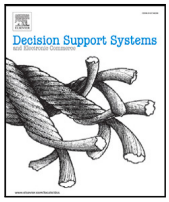
Decision Support Systems 181 (2024) 114228

Contents lists available at ScienceDirect

Decision Support Systems

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



Enhancing healthcare decision support through explainable AI models for risk prediction

Shuai Niu [a], Qing Yin [b], Jing Ma [a], Yunya Song [c], Yida Xu [d], Liang Bai [e], Wei Pan [f], Xian Yang [b,g,*]

[a] The Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China
[b] Alliance Manchester Business School, The University of Manchester, Oxford Rd, Manchester, M13 9PL, UK
[c] AI and Media Research Lab, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China
[d] The Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China
[e] The Computer and Information Technology School, Shanxi University, Shanxi Road, Tai Yuan, Shan Xi, China
[f] The Department of Computer Science, The University of Manchester, Oxford Rd, Manchester, M13 9PL, UK
[g] Data Science Institute, Imperial College London, South Kensington Campus, London, SW7 2AZ, UK

ARTICLE INFO

Keywords:
Explainable AI in healthcare
Healthcare decision support
Disease risk prediction
Modelling longitudinal patient data
Deep neural networks

ABSTRACT

Electronic health records (EHRs) are a valuable source of information that can aid in understanding a patient's health condition and making informed healthcare decisions. However, modelling longitudinal EHRs with heterogeneous information is a challenging task. Although recurrent neural networks (RNNs) are frequently utilized in artificial intelligence (AI) models for capturing longitudinal data, their explanatory capabilities are limited. Predictive clustering stands as the most recent advancement within this domain, offering interpretable indications at the cluster level for predicting disease risk. Nonetheless, the challenge of determining the optimal number of clusters has put a brake on the widespread application of predictive clustering for disease risk prediction. In this paper, we introduce a novel non-parametric predictive clustering-based risk prediction model that integrates the Dirichlet Process Mixture Model (DPMM) with predictive clustering via neural networks. To enhance the model's interpretability, we integrate attention mechanisms that enable the capture of local-level evidence in addition to the cluster-level evidence provided by predictive clustering. The outcome of this research is the development of a multi-level explainable artificial intelligence (AI) model. We evaluated the proposed model on two real-world datasets and demonstrated its effectiveness in capturing longitudinal EHR information for disease risk prediction. Moreover, the model successfully produced interpretable evidence to bolster its predictions.

1. Introduction

Decision support systems are evolving within healthcare to aid clinicians in intricate decision-making processes by leveraging information derived from clinical knowledge and patients' electronic health records (EHRs). EHRs constitute comprehensive repositories of diverse healthcare data, encompassing unstructured medical notes, clinical events, laboratory testing results, medical images, and other information generated across multiple hospital visits. Typical applications of utilizing EHRs to advance the precision medicine involve tasks such as disease risk prediction [1–5], statistical phenotype prediction [6], estimation of intensive care units (ICUs) stay duration [7], mortality prediction [6], survival prediction [8], and disease diagnosis [6,9]. However, the efficacy of computational models is constrained by their ability to handle high-dimensional, longitudinal, discrete, irregular, and heterogeneous EHRs [1].

This paper investigates a novel healthcare decision support model that extracts the representation of latent states from longitudinal EHRs to explore explainable patient trajectories for disease risk prediction. While various neural network-based methods have been developed to model longitudinal EHRs and learn latent states, such as recurrent neural networks (RNNs) [10,11] and convolutional neural networks (CNNs) [12,13], which are often regarded as data-driven black-box approaches [14]. Predictive clustering [15,16], on the other hand, has recently emerged as a novel approach for disease prediction tasks by clustering patients' latent health states into several groups and providing cluster-level explainable evidence for prediction results. In

* Corresponding author at: Alliance Manchester Business School, The University of Manchester, Oxford Rd, Manchester, M13 9PL, UK.
E-mail addresses: 20483007@life.hkbu.edu.hk (S. Niu), qing.yin-2@postgrad.manchester.ac.uk (Q. Yin), majing@comp.hkbu.edu.hk (J. Ma), yunyasong@hkbu.edu.hk (Y. Song), xuyida@hkbu.edu.hk (Y. Xu), bailiang@sxu.edu.cn (L. Bai), wei.pan@manchester.ac.uk (W. Pan), xian.yang@manchester.ac.uk (X. Yang).

https://doi.org/10.1016/j.dss.2024.114228

Available online 18 April 2024

this paper, we develop an explainable artificial intelligence (AI) model for disease risk prediction based on the fundamental principles of predictive clustering.

In the task of disease risk prediction, cluster assignment changes demonstrate the shifts in patients' latent health states and provide clusters of patients with similar health characteristics, aiding our understanding of the factors contributing to their health outcomes. Determining the optimal number of clusters when applying predictive clustering models to different EHR datasets is a significant challenge, as the number of clusters can significantly affect the results, and the optimal value can vary across datasets. To address this issue, we propose a neural non-parametric approach based on the Dirichlet Process Mixture Model (DPMM) [17] for predictive clustering in disease risk prediction tasks. DPMM is formulated as an infinite mixture model, and the stick-breaking construction of the Dirichlet process is employed. We can cluster data without pre-defining the exact number of clusters using DPMM. The model is trained using Stochastic Gradient Variational Bayes (SGVB) [18] to couple non-parametric clustering with neural networks.

To increase the level of explainability in our model, we explore both cluster-level explainability methods and local-level ones. The attention mechanism [19] is a popular local-level explainability method that has been used extensively in the literature to provide detailed explanations of prediction tasks. This mechanism assigns importance scores to input features and identifies crucial medical terms that contribute to disease risk prediction. Previous studies have shown the effectiveness of attention mechanisms for EHR data analysis [9,10,20–22]. Therefore, we incorporate attention mechanisms into our model to enhance its explainability. Notably, our focus is on modelling unstructured patient information, such as medical notes and auxiliary information from clinical events and laboratory testing results. The patient data found in medical notes is highly valuable and demands special consideration. However, due to the unstructured nature of this data, developing models to analyse it can be challenging, and there is a paucity of research in this domain. In order to transform the unstructured textual data derived from medical notes, we employ Clinical-BERT [23], which is a robust clinical language understanding model utilized as the text encoder. However, heterogeneous data from different modalities may have domain discrepancies that can lead to sub-optimal performance. To address this, we adopt soft Prompt learning [24–26] to reduce domain discrepancies between different modalities and better integrate heterogeneous information from medical notes and other modalities.

In this paper, we introduce a novel neural model named **Dir**ichlet Process-based **Pred**ictive Clustering (**DirPred**) for the purpose of disease risk prediction. The primary contributions of our work are outlined below:

- We present an explainable AI model designed to predict disease risk and enhance healthcare decision-making by incorporating multi-level explainability. This is achieved through the integration of predictive clustering and the attention mechanism.
- To capture the temporal dependencies in longitudinal EHRs, our predictive model comprises a prior module that encodes information from previous time steps and a posterior module that encodes observations at the current time step. The model parameters are learned through stochastic gradient descent and variational Bayesian inference.
- We address the challenge of encoding heterogeneous information from EHRs into a unified encoding space by proposing a soft Prompt learning-based data encoding approach.
- To validate the effectiveness of the proposed model, we apply DirPred to two publicly available EHR datasets, namely MIMIC-III [27] and N2C2-2014 [28].

2. Related work

2.1. AI models for disease risk prediction

In recent years, several AI models have been proposed for disease risk prediction in healthcare decision-making using EHR data, including time-aware, knowledge-aware, and attention-based models [2,29–34]. Given the longitudinal nature of EHR data, time-aware models utilize time information during the model construction process. For example, RetainEX [30] and ConCare [35] assumed that patient information may decay between consecutive visits and thus applied the information decay function to assist time-series data encoding for disease risk prediction. Knowledge-aware models aim to improve risk prediction performance by incorporating external information, such as medical knowledge graphs [33,34] and disease-related information [36]. Attention-based models have played a significant role in risk prediction along with time-aware and knowledge-aware models. The attention mechanism [19] is a popular approach to interpreting the results generated by deep neural networks. For example, RETAIN [10] introduced the attention mechanism to the RNNs-based predictive model to provide explainable results with high prediction accuracy. DIPOLE [20] adopted the attention-based bidirectional RNNs for diagnosis prediction. RAIM [7] and MNN [37] relied on the attention mechanism to assign different weights to different variants for information extraction and data aggregation. The work in [2,9,22] adopted a label-dependent attention approach to help capture clinical terms from medical notes.

2.2. Modelling EHR data

EHRs contain heterogeneous information from multiple modalities, reflecting patients' health states from different aspects. For the time-series laboratory testing data, many RNNs-based models have been developed [7,38,39]. Meanwhile, attention mechanisms have been increasingly introduced to generate explainable prediction results from medical notes [9,22]. To model multimodal EHRs, RAIM [7] applied RNNs and attention mechanisms to handle both laboratory testing data and Electrocardiogram (ECG) waveform data. LDAM [2], on the other hand, employed the label-dependent attention mechanism as the bridge to fuse laboratory testing data with medical notes, demonstrating that the inclusion of disease risk-related prompts can lead to better predictive performance.

Apart from heterogeneity, EHR data are also longitudinal, storing patient health information collected from multiple hospital visits. A variety of neural network-based models, such as RNNs and DSSMs, have been developed to extract information from longitudinal data. For example, GameNet [11], an RNNs-based model with an attention mechanism, was developed for disease diagnosis and drug recommendation. The work in [15] attempted to understand disease progression using deep predictive clustering, where encoded time-series data samples were clustered over time. The methods in [12,13] used multi-level CNNs to capture the complex changes of EHRs. Along with RNNs and CNNs-based methods, DSSMs [39–41] have also played an essential role in modelling longitudinal EHR data. For instance, the work in [39] developed an attentive deep Markov model to trace patients' latent states and predict disease risk from laboratory testing results. The work in [40] proposed a causal hidden Markov model to learn separate latent representations with different supervised tasks, including medical image reconstruction and risk prediction. In this paper, we primarily concentrate on utilizing unstructured patient data, comprising medical notes, for the purpose of disease risk prediction. As far as our understanding extends, only a few existing works focus on constructing deep neural networks to represent this input data longitudinally.

2.3. Predictive clustering in healthcare decision making

Traditional unsupervised clustering models, such as K-means and hierarchical clustering, often struggle to meet our expectations for

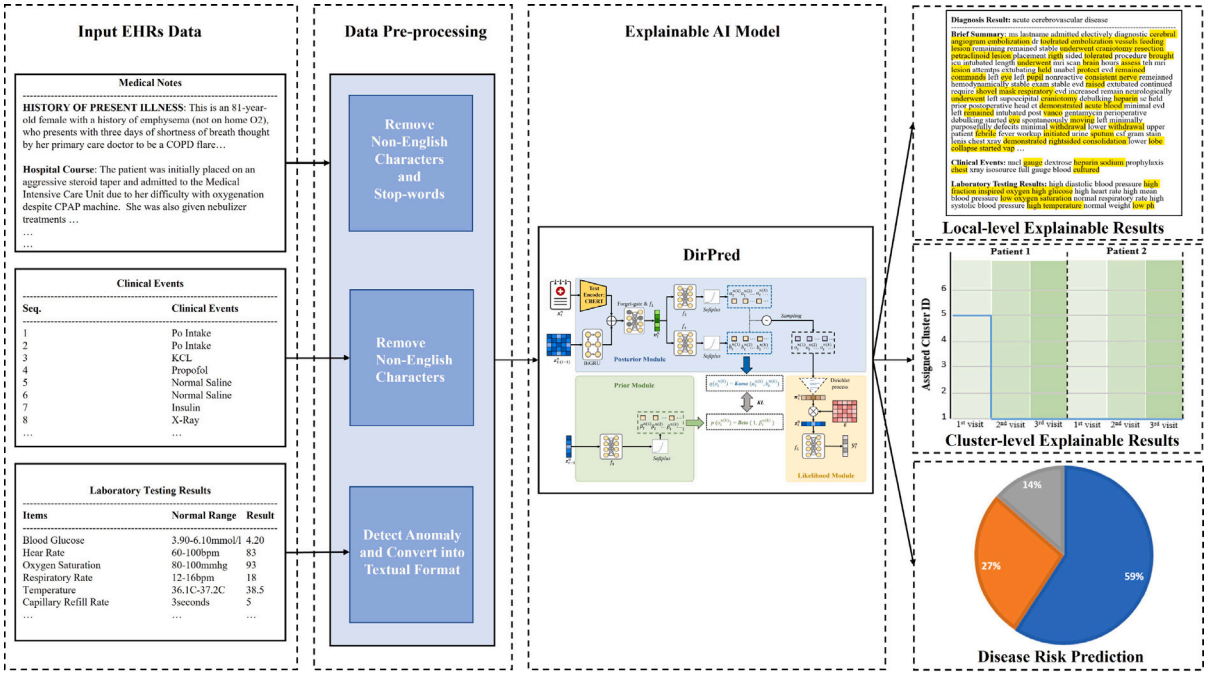


Fig. 1. The flowchart of applying the DirPred model for healthcare decision-making in disease risk prediction.

most prediction tasks. Furthermore, there have been limited efforts to use unsupervised cluster approaches for predicting risks at different time points, particularly for longitudinal EHRs. With the widespread use of neural networks, recent studies have explored leveraging their ability to learn latent representations from raw data. [15] introduced a predictive clustering model called ACTPC, which directed the unsupervised clustering process using a supervised task. Patients' latent health states were clustered into different groups, and the embedding of its cluster center characterized each group. As a continuation of the ACTPC framework, CAMELOT [16] was developed, replacing the non-differentiable selector network with an identifier network, thereby enabling end-to-end training. Nevertheless, these techniques necessitated a prior specification of the precise quantity of clusters and bestowed insufficient emphasis on modelling unstructured longitudinal medical records. This paper focuses on developing a non-parametric predictive clustering model by introducing a neural Dirichlet process to learn the number of clusters automatically.

3. Methodology

In this paper, we propose the DirPred model for disease risk prediction in healthcare decision-making, as illustrated in Fig. 1. The decision-making process comprises several key steps: collecting heterogeneous information from EHRs, pre-processing the data, modelling the data using DirPred, and generating multi-level explainable results with disease risk predictions. In the following sections, we provide a detailed description of the DirPred model. Firstly, we explain the fundamental concepts of predictive clustering and the Dirichlet process. Secondly, we describe the three main modules of DirPred. Lastly, we define the loss function for training the model.

3.1. Preliminary knowledge

3.1.1. Predictive clustering

Suppose each patient n is characterized by a sequence of EHRs collected from multiple hospital visits, where the data sample at each visit t is denoted as $\boldsymbol{x}_t^n$. In the risk prediction task, the presence of disease risks $\{\boldsymbol{y}_1^n, \ldots, \boldsymbol{y}_t^n\}$ are predicted using the information contained in $\{\boldsymbol{x}_1^n, \ldots, \boldsymbol{x}_t^n\}$. Let $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{\mathcal{K}\times D}$ denotes the embedding matrix of cluster centers, where $\mathcal{K}$ is the number of clusters, and D denotes the embedding size, set to the default value of 768 as in [42]. $\boldsymbol{\mathcal{E}}$ can be initialized by the K-means clustering results of all EHRs and updated by back-propagation through end-to-end training. In the setting of predictive clustering [15], $\boldsymbol{\mathcal{E}}$ is used to predict risks via:

$$\hat{\boldsymbol{y}}_t^n = f(\boldsymbol{\mathcal{E}}^T \boldsymbol{\pi}_t^n), \tag{1}$$

where $f(.)$ is the prediction network composed of fully connected layers, and $\boldsymbol{\pi}_t^n \in \mathbb{R}^{\mathcal{K}}$ is the distribution of cluster assignment for patient n at time t learned by neural networks via encoding the information from $\{\boldsymbol{x}_1^n, \ldots, \boldsymbol{x}_t^n\}$.

3.1.2. Dirichlet process

The Dirichlet Process Mixture Model creates clusters without predefining the number of clusters by assuming that the cluster assignment of each sample is generated via the Dirichlet process [43]. The stick-breaking construction of the Dirichlet process is represented as follows, which can be seen as a stick being broken into several pieces:

$$\pi_t^{n(k)} = \begin{cases} o_t^{n(1)}, & \text{if } k = 1 \\ o_t^{n(k)} \prod_{j<k} (1 - o_t^{n(j)}), & \text{for } k > 1. \end{cases} \tag{2}$$

Here, $\pi_t^{n(k)}$ is the kth element of $\boldsymbol{\pi}_t^n$ ranging from 0 to 1 satisfying $\sum_k^{\infty} \pi_t^{n(k)} = 1$, and $o_t^{n(k)}$ is drawn from a Beta distribution:

$$Beta(1, \beta_t^{n(k)}) = \beta_t^{n(k)} (1 - o_t^{n(k)})^{(\beta_t^{n(k)} - 1)}. \tag{3}$$

To infer the non-parametric distribution $\boldsymbol{\pi}_t^n$ coupled with deep neural networks, SGVB is widely adopted. As the Beta distribution does not meet the requirement of SGVB to have differentiable non-centered parametrization (DNCP), the work in [44] chose the Kumaraswamy (Kuma) distribution as the approximated posterior of $o_t^{n(k)}$, which is written as:

$$Kuma(a_t^{n(k)}, b_t^{n(k)}) = a_t^{n(k)} b_t^{n(k)} o_t^{n(k)(a_t^{n(k)} - 1)} (1 - o_t^{n(k) a_t^{n(k)}})^{(b_t^{n(k)} - 1)}. \tag{4}$$

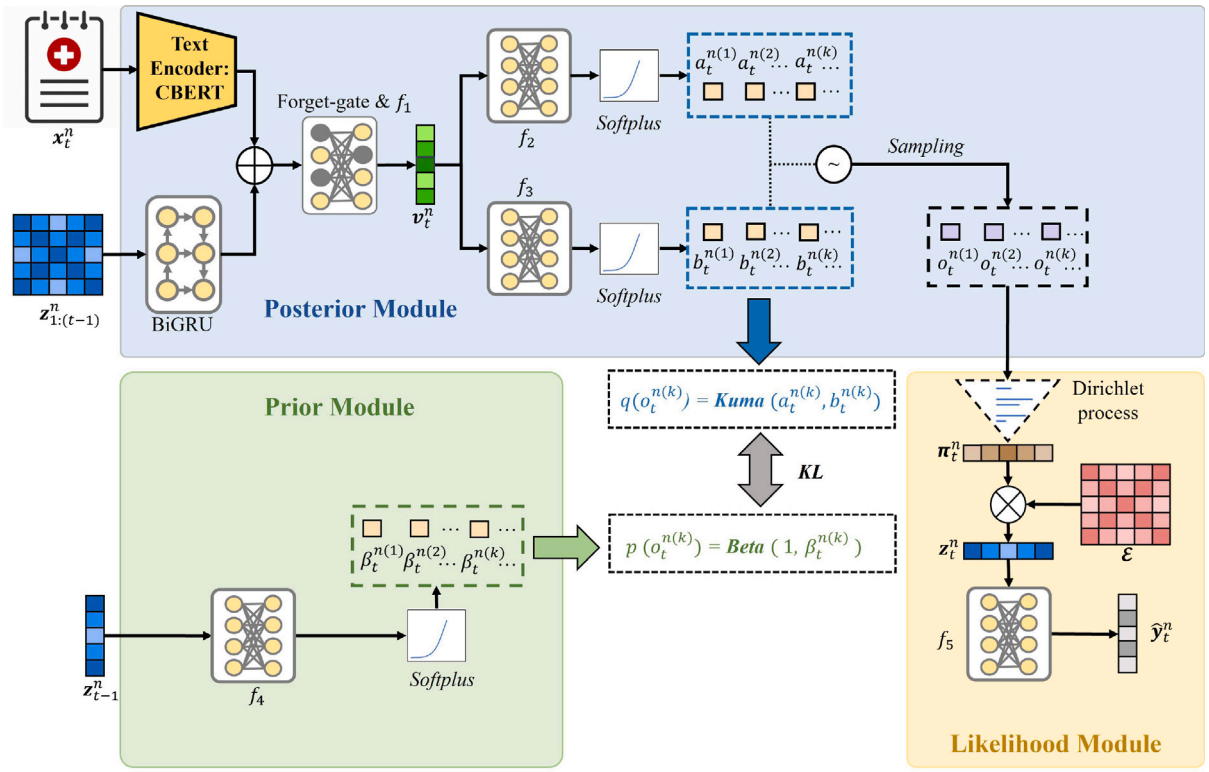


Fig. 2. The overview of the proposed DirPred model. Key variables are described as follows: $\boldsymbol{x}_t^n$ is the medical notes with auxiliary information of patient n at time t; $\boldsymbol{z}_t^n$ is the patient latent state at the tth hospital visit; $\boldsymbol{z}_{1:(t-1)}^n = [\boldsymbol{z}_1^n, \ldots, \boldsymbol{z}_{t-1}^n]$ contains all latent states from 1 to $t-1$ for patient n; $\boldsymbol{\pi}_t^n$ is the cluster assignment distribution generated from the stick-breaking construction process, whose kth element $\pi_t^{n(k)}$ ranges from 0 to 1 satisfying $\sum_k^{\infty} \pi_t^{n(k)} = 1$; $o_t^{n(k)}$ is the key variable to derive $\pi_t^{n(k)}$; $a_t^{n(k)}$ and $b_t^{n(k)}$ denote the parameters of the posterior distribution $q(o_t^{n(k)})$, and $\beta_t^{n(k)}$ is the parameter of prior distribution $p(o_t^{n(k)})$; $\boldsymbol{\mathcal{E}}$ is the embedding matrix of cluster centers; $\hat{\boldsymbol{y}}_t^n$ refers to the predicted risks.

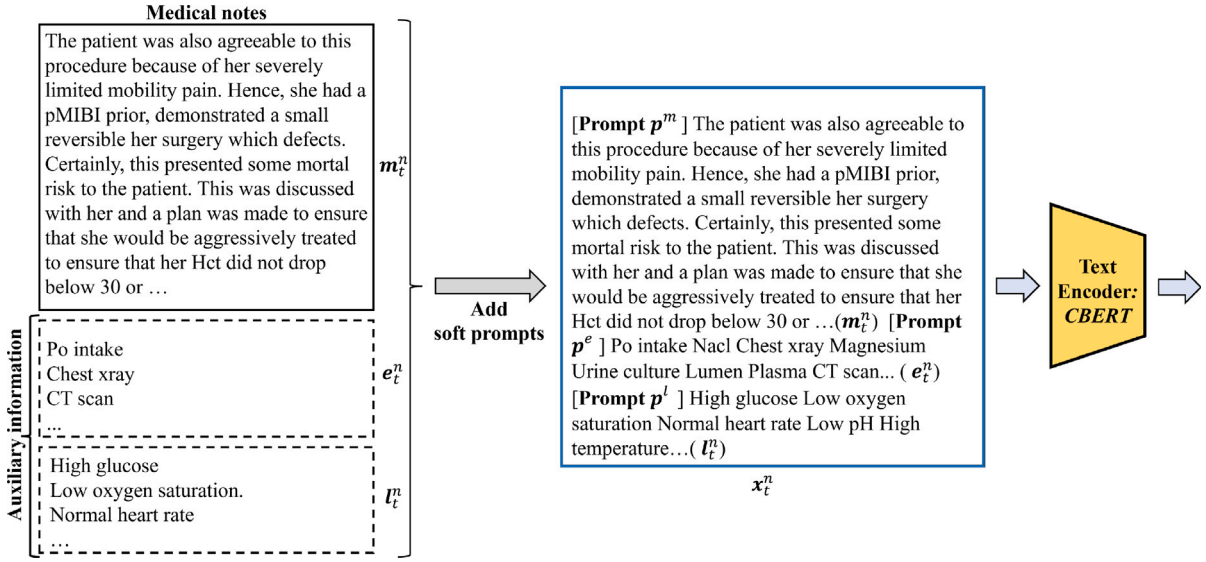


Fig. 3. The inputs of the text encoder $\boldsymbol{x}_t^n$ for patient n at time t in the posterior module. $\boldsymbol{m}_t^n$, $\boldsymbol{e}_t^n$ and $\boldsymbol{l}_t^n$ refers to the medical notes, clinical events, and descriptions of laboratory testing results. $\boldsymbol{p}^m$, $\boldsymbol{p}^e$, and $\boldsymbol{p}^l$ are their soft prompts shared across all inputs. *CBERT* denotes the language model Clinical-BERT.

For DNCP, we desire Kumaraswamy's closed-form inverse cumulative distribution function, where the samples can be drawn via the inverse transform [44]:

$$o_t^{n(k)} \sim \left(1 - u^{\frac{1}{b_t^{n(k)}}}\right)^{\frac{1}{a_t^{n(k)}}} \quad \text{where } u \sim \text{Uniform}(0, 1). \tag{5}$$

where the detail of parameters in Eq. (3), (4), and (5) will be elaborated in the following sections.

3.2. Our model

Fig. 2 provides an overview of our DirPred model. At the tth hospital visit of patient n, the current observation $\boldsymbol{x}_t^n$ encompasses unstructured health information primarily derived from medical notes, along with auxiliary details from clinical events and laboratory testing results. The patient latent state $\boldsymbol{z}_t^n$ is derived through predictive clustering utilizing the embedding matrix $\boldsymbol{\mathcal{E}}$ and the distribution of cluster assignment $\boldsymbol{\pi}_t^n$.

The prior distribution $p(o_t^{n(k)})$, a crucial variable in constructing the Dirichlet process for obtaining the cluster assignment $\boldsymbol{\pi}_t^n$, is generated based on the previous latent state $\boldsymbol{z}_{t-1}^n$. Conversely, the posterior distribution $q(o_t^{n(k)})$ is approximated by encoding the current observation $\boldsymbol{x}_t^n$ along with all previous latent states $\boldsymbol{z}_{1:(t-1)}^n$. Our model is trained by minimizing the Kullback–Leibler (KL) divergence between the prior and posterior distributions of $o_t^{n(k)}$, combining the risk prediction loss.

3.3. The posterior module

3.3.1. Encoding unstructured health information

In this subsection, we focus on describing the text encoder of the posterior module. Recent works in [2,22] have shown promising results by integrating medical notes with other auxiliary information. However, they still cannot resolve the difference among data from different modalities. To address this issue, we adopt a soft Prompt learning-based data encoding method.

The example input shown in Fig. 3 contains raw data from medical notes $\boldsymbol{m}_t^n$ and also auxiliary information including the descriptions of clinical events $\boldsymbol{e}_t^n$ and laboratory testing results in $\boldsymbol{l}_t^n$. The descriptions of clinical events $\boldsymbol{e}_t^n$ are obtained by concatenating all clinical events into a sequence. For laboratory testing results, we apply the boxplot anomaly detection method [45] to find all abnormal information and convert it into textual descriptions $\boldsymbol{l}_t^n$. $\boldsymbol{x}_t^n$ collects all this information as $\boldsymbol{x}_t^n = \{\boldsymbol{p}^m, \boldsymbol{m}_t^n, \boldsymbol{p}^e, \boldsymbol{e}_t^n, \boldsymbol{p}^l, \boldsymbol{l}_t^n\}$, where $\boldsymbol{p}^m$, $\boldsymbol{p}^e$, and $\boldsymbol{p}^l$ are soft prompts. These soft prompts, which are modality-specific, are tokens that are shared across all samples. The embeddings of the prompts are learnable in the training process with the purpose of mitigating differences from different modalities.

3.3.2. Neural Dirichlet process

Fig. 2 also illustrates the structure of the posterior module related to the neural Dirichlet process. Suppose $\boldsymbol{z}_{1:(t-1)}^n = [\boldsymbol{z}_1^n, \ldots, \boldsymbol{z}_{(t-1)}^n]$ contains all previous latent states of the patient n, which is encoded together with $\boldsymbol{x}_t^n$ to get a fused embedding vector $\boldsymbol{v}_t^n \in \mathbb{R}^D$:

$$\boldsymbol{v}_t^n = f_1(g(CBERT(\boldsymbol{x}_t^n)) \oplus BiGRU(\boldsymbol{z}_{1:(t-1)}^n)), \tag{6}$$

where $CBERT(.)$ denotes the Clinical-BERT encoder [23], $BiGRU(.)$ is the bidirectional gate recurrent unit(GRU), $f_1(.)$ is a fully connected network, $g(.)$ is the forget gate adopted from the long short-term memory (LSTM) [46] and $\oplus$ is the concatenation operator. $\boldsymbol{v}_t^n$ is then fed into two parallel fully connected networks $f_2(.)$ and $f_3(.)$ with the $softplus$ activation function: the outputs are the parameters of the posterior distribution $q(o_t^{n(k)}) = Kuma(a_t^{n(k)}, b_t^{n(k)})$, which are:

$$[a_t^{n(1)}, \ldots, a_t^{n(k)}, \ldots, a_t^{n(K)}] = softplus(f_2(\boldsymbol{v}_t^n)) = log(1 + exp(f_2(\boldsymbol{v}_t^n))) \tag{7}$$

and

$$[b_t^{n(1)}, \ldots, b_t^{n(k)}, \ldots, b_t^{n(K)}] = softplus(f_3(\boldsymbol{v}_t^n)) = log(1 + exp(f_3(\boldsymbol{v}_t^n))), \tag{8}$$

where the softplus activation function is to make sure the parameters $a_t^{n(k)}$ and $b_t^{n(k)}$ are positive. With $a_t^{n(k)}$ and $b_t^{n(k)}$, for each k then $o_t^{n(k)}$ can be sampled from Eq. (5) to generate the cluster assignment distribution $\boldsymbol{\pi}_t^n$ via Eq. (2). Please note that K here does not refer to the exact number of clusters. Instead of implying a finite-dimensional prior, K here is the truncation parameter. $o_t^{n(K)}$ is always set to one to ensure $\sum_k^K \pi_t^{n(k)} = 1$, where $\pi_t^{n(K)}$ represents the total probability of K to ∞ clusters.

3.4. The prior module

As described in [44], the prior distribution of $o_t^{n(k)}$ is assumed to follow the Beta distribution, that is $p(o_t^{n(k)}) = Beta(1, \beta_t^{n(k)})$. Its parameter $\beta_t^{n(k)}$ is obtained by encoding the latent states at the previous time step:

$$[\beta_t^{n(1)}, \ldots, \beta_t^{n(k)}, \ldots, \beta_t^{n(K)}] = softplus(f_4(\boldsymbol{z}_{t-1}^n)), \tag{9}$$

where $f_4(.)$ refers to a fully connected network.

3.5. The likelihood module

In the likelihood module, the embedding matrix of cluster centers $\boldsymbol{\mathcal{E}}$ together with the cluster distribution vector $\boldsymbol{\pi}_t^n$ are used to generate the patient latent state $\boldsymbol{z}_t^n$ via:

$$\boldsymbol{z}_t^n = \boldsymbol{\mathcal{E}}^T \boldsymbol{\pi}_t^n, \tag{10}$$

where $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{K \times D}$, K is the truncation parameter, $\boldsymbol{z}_t^n \in \mathbb{R}^D$, and $(.)^T$ is the transpose operator. $\boldsymbol{z}_t^n$ is then fed into the fully connected network $f_5(.)$ to get the risk prediction results as:

$$\hat{\boldsymbol{y}}_t^n = f_5(\boldsymbol{z}_t^n). \tag{11}$$

3.6. Learning objective

With the adoption of the variational Bayesian inference, the loss is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\left\{ \mathbb{E}_{q_\phi(o_1^n|x_1^n)}[log p_\theta(\boldsymbol{y}_1^n|\boldsymbol{o}_1^n)] + KL(q_\phi(\boldsymbol{o}_1^n|\boldsymbol{x}_1^n) \parallel p_\theta(\boldsymbol{o}_1^n)) - \sum_{t=2}^{T_n} \mathbb{E}_{q_\phi(o_t^n|\boldsymbol{z}_{1:(t-1)}^n, \boldsymbol{x}_t^n)}[log p_\theta(\boldsymbol{y}_t^n|\boldsymbol{o}_t^n)] + \sum_{t=2}^{T_n} KL(q_\phi(\boldsymbol{o}_t^n|\boldsymbol{z}_{1:(t-1)}^n, \boldsymbol{x}_t^n) \parallel p_\theta(\boldsymbol{o}_t^n)) \right\}, \tag{12}$$

where N is the total number of patients, and T_n is the number of hospital visits for patient n. The likelihood $p_\theta(\boldsymbol{y}_t^n|\boldsymbol{o}_t^n)$ is calculated as:

$$p_\theta(\boldsymbol{y}_t^n|\boldsymbol{o}_t^n) = \frac{1}{L}\sum_{i}^{L} y_{t,i}^n \cdot \log(\hat{y}_{t,i}^n) + (1 - y_{t,i}^n) \cdot \log(1 - \hat{y}_{t,i}^n), \tag{13}$$

where L is the number of classes (i.e. disease risks). The posterior and prior of $\boldsymbol{o}_t^n$ are $q_\phi(\boldsymbol{o}_t^n|\boldsymbol{z}_{1:(t-1)}^n, \boldsymbol{x}_t^n) = Kuma(a_t^{n(k)}, b_t^{n(k)})$, and $p_\theta(\boldsymbol{o}_t^n) = Beta(1, \beta_t^{n(k)})$, respectively. ϕ and θ represent the parameters of neural networks for the distribution approximation. The KL divergence between the prior and posterior distributions of $o_t^{n(k)}$ can be represented as [44]:

$$\begin{aligned}\mathbb{E}_{q(o_t^{n(k)})}\left[log\frac{q(o_t^{n(k)})}{p(o_t^{n(k)})}\right] &= \frac{a_t^{n(k)} - 1}{a_t^{n(k)}}\left(-\gamma - \Psi(b_t^{n(k)}) - \frac{1}{b_t^{n(k)}}\right) \\ &+ log a_t^{n(k)} b_t^{n(k)} + log\ B(1, \beta_t^{n(k)}) \\ &- \frac{b_t^{n(k)} - 1}{b_t^{n(k)}} + (\beta_t^{n(k)} - 1) b_t^{n(k)} \sum_{m=1}^{\infty} \frac{1}{m + a_t^{n(k)} b_t^{n(k)}} B\left(\frac{m}{a_t^{n(k)}}, b_t^{n(k)}\right),\end{aligned} \tag{14}$$

where γ is Euler's constant, $\Psi(.)$ is the Digamma function, $B(.)$ is the Beta function and the infinite sum results from the Taylor expansion.

The training procedure to optimize DirPred by minimizing the loss defined in Eq. (12) is shown in Algorithm 1.

4. Experiments

4.1. Dataset

As summarized in Table 1, our model and all comparative baselines are trained and evaluated on two publicly accessible de-identified medical datasets, MIMIC-III[1] and N2C2-2014.[2]

[1] https://physionet.org/content/mimiciii/1.4/

[2] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

Algorithm 1 The DirPred model

Input Given the EHR data $\boldsymbol{x}_t^n$ for $t \in \{1, \ldots, T_n\}$ and $n \in \{1, \ldots, N\}$, where $\boldsymbol{x}_t^n$ refers to the patient's medical notes with auxiliary information.
while not converge **do**
  for Each patient n **do**
    for Each time t **do**
      Encode $\boldsymbol{x}_t^n$ using the Clinical-BERT.
      Encode the information from previous latent states $\boldsymbol{z}_{1:(t-1)}^n$ using the bidirectional GRU.
      Combine the outputs from the above two steps using Eq. (6) and then generate $a_t^{n(k)}$ and $b_t^{n(k)}$ for all k using Eq. (7) and Eq. (8).
      Sample $o_t^{n(k)}$ using $a_t^{n(k)}$ and $b_t^{n(k)}$ for all k as defined in Eq. (5).
      Get the cluster assignment distribution $\boldsymbol{\pi}_t^n$ using $\{o_t^{n(k)}\}_{k=1}^K$ from Eq. (2).
      Generate the patient latent state $\boldsymbol{z}_t^n$ using $\boldsymbol{\pi}_t^n$ and the embedding matrix of cluster centers $\boldsymbol{\mathcal{E}}$ from Eq. (10).
      Obtain the prediction results using $\boldsymbol{z}_t^n$ from Eq. (11).
      Calculate the cross-entropy loss of the risk prediction results.
      Derive the parameter $\beta_t^{n(k)}$ of the prior distribution of $o_t^{n(k)}$ for all k using Eq. (9).
      Calculate the KL divergence between the prior and the posterior of $o_t^{n(k)}$ as described in Eq. (14).
    end for
  end for
  Update parameters by minimizing the loss defined in Eq. (12) for patients in each batch.
end while

Table 1
The summary of EHR datasets.

Dataset	MIMIC-III	N2C2-2014
# Records	22,220	1,304
# Patients	19,017	296
# Patients with Multiple Visits	3,740	296
Avg. # Visits	2.61	4.41
# Records from Patients with Multiple Visits	9,759	1,304
Data Examples	**Text**: "This year old woman has a history of COPD. Over the past five years she has had progressive difficulties with her breathing. She was admitted to hospital for respiratory failure due to a COPD exacerbation. Due to persistent hypoxemia, she required intubation and a eventual bronchoscopy on revealed marked narrowing of the airways on expiration consistent with tracheomalacia. She ..." **Event**: "dextrose gauge furosemide lasix po intake nacl magnesium sulfate chest xray..." **Lab**:"high glucose high fraction inspired oxygen high glucose normal heart rate low oxygen saturation..."	**Text**: "The patient is year old male with complaints of chest pain and throat tightness. The patient reported that he was stuck in traffic for about hours last night and apparently got very tense. He felt some heat from his car thought that it was overheating and then developed some chest pain and throat tightness. He really described what seems to be fleeting chest tightness and no diaphoresis no shortness of breath and no arm ..."

4.1.1. The MIMIC-III dataset

MIMIC-III [27] is one of the largest publicly accessible EHR datasets, comprising both structured and unstructured health information collected during multiple hospital visits of patients. It contains 22,220 records of patient visit information that encompass both medical notes (hospital brief course) and auxiliary information collected from 19,017 distinct individuals across various single or multiple hospital visits. We extracted hospital visit records and risk indicators from the MIMIC-III dataset using the data extraction method described in [47]. Given that our model encodes patients' longitudinal information, we extracted a subset of the MIMIC-III dataset that comprises records from 3740 patients who had two or more hospital visits, totalling 9759 records. The average number of visits for patients in this subset is 2.61.

Clinical events and laboratory testing results serve as auxiliary information that is fed into the predictive model alongside medical notes. Specifically, clinical events are transformed into a sequence of non-repeating phrases, while the boxplot anomaly detection method [45] is applied to laboratory data to obtain text descriptions of anomalies. Examples of the extracted data are presented in Table 1. We pre-process the three input textual data by removing numbers, noise, and stop words. To evaluate the performance of our proposed model in the disease risk prediction task, we use three upper-level categories of risk indicators, Chronic disease risk, Acute disease risk, and Mixed disease risk as our prediction targets [47]. We adopt the same data-splitting strategy as in [47], whereby we obtain training and test datasets at a 4:1 ratio for performance evaluation.

4.1.2. The N2C2-2014 dataset

Different from MIMIC-III, the N2C2-2014 dataset does not include such abundant auxiliary information. Instead, it consists of 1304 medical notes from 296 patients who have undergone two or more hospital visits, with an average visit count of 4.41. We applied the same data pre-processing strategy as used for the MIMIC-III dataset, which involved removing numbers, noise, and stop words. Our prediction target was four disease risk-related factors: Hyperlipidemia, Hypertension, Coronary artery disease, and Diabetes. We split the training and test datasets at a 4:1 ratio for performance evaluation.

4.2. Baseline methods

We compared DirPred with other baseline methods, which fall into three categories: Class 1 methods are purely supervised models for risk prediction; Class 2 methods adopt clustering approaches to assist the supervised prediction tasks; Class 3 methods are ablated versions of DirPred.

The Class 1 baseline methods are listed as follows:

- **SVM** and **XGBOOST**: These are two conventional machine learning methods which use the word2vec representations of inputs for predictive model construction.
- **CAML**: CAML [9] is a state-of-the-art interpretable medical textual classification model. It integrates label-embedding and cross-attention mechanisms to provide an interpretable medical text classification model.
- **CAML$^+$**: The encoding layer of CAML is replaced with Clinical-BERT [23] for a fair comparison with our model.
- **CAML^{++}**: To encode the longitudinal information from EHRs, the time-aware attention mechanism from ConCare [35] is integrated into CAML$^+$.
- **RETAIN**: RETAIN [10] extracts information from longitudinal EHRs in reverse time order by using two RNNs and self-attention mechanisms.

- **RETAIN$^+$**: For a fair comparison, the language model Clinical-BERT is introduced to RETAIN for text embedding.
- **DIPOLE**: DIPOLE [20] replaces the two RNNs layers in RETAIN with Bi-directional RNNs [48]. It also adopts an attention mechanism to help integrate information from both past and future hospital visits.
- **DIPOLE$^+$**: The encoding layer of DIPOLE is replaced with Clinical-BERT for text embedding.

The Class 2 baseline methods are listed as follows:

- **Deep K-means**: K-means is a conventional unsupervised clustering method. To be capable of handling unstructured data and performing risk prediction, we adopt a deep neural network version of K-means by utilizing the Clinical-BERT [23] and fully connected networks. All data will be first clustered into groups via K-means to generate center embeddings, based on which a predictive model will be built to predict disease risk.
- **CAMLOT**: ACTPC [15] and CAMELOT [16] are two recent predictive clustering models for disease risk prediction. While CAMELOT has demonstrated improved performance and training strategies, it was not designed to handle unstructured data obtained from multiple hospital visits; instead, it focused on modelling numerical time-series health monitoring signals. In light of this observation, we have modified CAMELOT by replacing its RNNs-based encoding module with Clinical-BERT, thereby enabling it to handle unstructured data.

The ablated versions of our model in Class 3 are listed as follows:

- **DirPred-I**: DirPred-I uses only medical notes as input data without auxiliary information. We replaced the non-parametric clustering approach with a parametric clustering approach [49]. For MIMIC-III and N2C2-2014, the number of clusters is set to 8 and 16, respectively. We followed the strategy adopted in [15], where the number of clusters is set to 2^L, with L representing the number of disease risks.
- **DirPred-II**: DirPred-II takes the same data input as DirPred-I, while the clustering part follows the non-parametric approach proposed in DirPred.
- **DirPred-III**: DirPred-III removes the prior module from DirPred to check the impacts of modelling longitudinal information on the performance of risk prediction.

For all comparative models, the learning rate is set to $1e^{-5}$, the token length is 300, the embedding size of Clinical-BERT is 768, and the size of the latent state is 384. The ADAM optimizer is chosen as the optimizer for model training. We also adopt the dropout strategy with a dropout rate of 0.3 and the gradient clip strategy with a clip value of 125. All comparative models are trained five times with a fixed set of five different seeds, and the results were presented in terms of average indicator performance. All models are implemented using PyTorch on an NVIDIA TESLA V100 GPU. The source code of our model is publicly accessible.[3]

4.3. Evaluation metrics

The performance of risk prediction is assessed using metrics such as precision, recall, F1 score, accuracy, and AUROC score.

Precision is the ratio of positive predictions which are correctly identified, which is expressed as,

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

Recall is the ratio of all true positives are correctly identified, which is expressed as,

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

F1 score is the harmonic mean of precision and recall, which is expressed as,

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{17}$$

Accuracy is the ratio of correct predictions, which is expressed as,

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

TP, TN, FP, and FN represent True Positives, True Negative, False Positives, and False Negative. True Positives are for accurate positive class predictions, True Negatives are for accurate negative class predictions, False Positives are for inaccurate positive class predictions, and False Negatives are for inaccurate negative class predictions.

Area Under the Receiver Operating Characteristic Curve (AUROC) score is a classification performance measurement that summarizes a value of the Receiver Operating Characteristic Curve.

In addition, to provide comprehensive evaluation results, we compute all metrics from both micro-average and macro-average. The difference between micro-average and macro-average is that the former takes into account the total outcome contributions from all classes to calculate the average, while the latter calculates the average for each class independently and then aggregates.

4.4. Evaluation results

4.4.1. Comparison with Class 1 models

From Table 2 and Table 3, our DirPred model outperforms all Class 1 methods with higher F1, AUROC, and ACC scores in the risk prediction task for both datasets. We can also observe that compared with conventional machine learning approaches such as SVM and XGBOOST, the neural network-based models, especially CAML$^+$, exhibit stronger predictive power. Although SVM and XGBOOST achieve high recall values (close to 1.0), their precision scores are much lower than the others. Therefore, when considering the overall performance represented by F1 and AUROC scores, these models are found to be not as good as the CAML$^+$ model.

Furthermore, we observe that longitudinal models, such as CAML^{++}, RETAIN, and DIPOLE, have better performance than SVM, XGBOOST, and CAML. Among these models, DIPOLE$^+$ has the best F1, AUROC, and ACC scores. This observation indicates that historical information generated from previous hospital visits can improve the performance of disease risk prediction. Moreover, RNNs are useful in extracting longitudinal information, as evidenced by the evaluation performance of the RETAIN and DIPOLE models.

Additionally, we find that the performance can be enhanced by integrating Clinical-BERT [23] into predictive models. Specifically, the F1, AUROC, and ACC scores from CAML, RETAIN, and DIPOLE can be increased when the pre-trained language model is added. The improvement in the N2C2-2014 dataset is particularly significant. For example, the ACC score of N2C2-2014 increased from 0.2957 to 0.5923 by introducing Clinical-BERT to DIPOLE.

4.4.2. Comparison with Class 2 models

Table 2 and Table 3 show that Deep K-means and CAMELOT models have similar performance on the MIMIC-III dataset, while CAMELOT outperforms Deep K-means on the N2C2-2014 dataset with higher F1, AUROC, and ACC scores. However, both models fail to show better performance than the Class 1 methods. This is because Class 2 methods are not designed for modelling unstructured data, and providing cluster-level evidence can sacrifice predictive performance. It is worth noting that our DirPred model, which is also a predictive clustering-based approach, significantly outperforms CAMELOT and Deep K-means models.

[3] https://github.com/Healthcare-Data-Mining-Laboratory/DirPred.git

Table 2
The risk prediction results for the MIMIC-III dataset. The best results are highlighted in bold, and the second-best results are marked with an asterisk (*).

Models	MIMIC-III								
	Micro				Macro				ACC
	Precision	Recall	F1	AUROC	Precision	Recall	F1	AUROC	
SVM	0.7735	0.9998	0.8722	0.5004	0.7735	0.9998	0.8703	0.5003	0.5111
XGBOOST	0.7736	0.9999	0.8723	0.5005	0.7736	0.9999	0.8705	0.5003	0.5106
CAML	0.7849	0.9718	0.8777	0.6702	0.7809	0.9680	0.8634	0.6270	0.4939
CAML^{+}	0.8467	0.9294	0.8860	0.8350	0.8464	0.9229	0.8796	0.8165	0.5363
CAML^{++}	0.8445	0.9512	0.8895	0.8428	0.8385	0.9458	0.8853	0.8187	0.5471
RETAIN	0.8473	0.9206	0.8824	0.8131	0.8452	0.9176	0.8799	0.7956	0.5476
RETAIN^{+}	0.8336	0.9499	0.8879	0.8349	0.8326	0.9457	0.8848	0.8289	0.5556
DIPOLE	0.8401	0.9323	0.8904	0.8317	0.8384	0.9445	0.8879	0.8226	0.5397
DIPOLE^{+}	0.8581	0.9355	0.8948	0.8470	0.8572	0.9335	0.8929	0.8283	0.5633
Deep K-means	0.7735	0.9999	0.8624	0.6055	0.7735	0.9998	0.8601	0.5001	0.5103
CAMELOT	0.7803	0.9650	0.8618	0.5070	0.7803	0.9652	0.8600	0.5080	0.5108
DirPred-I	0.8272	0.9797	0.8971	0.8532	0.8262	0.9768	0.8952	0.8494	0.5651
DirPred-II	0.8778	0.9223	0.8994*	0.8692*	0.8773	0.9192	0.8972*	0.8576*	0.5980*
DirPred-III	0.8706	0.9223	0.8956	0.8514	0.8690	0.9194	0.8933	0.8470	0.5819
DirPred	0.8722	0.9340	**0.9022**	**0.8778**	0.8714	0.9307	**0.8997**	**0.8625**	**0.6041**

Table 3
The risk prediction results for the N2C2-2014 dataset. The best results are highlighted in bold, and the second-best results are marked with an asterisk (*).

Models	N2C2-2014								
	Micro				Macro				ACC
	Precision	Recall	F1	AUROC	Precision	Recall	F1	AUROC	
SVM	0.6068	0.9972	0.7539	0.5074	0.6063	0.9973	0.7441	0.5073	0.1869
XGBOOST	0.6062	0.9948	0.7540	0.5080	0.6070	0.9948	0.7437	0.5076	0.1826
CAML	0.6846	0.8502	0.7565	0.7333	0.6525	0.8142	0.7078	0.6572	0.2348
CAML^{+}	0.8908	0.9136	0.9007	0.9428	0.8893	0.9022	0.8949	0.9462	0.5627
CAML^{++}	0.8376	0.9275	0.8903	0.9239	0.8543	0.9175	0.8877	0.9221	0.5122
RETAIN	0.7556	0.8466	0.7949	0.8194	0.7392	0.8162	0.7716	0.7949	0.3126
RETAIN^{+}	0.8859	0.9118	0.8994	0.9204	0.8902	0.9023	0.8973	0.9082	0.5347
DIPOLE	0.7669	0.8341	0.7990	0.8253	0.7546	0.8000	0.7715	0.8118	0.2957
DIPOLE^{+}	0.8961	0.9201	0.9036	0.9376	0.8975	0.9128	0.8977	0.9226	0.5923
Deep K-means	0.6781	0.8272	0.7452	0.6774	0.5074	0.7453	0.6003	0.5187	0.1308
CAMELOT	0.6057	0.9999	0.7544	0.6803	0.6057	0.9999	0.7439	0.5754	0.1905
DirPred-I	0.9175	0.9344	0.9273*	0.9382*	0.9104	0.9275	0.9185*	0.9296*	0.6459
DirPred-III	0.8972	0.9306	0.9136	0.9346	0.8971	0.9209	0.9077	0.9284	0.6467*
DirPred / DirPred-II	0.9270	0.9379	**0.9323**	**0.9592**	0.9209	0.9360	**0.9278**	**0.9638**	**0.7050**

★ Please note that DirPred-II for the N2C2-2014 dataset is equivalent to DirPred because only medical notes data from the N2C2-2014 dataset are used.

This observation implies that our model is the state-of-the-art predictive clustering method in the disease risk prediction task. The improvements are attributed to the development of the neural Dirichlet process model.

4.4.3. Comparison with Class 3 models

When comparing the non-parametric clustering-based approaches (i.e. DirPred-II and DirPred) with the parametric clustering approaches (i.e. DirPred-I), the results from Table 2 and Table 3 suggest that the former outperforms the latter on both datasets. This observation indicates the superior performance of the non-parametric clustering method for the disease prediction task. For MIMIC-III, which contains additional clinical events and laboratory results, we further investigate the impact of including them in our model. By comparing DirPred-II (the ablated version of DirPred without using additional information) with DirPred, we find that incorporating auxiliary information can improve performance. We also demonstrate the effectiveness of the prior module by introducing the ablated version, DirPred-III. By comparing DirPred-III with DirPred, we can see that the predictive performance is significantly reduced without the prior module, especially for the N2C2-2014 dataset.

4.5. Broadening model scope: Capturing intra-visit variations

To further explore the capabilities of our DirPred model with varying medical data inputs, we have undertaken an experiment specifically focusing on its proficiency in capturing intra-visit patients' latent health state variations related to acute disease risk, such as sepsis. We undertake a comparative analysis between our model, DirPred, and the baseline model DIPOLE (it has demonstrated the highest performance evaluation scores across all baseline models) on the MIMIC-III dataset (while the N2C2-2014 dataset does not include time-series input features). To demonstrate the ability of our model to capture intra-visit variations, we have implemented a transition from utilizing textual inputs for each patient hospital visit to employing continuous time-series laboratory testing results within each visit. Furthermore, we have substituted the text encoder with an RNNs-based network—GRU. The results presented in Table 4 show that our DirPred model attains the highest evaluation scores across all metrics, thereby substantiating its superior capability in capturing intra-visit variations.

5. Discussion

5.1. Local-level explainability

Our proposed model can assist clinicians and patients in extracting valuable features from EHR data by providing local-level evidence through an attention mechanism. Fig. 4 displays the local-level explainable results of the attention mechanism. We randomly selected three patient cases from the test dataset to demonstrate the model's performance, where the highlighted input features are the ones that gained high attention scores in the attention mechanism.

For example, patient case 1 is at risk of "acute cerebrovascular disease". Our model highlights words or phrases from medical notes,

Table 4
Evaluation performance on an acute disease risk (sepsis) with continuous time-series laboratory testing results as the input. The best results are highlighted in bold.

Models	MIMIC-III								
	Micro				Macro				ACC
	Precision	Recall	F1	AUROC	Precision	Recall	F1	AUROC	
DIPOLE	0.8220	0.8220	0.8220	0.6955	0.5366	0.5005	0.4543	0.6955	0.8220
DirPred	0.8261	0.8261	**0.8261**	**0.7130**	0.9127	0.5080	**0.4680**	**0.7130**	**0.8261**

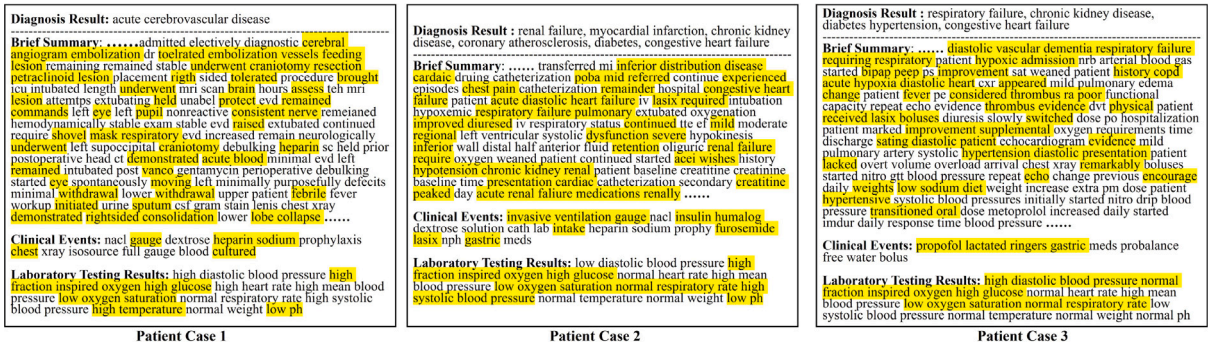


Fig. 4. Local-level explainable results were obtained from the attention mechanism for three randomly selected patient cases. The highlighted input features gained high attention scores (top 20%) in the attention mechanism and provided important information for disease risk prediction.

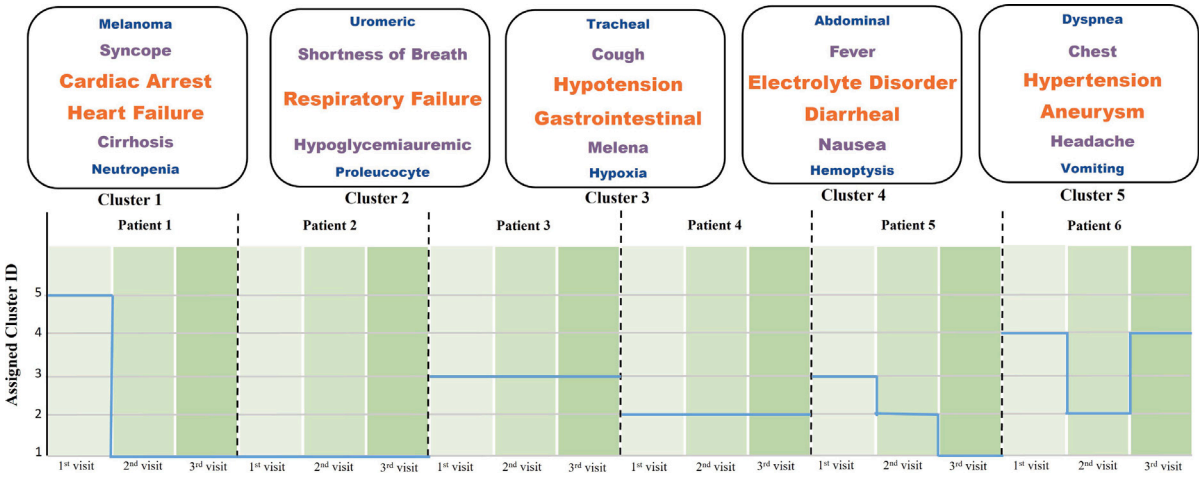
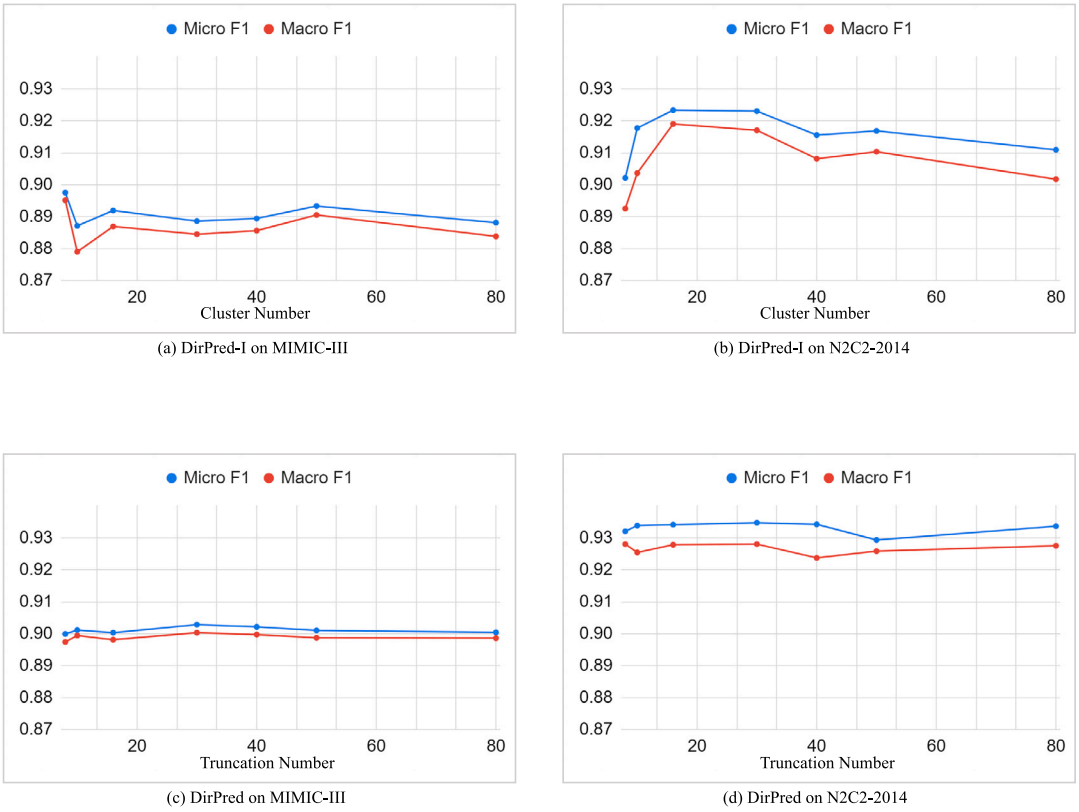


Fig. 5. The cluster assignments and medical terminologies of six randomly selected patients from their first three hospital visits.

such as "cerebral angiogram embolization", "underwent craniotomy resection petra clinoid lesion", and "craniotomy", which are explicitly semantically related to diagnosing diseases. For clinical events, "heparin sodium" is a medication that is related to cerebrovascular disease treatment [50]. Regarding laboratory testing results, parameters such as "fraction inspired oxygen", "glucose", "oxygen saturation", "temperature", and "pH" are commonly measured in patients with acute cerebrovascular disease [51]. Similar findings were also observed in the other two patient cases.

5.2. Cluster-level explainability

We present the results of clustering patients' latent health states using our predictive clustering-based model, which not only predicts disease risks but also groups latent states into different clusters to provide cluster-level explainable evidence. Fig. 5 displays the assignments of six patients randomly selected from the test dataset. By examining the cluster assignments indicated by Cluster ID, we observe that the trajectories of patients' latent health states (i.e., cluster assignments at different time points) change across different hospital visits. Specifically, patient 1 first stays in cluster 5 and then remains in cluster 1 for the remaining two visits; patients 2, 3, and 4 are assigned to clusters 1, 3, and 2, respectively, for all three hospital visits; patient 5 is assigned to clusters 3, 2, and 1 in three consecutive visits; patient 6 is assigned to cluster 4 at the 1st and 3rd hospital visits and cluster 2 at the 2nd hospital visit.



(a) DirPred-I on MIMIC-III (b) DirPred-I on N2C2-2014

(c) DirPred on MIMIC-III (d) DirPred on N2C2-2014

Fig. 6. The performance of disease risk prediction under different settings.

The detected clusters by predictive clustering can be used to explain patients' health states by analysing each cluster's common characteristics. Fig. 5 shows the most frequent medical terminologies associated with EHRs in each cluster. We find that most terminologies in each cluster fall under one broad disease category. For example, cluster 1 is related to cardiovascular diseases, where "cardiac arrest" is a typical symptom of cardiovascular disease [52], "cirrhosis" may be caused by heart disease [53], "neutropenia" could aggravate acute cardiovascular diseases [54], and "melanoma" demonstrated that cardiac metastases occurred in up to 65% of the cases [55]. Cluster 2, is related to respiratory system diseases, and sometimes severe "hypoglycemia" can lead to respiratory failure [56]. Similar findings are also observed for the other three clusters. By understanding these clusters and analysing the changes in cluster assignments for each patient over time, we can gain insights into the changes in their health states, which in turn can support the changes in the risk prediction results.

5.3. Sensitivity analysis

We have conducted a sensitivity analysis to investigate how the number of clusters, denoted by $\mathcal{K}$, and the truncation parameter, denoted by K, would affect the performance of DirPred-I and DirPred, respectively. Fig. 6 illustrates the F1 scores obtained by varying the values of $\mathcal{K}$ and K within the range of [8, 10, 16, 30, 40, 50, 80]. It can be observed that the F1 curves of DirPred-I display more significant fluctuations than those of DirPred across different settings. For DirPred-I, the optimal number of clusters leading to the highest F1 scores for the MIMIC-III and N2C2-2014 datasets are 8 and 16, respectively. These values align with the recommendations proposed by [15] that we have adopted in the previous experiments. In contrast, for DirPred, the F1 values remain relatively stable when changing the truncation parameter. This observation suggests that the performance of DirPred is less sensitive to the choice of truncation parameter, which is a notable advantage of non-parametric clustering methods.

6. Conclusions

In this paper, we have proposed a novel explainable AI model, the non-parametric predictive clustering, for disease risk prediction in healthcare decision-making. Our model utilizes longitudinal medical notes along with auxiliary information and provides multi-level explainable evidence simultaneously. The predictive clustering approach groups latent health states and uses the weighted representation of cluster centers for risk prediction without pre-defining the exact number of clusters. To accomplish this, we adopt the Dirichlet process mixture model as a non-parametric clustering approach. To effectively couple non-parametric processes with neural networks, the model is trained using the stochastic gradient descent variational Bayesian inference method. The posterior of the parameters in the non-parametric clustering algorithm is approximated using both the current and historical information. To encode heterogeneous information from multiple modalities of EHRs, we adopt the soft Prompt learning approach for data fusion. In order to capture temporal dependencies and construct a dynamic model, a prior network is specifically engineered to furnish prior parameters derived from the previous latent state. In our experiments, our model demonstrates superior predictive performance over state-of-the-art comparative models on two popular real-world EHR datasets. Additionally, it provides local-level and cluster-level explainable evidence to identify valuable information contained in EHR data and interpret patients' latent health states.

CRediT authorship contribution statement

Shuai Niu: Validation, Software, Resources, Methodology, Data curation, Conceptualization, Visualization, Writing – original draft, Writing – review & editing. **Qing Yin:** Conceptualization, Methodology, Writing – review & editing. **Jing Ma:** Supervision, Writing – review & editing. **Yunya Song:** Writing – review & editing. **Yida Xu:** Conceptualization, Methodology, Writing – review & editing. **Liang**

Bai: Methodology, Validation. **Wei Pan:** Data curation, Investigation, Resources, Writing – review & editing. **Xian Yang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Conceptualization, Investigation, Methodology.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Liang Bai reports financial support was provided by National Science and Technology Major Project. Liang Bai reports financial support was provided by National Natural Science Foundation of China. Liang Bai reports financial support was provided by Fundamental Research Program of Shanxi Province. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The link of data and code is announced in the manuscript.

Acknowledgements

This work is supported by the National Science and Technology Major Project (No. 2021ZD0113303), the National Natural Science Foundation of China (No. 62276159), and the Fundamental Research Program of Shanxi Province (No. 202303021223004).

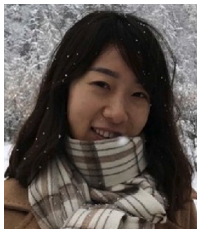
References

[1] Y. Cheng, F. Wang, P. Zhang, J. Hu, Risk prediction with electronic health records: A deep learning approach, in: Proceedings of the 2016 SIAM International Conference on Data Mining, SIAM, 2016, pp. 432–440.
[2] 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.
[3] S. Schallmoser, T. Zueger, M. Kraus, M. Saar-Tsechansky, C. Stettler, S. Feuerriegel, Machine learning for predicting micro-and macrovascular complications in individuals with prediabetes or diabetes: Retrospective cohort study, J. Med. Internet Res. 25 (2023) e42181.
[4] T. Zueger, S. Schallmoser, M. Kraus, M. Saar-Tsechansky, S. Feuerriegel, C. Stettler, Machine learning for predicting the risk of transition from prediabetes to diabetes, Diabetes Technol. Ther. 24 (11) (2022) 842–847.
[5] W.-Y. Hsu, A decision-making mechanism for assessing risk factor significance in cardiovascular diseases, Decis. Support Syst. 115 (2018) 64–77.
[6] 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.
[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] K. Topuz, F.D. Zengul, A. Dag, A. Almehmi, M.B. Yildirim, Predicting graft survival among kidney transplant recipients: A Bayesian decision support model, Decis. Support Syst. 106 (2018) 97–109.
[9] 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.
[10] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, Adv. Neural Inf. Process. Syst. 29 (2016).
[11] J. Shang, C. Xiao, T. Ma, H. Li, J. Sun, Gamenet: Graph augmented memory networks for recommending medication combination, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, 2019, pp. 1126–1133.
[12] N. Razavian, D. Sontag, Temporal convolutional neural networks for diagnosis from lab tests, 2015.
[13] Z. Che, Y. Cheng, Z. Sun, Y. Liu, Exploiting convolutional neural network for risk prediction with medical feature embedding, in: NIPS Workshop on Machine Learning for Health, 2017.
[14] R. Krishnan, U. Shalit, D. Sontag, Structured inference networks for nonlinear state space models, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, No. 1, 2017.
[15] C. Lee, M. Van Der Schaar, Temporal phenotyping using deep predictive clustering of disease progression, in: International Conference on Machine Learning, PMLR, 2020, pp. 5767–5777.
[16] H. Aguiar, M. Santos, P. Watkinson, T. Zhu, Learning of cluster-based feature importance for electronic health record time-series, in: International Conference on Machine Learning, PMLR, 2022, pp. 161–179.
[17] O. Dinari, O. Freifeld, Sampling in Dirichlet process mixture models for clustering streaming data, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 818–835.
[18] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: International Conference on Learning Representations, 2013.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[20] 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.
[21] 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.
[22] S. Niu, Y. Song, Q. Yin, Y. Guo, X. Yang, Label-dependent and event-guided interpretable disease risk prediction using EHRs, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2021, pp. 1473–1476.
[23] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.
[24] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348.
[25] 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.
[26] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023, arXiv preprint arXiv:2301.12597.
[27] 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.
[28] V. Kumar, A. Stubbs, S. Shaw, Ö. Uzuner, Creation of a new longitudinal corpus of clinical narratives, J. Biomed. Inform. 58 (2015) S6–S10.
[29] S. Niu, J. Ma, L. Bai, Z. Wang, L. Guo, X. Yang, EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation, Inf. Fusion 102 (2024) 102069.
[30] B.C. Kwon, M.-J. Choi, J.T. Kim, E. Choi, Y.B. Kim, S. Kwon, J. Sun, J. Choo, Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records, IEEE Trans. Vis. Comput. Graphics 25 (1) (2018) 299–309.
[31] X.S. Zhang, F. Tang, H.H. Dodge, J. Zhou, F. Wang, Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2487–2495.
[32] 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.
[33] E. Choi, C. Xiao, W. Stewart, J. Sun, Mime: Multilevel medical embedding of electronic health records for predictive healthcare, Adv. Neural Inf. Process. Syst. 31 (2018).
[34] 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.
[35] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, J. Gao, Concare: Personalized clinical feature embedding via capturing the healthcare context, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 01, 2020, pp. 833–840.
[36] 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.
[37] Z. Qiao, X. Wu, S. Ge, W. Fan, MNN: multimodal attentional neural networks for diagnosis prediction, Extraction 1 (2019) A1.
[38] A.M. Alaa, M. van der Schaar, Attentive state-space modeling of disease progression, Adv. Neural Inf. Process. Syst. 32 (2019).
[39] 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.

[40] J. Li, B. Wu, X. Sun, Y. Wang, Causal hidden Markov model for time series disease forecasting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12105–12114.
[41] S. Niu, J. Ma, Q. Yin, L. Bai, C. Li, X. Yang, A deep clustering-based state-space model for improved disease risk prediction in personalized healthcare, Ann. Oper. Res. (2024) 1–26.
[42] 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.
[43] J. Sethuraman, A constructive definition of dirichlet priors, Statist. Sinica (1994) 639–650.
[44] E. Nalisnick, P. Smyth, Stick-breaking variational autoencoders, in: International Conference on Learning Representations, 2017.
[45] P.J. Rousseeuw, M. Hubert, Anomaly detection by robust statistics, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8 (2) (2018) e1236.
[46] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[47] H. Harutyunyan, H. Khachatrian, D.C. Kale, G. Ver Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, Sci. Data (ISSN: 2052-4463) 6 (1) (2019) 96, http://dx.doi.org/10.1038/s41597-019-0103-9.
[48] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.
[49] D. Zhang, F. Nan, X. Wei, S.-W. Li, H. Zhu, K. Mckeown, R. Nallapati, A.O. Arnold, B. Xiang, Supporting clustering with contrastive learning, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5419–5430.
[50] M. Camerlingo, P. Salvi, G. Belloni, T. Gamba, B.M. Cesana, A. Mamoli, Intravenous heparin started within the first 3 hours after onset of symptoms as a treatment for acute nonlacunar hemispheric cerebral infarctions, Stroke 36 (11) (2005) 2415–2420.
[51] E.C. Jauch, J.L. Saver, H.P. Adams Jr., A. Bruno, J. Connors, B.M. Demaerschalk, P. Khatri, P.W. McMullan Jr., A.I. Qureshi, K. Rosenfield, et al., Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association, Stroke 44 (3) (2013) 870–947.
[52] K. Patel, J.E. Hipskind, S.W. Akers, Cardiac arrest (nursing), 2021.
[53] M. Rodriguez Ziccardi, V.S. Pendela, M. Singhal, Cardiac Cirrhosis, StatPearls Publishing, 2021.
[54] D.S. Gaul, S. Stein, C.M. Matter, Neutrophils in Cardiovascular Disease, Oxford University Press, 2017.
[55] D.L. Glancy, W.C. Roberts, The heart in malignant melanoma: a study of 70 autopsy cases, Am. J. Cardiol. 21 (4) (1968) 555–571.
[56] M.A. Baig, S. Ali, J. Rasheed, M. Bergman, V. Privman, Severe hypoglycemia in a nondiabetic patient leading to acute respiratory failure, J. Natl. Med. Assoc. 98 (8) (2006) 1362.

Mr. Niu Shuai is a Ph.D. candidate at Department of Computer Science, Hong Kong Baptist University advised by Dr. Ma Jing. He received his Master's degree in the school of Computer Science at the University of Manchester. His research interests include data/text mining in healthcare, natural language processing, and computer vision. Now his research primarily lies in explainable AI in healthcare.

Miss. Yin Qing is a Ph.D. candidate at Alliance Manchester Business School, the University of Manchester advised by Dr. Xian Yang. She received her Master's degree in Operations Research and Cybernetics from the Center for Applied Mathematics of Tianjin University. Her research interests include artificial intelligence in healthcare, natural language processing, and data mining. Now her research primarily lies in unsupervised deep models for real-world data.



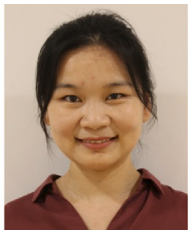
Dr. Ma Jing received her Ph.D. degree from the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong in 2020. Her research include Natural Language Processing, Social Network Analysis and Mining, Rumor Detection and Fact Verification. During Dec 2018—Aug 2019, she was a visiting scholar at Nanyang Technological University (NTU), Singapore. During Dec 2019—Feb 2020, she was a visiting scholar at Institute for Basic Science (IBS), South Korea. Before that, she obtained both of her B.E. and M.E. degree from Beijing University of Posts and Telecommunications in 2013 and 2016 respectively. In recent years, she served as Program Committee Member for ACL, EMNLP, CIKM, AAAI, etc, and was invited to review for journals such as TIST, TKDE, TOMM, etc.

Prof. Song Yunya is a Professor in the Department of Journalism and Director of AI and Media Research Lab at the School of Communication, Hong Kong Baptist University. Her research cuts across global communication, social computing, computer-mediated networks, digital media, cyber-psychology, and behaviour. Her journal articles have appeared in leading SSCI-Indexed journals, not only in the field of communication but also in other disciplines. She currently serves as the Editor of Communication & Society, and as an Associate Editor of the Journal of Contemporary Eastern Asia.

Prof. Richard Yida Xu is a professor in the Department of Mathematics at Hong Kong Baptist University. He received the B.E. degree in computer engineering from the University of New South Wales, Australia in 2001, and Ph.D. degree in computer science from the University of Technology Sydney in 2006. His current research interests include deep learning theory, Bayesian nonparametrics and machine learning applications. He has published several papers in his research fields, including IEEE Trans Cybern, IEEE TKDE, ECCV, AAAI, IJCAI.

Dr. Bai Liang received his Ph.D degree in Computer Science from Shanxi University in 2012. He is currently an Associate Professor with institute of intelligent information processing, Shanxi University. His research interest is in the areas of machine learning and data mining. He has published several papers in his research fields, including IEEE TPAMI, IEEE TKDE, IEEE TFS, ICML, KDD, AAAI.

Dr. Pan Wei is a Senior Lecturer (Associate Professor) in Machine Learning at the Department of Computer Science and a member of Centre for AI Fundamentals and Centre for Robotics and AI, The University of Manchester, UK. Before that, He was an Assistant Professor in Robot Dynamics at the Department of Cognitive Robotics and co-director of Delft SELF AI Lab, TU Delft, Netherlands and a Project Leader at DJI, China. He is an area chair or associate editor of CoRL, IEEE RAL, ICRA, IROS, ACM TOPML, IET CSR and AAAI. He received his degrees from Imperial College London, University of Science and Technology of China and Harbin Institute of Technology. He has a broad interest in robot control using Bayesian machine learning and the principles of dynamic control.



Dr Xian Yang is currently a lecturer (Assistant Professor) at Alliance Manchester Business School, the University of Manchester. Before joining AMBS, she worked as an Assistant Professor at the Department of Computer Science from Hong Kong Baptist University, a researcher at Microsoft Research Asia and a research fellow at the Data Science Institute of Imperial College London. Dr Xian Yang received her Ph.D. degree from the Department of Computing at Imperial College London in 2016. Her research interests include artificial intelligence in healthcare, natural language processing, data mining and computational epidemiology.