Firstly, the hyperparameter λ serves as a pivotal factor in balancing the trade-off between two objective loss functions. Secondly, the number of beams and the number of temperatures emerge as crucial parameters. The former pertains to the count of candidate sequences taken into consideration throughout the generation process. Conversely, the latter parameter operates within the sampling process, regulating the degree of randomness and creativity of the generated text. Lastly, the prompt length is indicative of the scope encompassing the learnable soft prompts. Specifically, Fig. 7(a) shows the results obtained by varying the loss balance hyperparameter λ within the range of [0.01, 0.1, 1, 10, 100]. Fig. 7(b) depicts the results obtained with varying numbers of beams [1, 2, 3, 4]. Fig. 7(c) represents the results obtained with different temperature values, ranging from [0.2, 0.4, 0.6, 1.2, 1.6]. Fig. 7(d) displays the results obtained with different prompt embedding lengths N_p, using settings ranging from [4, 8, 12]. The results demonstrate that there is minimal significant fluctuation in the F1 scores and Accuracy, indicating that our EHR-KnowGen model is not highly sensitive to these specific hyperparameters.

4.4. Model interpretability

One of the noteworthy contributions of this study is the development of a disease diagnosis approach that yields multi-level interpretation results. The obtained explainable outcomes from our model, incorporating the knowledge attention mechanism and knowledge calibration module, are illustrated in Fig. 8. To demonstrate the effectiveness of our model, we randomly selected three patient cases from the test dataset. In Fig. 8, the highlighted input features correspond to those receiving high attention scores through the attention mechanism. The phrases highlighted in red represent the top 20% of important phrases, while the pink-highlighted phrases represent the top 40% of important phrases.

For example, in the case of the first patient diagnosed with "anaemia unspecified" and "other postoperative infection", our model assigns higher scores to phrases such as "secondary congestive heart", "apnea", "BiPAP", "low oxygen saturation", "high respiratory rate" and "low temperature". Research indicates that approximately one-third of congestive heart failure cases involve anaemia [45]. Additionally, anaemia can lead to reduced venous haemoglobin saturation and decreased tissue oxygen saturation [46], manifesting symptoms such as tachypnea and apnea [47]. Furthermore, individuals with anaemia often experience sensations of coldness, resulting in lowered body temperature [48]. The use of BiPAP is associated with an elevated risk of postoperative infection [49]. In addition, with respect to coarse-grained disease distribution, our model assigns the highest score to "deficiencies and other anaemia" as the coarse-grained cause of the diagnosis of the fine-grained disease "anaemia unspecified". The second highest score is given to the category of "complications of surgical procedures or medical care", which underlies the diagnosis of "other

Table 3
Ablation results of risk prediction for the MIMIC-III dataset.

Models	Modality	Micro			Macro			ACC
		Pre	Recall	F1	Pre	Recall	F1	
EHR-KnowGen-I	M,E,L	0.3736	0.2216	0.2781	0.3055	0.1749	0.2112	**0.3856**
EHR-KnowGen-II	M,E,L	0.3359	0.2499	0.2865	0.2847	0.1906	0.2170	0.3613
EHR-KnowGen-III	M,E,L	0.3194	0.2530	0.2829	0.2701	0.2015	0.2113	0.3457
EHR-KnowGen-IV	M,E,L	0.3356	0.2390	0.2792	0.3062	0.1853	0.2187	0.3605
EHR-KnowGen	M,E,L	0.2742	0.3228	**0.2965**	0.2354	0.2537	**0.2376**	0.3813

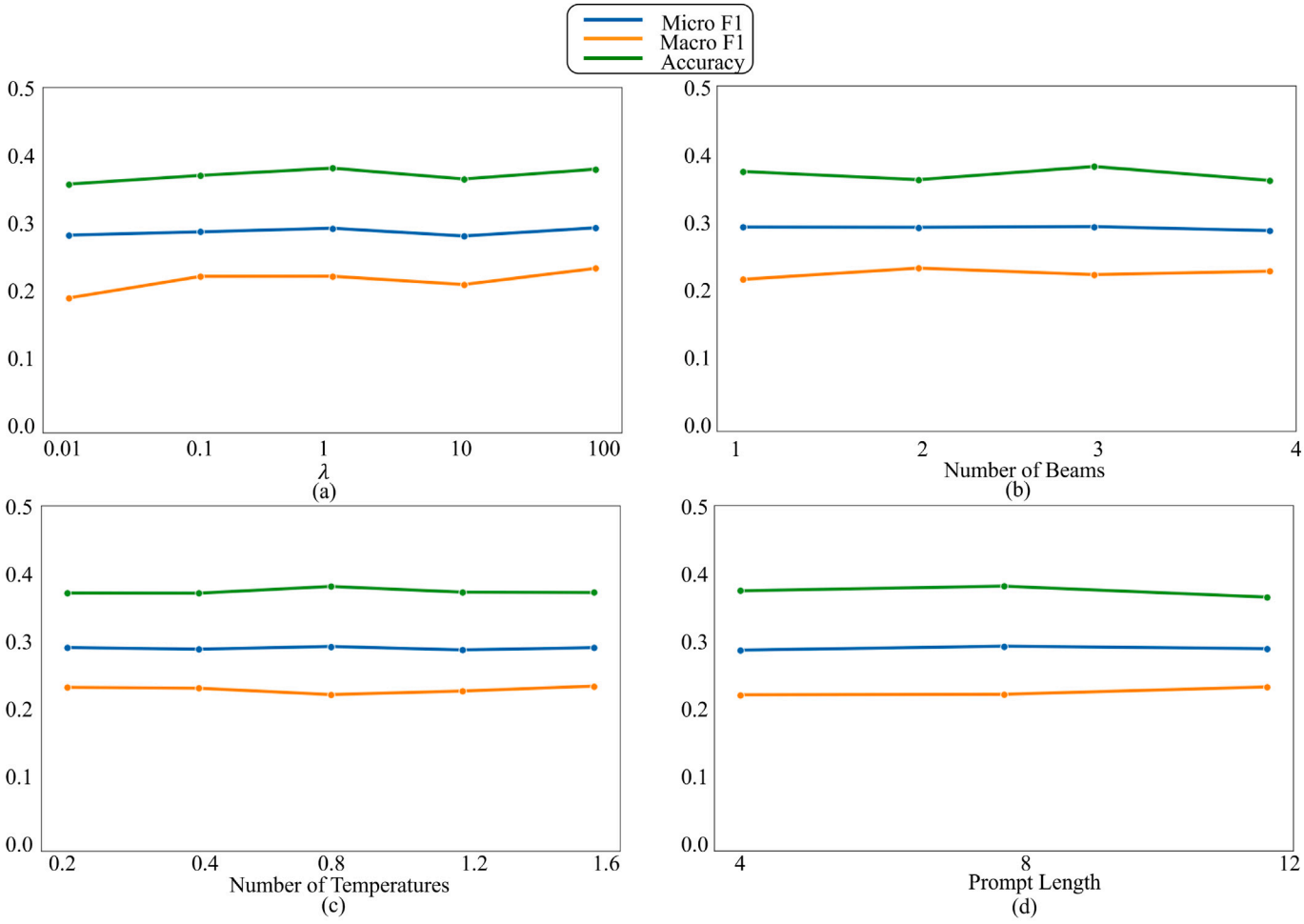

Fig. 7. The sensitive analysis with different hyper-parameters settings of our model EHR-KnowGen on the MIMIC-III dataset.

postoperative infection". Similar patterns were observed in the other two patient cases.

This analysis effectively shows how our model delivers interpretable and contextually relevant information for the multi-disease diagnosis, highlighting the importance of specific phrases and their associations with the diagnosed conditions.

4.5. Applicability in different scenarios

4.5.1. Applicability to less frequent diseases

To demonstrate the effectiveness of our EHR-KnowGen model in the context of rare disease diagnosis, we conducted an evaluation using the MIMIC-III dataset. Specifically, Fig. 9 shows the average F1 score for the Top 10% and Top 30% least common diseases. Upon analyzing this figure, it is evident that the comparative generative models PromptEHR and LLaMA did not exhibit superiority when compared to the comparative discriminant models in this specific context. However, our generative model EHR-KnowGen achieved the highest F1 scores for both the Top 10% and 30% least common diseases among all the comparative models. This outcome underscores the significance of the knowledge adaptor component in guiding disease generation and its crucial role in enhancing the performance of our model. This analysis highlights the effectiveness of EHR-KnowGen, specifically in the realm of rare disease diagnosis, and emphasizes the importance of incorporating the knowledge adaptor for improved disease generation outcomes.

4.5.2. Applicability to single-modal EHRs

To assess the robustness of our EHR-KnowGen model, we conducted an evaluation on an additional publicly available EHR dataset known as N2C2-2014 [24]. This dataset was specifically curated for natural language processing (NLP) research and consists of EHRs along with corresponding annotations. It encompasses 1304 medical notes obtained from 296 individuals. To prepare the dataset for our task, we removed stop-words and non-alphabetic characters from the medical notes. Our study focused on predicting four major disease diagnoses:

Fig. 8. Attention results obtained from the knowledge-attention module and knowledge calibration module for three randomly selected patient cases. Important data inputs are highlighted by the attention mechanism, where red represents the top 20% attention scores and pink represents the top 40% attention scores, which provide important information for multi-morbidity diagnosis.

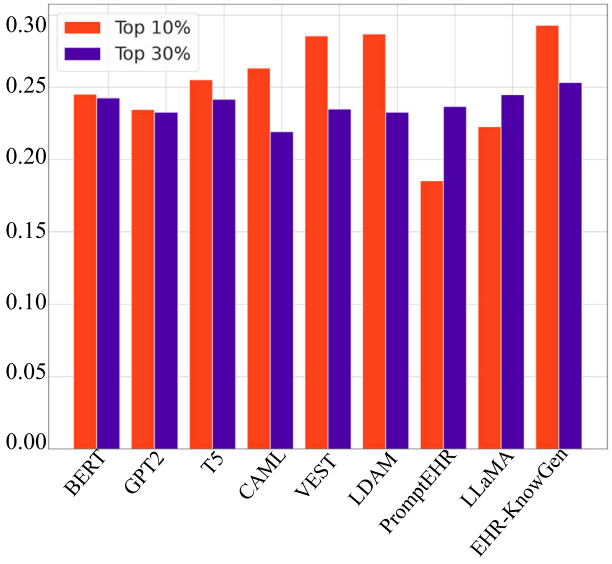

Fig. 9. The distribution of F1 scores for the Top 10% and the Top 30% of least common diseases.

"hyperlipidemia", "hypertension", "coronary artery disease", and "diabetes", which are due to their relevance to diseases of interest in all diagnoses. Performance assessment of the predictive models was carried out using an 4:1 training-test dataset split.

Table 4 presents the evaluation results of our EHR-KnowGen model alongside other comparative models. In contrast to the evaluation conducted on the MIMIC-III dataset, where LLaMA demonstrated the best performance in terms of F1 scores and Accuracy with other baseline models, here we found that GPT2 and T5 achieved similar results and exhibited the highest evaluation performance across all baseline models. This difference can be attributed to the requirement of a large volume dataset for the HFRL approach utilized by LLaMA, whereas the N2C2-2014 dataset consisted of only 1304 medical notes. Despite the relatively small size of the N2C2-2014 dataset, EHR-KnowGen outperformed all other comparative models in terms of Micro and Macro F1 scores and Accuracy. This result underscores the superior structure of our model for multi-disease diagnosis tasks.

5. Conclusions

This paper introduces EHR-KnowGen, a novel generative multimodal language model specifically designed for disease diagnosis using medical notes, clinical events, and laboratory testing results

Table 4
Risk prediction results for the N2C2 dataset.

Models	N2C2-2014						
	Micro			Macro			ACC
	Pre	Recall	F1	Pre	Recall	F1	
BERT	0.9154	0.9386	0.9269	0.9099	0.9309	0.9200	0.7018
GPT2	0.9378	0.9496	0.9436	0.9324	0.9427	0.9372	0.7631
T5	0.9329	0.9548	0.9437	0.9345	0.9455	0.9394	0.7544
CAML	0.9030	0.9677	0.9343	0.9015	0.9622	0.9305	0.7304
LLaMA	0.9317	0.9168	0.9242	0.9314	0.9003	0.9142	0.6839
EHR-KnowGen	0.9449	0.9534	**0.9492**	0.9389	0.9459	**0.9422**	**0.7826**

from EHRs. The primary objective of our model is to efficiently extract multimodal information from EHRs while providing interpretable evidence to support disease diagnosis. To achieve this goal, we propose a disease-related information extractor module that utilizes fine-grained domain knowledge to extract disease-related features from latent multimodal EHR representations. Furthermore, we incorporate coarse-grained domain knowledge to calibrate these extracted features, thereby enhancing interpretability. The resulting fine-grained and coarse-grained knowledge distributions serve as valuable evidence for disease diagnosis. Moreover, we leverage the latent representations of fine-grained and coarse-grained knowledge to guide the process of generating disease diagnosis results. Through extensive experimentation, our model consistently outperforms state-of-the-art generative language models and other discriminant models when applied to real-world EHR datasets like MIMIC-III, demonstrating its superiority in disease diagnosis. Remarkably, our model exhibits impressive performance even for less frequent diseases, showing its robustness and effectiveness as a knowledge-enhanced generative language model. Furthermore, our model also demonstrates strong capabilities in single-modality disease diagnosis, as evidenced by its performance on the N2C2-2014 dataset. These findings emphasize the effectiveness of EHR-KnowGen in accurately diagnosing diseases across diverse modalities, positioning it as a valuable and reliable tool in the field of healthcare.

CRediT authorship contribution statement

Shuai Niu: Conceptualization, Methodology, Data curation, Software, Writing – original draft. **Jing Ma:** Writing – review & editing, Supervision. **Liang Bai:** Formal analysis. **Zhihua Wang:** Formal analysis, Visualization. **Li Guo:** Validation. **Xian Yang:** Conceptualization, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors 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

Two datasets are publicly available at https://physionet.org/content/mimiciii/1.4/ and https://portal.dbmi.hms.harvard.edu/projects/n2c2-2014/. The source code is available at https://github.com/Healthcare-Data-Mining-Laboratory/EHR-KnowGen.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2021ZD0113303), and the National Natural Science Foundation of China (Nos. 62022052, 62276159).

References

[1] 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.
[2] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, L. Carin, Joint embedding of words and labels for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2321–2331.
[3] Z. Qiao, X. Wu, S. Ge, W. Fan, MNN: multimodal attentional neural networks for diagnosis prediction, Extraction 1 (2019) A1.
[4] S. Niu, Y. Song, Y. Qin, Y. Guo, X. Yang, Label-dependent and event-guided interpretable disease risk prediction using EHRs, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, BIBM, 2021.
[5] 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.
[6] J. Luo, M. Ye, C. Xiao, F. Ma, Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 647–656.
[7] Y. Xu, S. Biswal, S.R. Deshpande, K.O. Maher, J. Sun, Raim: Recurrent attentive and intensive model of multimodal patient monitoring data, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2565–2573.
[8] H. Harutyunyan, H. Khachatrian, D.C. Kale, G. Ver Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, Sci. Data 6 (1) (2019) 1–18.
[9] S. Niu, Y. Qin, Y. Song, Y. Guo, X. Yang, Label dependent attention model for disease risk prediction using multimodal electronic health records, in: Proceedings of the IEEE Conference on Data Mining, 2021, pp. 455–464.
[10] Z. Wang, Z. Wu, D. Agarwal, J. Sun, Medclip: Contrastive learning from unpaired medical images and text, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2022, pp. 3876–3887.
[11] Y. Meng, W. Speier, M.K. Ong, C.W. Arnold, Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression, IEEE J. Biomed. Health Inf. 25 (8) (2021) 3121–3129.
[12] E. Alsentzer, J.R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, W. Redmond, M.B. McDermott, Publicly available clinical BERT embeddings, in: NAACL HLT 2019, 2019, p. 72.
[13] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.
[14] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, NPJ Digit. Med. 4 (1) (2021) 86.
[15] B. Yang, L. Wu, How to leverage the multimodal EHR data for better medical prediction? in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 4029–4038.
[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, 2023, arXiv preprint arXiv:2302.13971.
[17] Z. Wang, J. Sun, PromptEHR: Conditional electronic healthcare records generation with prompt learning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2022, pp. 2873–2885.
[18] E. Choi, S. Biswal, B. Malin, J. Duke, W.F. Stewart, J. Sun, Generating multi-label discrete patient records using generative adversarial networks, in: Machine Learning for Healthcare Conference, PMLR, 2017, pp. 286–305.
[19] 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.
[20] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, J. Gao, Kame: Knowledge-based attention model for diagnosis prediction in healthcare, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 743–752.
[21] C. Yin, R. Zhao, B. Qian, X. Lv, P. Zhang, Domain knowledge guided deep learning with electronic health records, in: 2019 IEEE International Conference on Data Mining, ICDM, IEEE, 2019, pp. 738–747.
[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (1) (2020) 5485–5551.
[23] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.

[24] V. Kumar, A. Stubbs, S. Shaw, Ö. Uzuner, Creation of a new longitudinal corpus of clinical narratives, J. Biomed. Informat. 58 (2015) S6–S10.
[25] E. Choi, M.T. Bahadori, J.A. Kulas, A. Schuetz, W.F. Stewart, J. Sun, RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, pp. 3512–3520.
[26] Y. Ozyurt, M. Kraus, T. Hatt, S. Feuerriegel, AttDMM: an attentive deep Markov model for risk scoring in intensive care units, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 3452–3462.
[27] A.M. Alaa, M. van der Schaar, Attentive state-space modeling of disease progression, Adv. Neural Inf. Process. Syst. 32 (2019).
[28] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, Y. Liu, Making sense of spatio-temporal preserving representations for EEG-based human intention recognition, IEEE Trans. Cybern. 50 (7) (2019) 3033–3044.
[29] S. Wang, X. Liu, E. Zhu, C. Tang, J. Liu, J. Hu, J. Xia, J. Yin, Multi-view clustering via late fusion alignment maximization, in: IJCAI, 2019, pp. 3778–3784.
[30] 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.
[31] C. Ding, S. Sun, J. Zhao, MST-GAT: A multimodal spatial–temporal graph attention network for time series anomaly detection, Inf. Fusion 89 (2023) 527–536.
[32] G.M. Dimitri, S. Spasov, A. Duggento, L. Passamonti, P. Lió, N. Toschi, Multi-modal and multicontrast image fusion via deep generative models, Inf. Fusion 88 (2022) 146–160.
[33] N. Bahador, J. Jokelainen, S. Mustola, J. Kortelainen, Multimodal spatio-temporal-spectral fusion for deep learning applications in physiological time series processing: A case study in monitoring the depth of anesthesia, Inf. Fusion 73 (2021) 125–143.
[34] 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.
[35] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.
[36] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.
[37] M.U. Khattak, H. Rasheed, M. Maaz, S. Khan, F.S. Khan, Maple: Multi-modal prompt learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19113–19122.
[38] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3045–3059.
[39] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, A. Zhang, Risk prediction on electronic health records with prior medical knowledge, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1910–1919.
[40] P.J. Rousseeuw, M. Hubert, Anomaly detection by robust statistics, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8 (2) (2018) e1236.
[41] H. Harutyunyan, H. Khachatrian, D.C. Kale, G. Ver Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, Sci. Data 6 (1) (2019) 96.
[42] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: EMNLP, 2014.
[43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.
[44] K. Ramesh, C. Xing, W. Wang, D. Wang, X. Chen, Vset: A multimodal transformer for visual speech enhancement, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6658–6662.
[45] D.S. Silverberg, D. Wexler, A. Iaina, The role of anemia in the progression of congestive heart failure. Is there a place for erythropoietin and intravenous iron? J. Nephrol. 17 (6) (2004) 749–761.
[46] M. Meznar, R. Pareznik, G. Voga, Effect of anemia on tissue oxygenation saturation and the tissue deoxygenation rate during ischemia, Crit. Care 13 (Suppl 1) (2009) P238.
[47] J. Duan, X. Kong, Q. Li, S. Hua, S. Zhang, X. Zhang, Z. Feng, Association between anemia and bronchopulmonary dysplasia in preterm infants, Sci. Rep. 6 (1) (2016) 22717.
[48] H. Ludwig, K. Strasser, Symptomatology of anemia, in: Seminars in Oncology, Vol. 28, Elsevier, 2001, pp. 7–14.
[49] S.S. Ahmed, M.S. Yousuf, K. Samad, H. Ullah, K.M. Siddiqui, Factors influencing the use of postoperative bilevel positive airway pressure (BiPAP) in patients undergoing adult cardiac surgery: A retrospective cohort study, Health Sci. Rep. 5 (6) (2022) e873.