RAPT: Pre-training of Time-Aware Transformer for Learning Robust Healthcare Representation

Houxing Ren, Jingyuan Wang*
State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, China
Peng Cheng Laboratory, Shenzhen, China
{renhouxing,jywang}@buaa.edu.cn

Wayne Xin Zhao
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China
batmanfly@gmail.com

Ning Wu
MOE Engineering Research Center of Advanced Computer Application Technology, School of Computer Science and Engineering, Beihang University, Beijing, China
wuning@buaa.edu.cn

ABSTRACT

With the development of electronic health records (EHRs), prenatal care examination records have become available for developing automatic prediction or diagnosis approaches with machine learning methods. In this paper, we study how to effectively learn representations applied to various downstream tasks for EHR data. Although several methods have been proposed in this direction, they usually adapt classic sequential models to solve one specific diagnosis task or address unique EHR data issues. This makes it difficult to reuse these existing methods for the early diagnosis of pregnancy complications or provide a general solution to address the series of health problems caused by pregnancy complications.

In this paper, we propose a novel model RAPT, which stands for RepresentAtion by Pre-training time-aware Transformer. To associate pre-training and EHR data, we design an architecture that is suitable for both modeling EHR data and pre-training, namely time-aware Transformer. To handle various characteristics in EHR data, such as insufficiency, we carefully devise three pre-training tasks to handle data insufficiency, data incompleteness and short sequence problems, namely similarity prediction, masked prediction and reasonability check. In this way, our representations can capture various EHR data characteristics. Extensive experimental results for four downstream tasks have shown the effectiveness of the proposed approach. We also introduce sensitivity analysis to interpret the model and design an interface to show results and interpretation for doctors. Finally, we implement a diagnosis system for pregnancy complications based on our pre-training model. Doctors and pregnant women can benefit from the diagnosis system in early diagnosis of pregnancy complications.

CCS CONCEPTS

• **Applied computing → Health informatics**.

KEYWORDS

Healthcare Informatics, Representation Learning, Pre-training

ACM Reference Format:
Houxing Ren, Jingyuan Wang, Wayne Xin Zhao, and Ning Wu. 2021. RAPT: Pre-training of Time-Aware Transformer for Learning Robust Healthcare Representation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548.3467069

1 INTRODUCTION

Pregnancy complications, such as gestational diabetes and hypertension, create severe threats to the health of pregnant women. It has been reported that approximately 300,000 women died due to complications in pregnancy and childbirth in 2017 [27]. Therefore, it is important to accurately predict possible symptoms in pregnant women at an early stage. With the development of electronic health records (EHRs), prenatal care examination records have become available for developing automatic prediction or diagnosis approaches with machine learning methods [35].

In the literature, many studies on EHR data mining or automatic diagnosis have been proposed. As EHR data can be formed as sequences, sequence-to-sequence models such as recurrent neural network (RNN) and Transformer are widely used as the mainstream solutions [3, 9, 10, 23, 24, 39]. Previous methods mainly adapted these classic sequential models to solve one specific diagnosis task or address unique EHR data issues. This makes it difficult to reuse these existing methods for the early diagnosis of pregnancy complications or provide a general solution to address the series of health problems caused by pregnancy complications. To achieve this purpose, a fundamental research question is how to derive effective data representations from EHR data, which can capture the major data characteristics of examination records.

However, EHR data are rather complicated, and it is not easy to design effective representation learning methods. There are at least three major challenges to address for modeling EHR data. First, EHR data dynamically change with irregular time intervals. During pregnancy, physical characteristics, such as body weights, fundal heights and abdominal girths, may change substantially. Furthermore, examination records of prenatal care correspond to irregularly distributed samples of women’s physical characteristics during the entire pregnancy. It is difficult to effectively extract and learn time-aware representations from such dynamic, irregular and

[*] Corresponding author.

KDD ’21, August 14–18, 2021, Virtual Event, Singapore
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467069

unstable prenatal care data. Second, different pregnancy complications usually correspond to varying factors or indicators. For example, gestational diabetes is more sensitive to timesteps, while hypertension is more sensitive to examination records of specific weeks. Third, the EHR data tend to be sparse or incomplete. For example, examination records can be obtained only when pregnant women are physically examined, which is limited by the number of patients visiting the hospital.

To address these issues, the focus of this paper is to design a general and robust representation learning method for various medical downstream tasks related to pregnancy complications. Inspired by recent progress in natural language processing [14, 28], we aim to introduce a successful pre-training technique to learn effective and robust representations for various medical tasks. For this purpose, there are two important issues to consider. First, we need to design a suitable network architecture that can effectively model EHR data. Although early studies have proposed several neural network architectures, we argue that they are not the most suitable form for pre-training on EHR data. Second, we need to design specific pre-training tasks that can effectively extract data characteristics and address EHR data issues (*e.g.,* insufficiency).

For this purpose, we propose a novel representation learning method for medical data, namely RAPT, standing for RepresentAtion by Pre-training time-aware Transformer. To develop a suitable neural architecture, we extend the transformer encoder [32] by introducing a time-aware multi-head attention mechanism, which effectively handles irregular time intervals. Furthermore, we design three pre-training tasks for medical data related to pregnancy complications: (1) *similarity prediction*, especially for addressing data insufficiency, (2) *masked prediction*, especially for addressing data incompleteness, and (3) *reasonability check*, especially for addressing short sequence problems. As the entire solution, the proposed neural architecture can be effectively trained with three special pre-training tasks, which provides an effective representation learning approach for medical data.

The main contribution of the paper is that we design a novel network architecture which is suitable for modeling EHR data and we propose a new learning paradigm for modeling EHR data. To the best of our knowledge, it is the first time that pre-training is applied for representation learning on EHR data. Our model can derive robust representations by addressing the issues of data insufficiency, data incompleteness, short sequence problems by pre-training. The proposed representation learning method can then be fine-tuned according to specific downstream tasks. We construct extensive experiments on four typical medical tasks and present an application system for assisting doctors in diagnosis. Experimental results demonstrate the effectiveness of the proposed representation learning method. We believe doctors and pregnant women can benefit from our model and the model-based diagnosis system in pregnancy complication early diagnosis applications.

2 PRELIMINARIES

In this section, we introduce the background and notations used throughout the paper and formally define our dataset.

Examination Record. During pregnancy, pregnant women need to visit the hospital multiple times for prenatal care and each visit is composed of multiple examinations. We define the examination results of the t-th prenatal care as a vector $\boldsymbol{c}_t = (e_1, e_2, \ldots, e_{N_e})$, where e_i denotes the i-th examination record and N_e denotes the number of examination records.

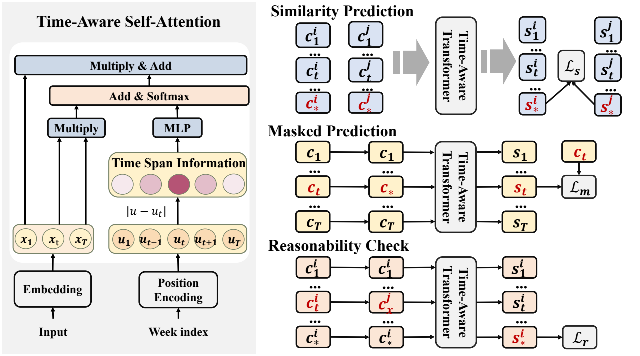


Figure 1: Overview of the proposed RAPT method.

Prenatal Care Sequence Data. Prenatal care refers to medical checkups to keep pregnant women and their babies healthy during pregnancy. We represent the prenatal care records for a woman during her pregnancy as a visit sequence of chronologically ordered events with irregular time intervals. For the i-th pregnant woman, her visit sequence of prenatal care is denoted as $< \boldsymbol{C}^{(i)}, \boldsymbol{\tau}^{(i)} >$, where $\boldsymbol{C}^{(i)} = \left(\boldsymbol{c}_1^{(i)}, \boldsymbol{c}_2^{(i)}, \ldots, \boldsymbol{c}_{T_i}^{(i)}\right)$ is the visit sequence of prenatal care of the i-th pregnant woman, and $\boldsymbol{\tau}^{(i)} = \left(\tau_1^{(i)}, \tau_2^{(i)}, \ldots, \tau_{T_i}^{(i)}\right)$ is the week index of each visit and $\tau_t^{(i)}$ denotes the week index of the t-th visit. For different pregnant women, the τ_t for the same t might be different. Moreover, due to some uncertain factors, such as premature delivery or skipping some examinations, the total number of visits, denoted as T_i , for different pregnant women are also likely to be different.

3 MODEL

In this section, we present the proposed RAPT model. Our core idea is to learn robust representations by pre-training. The overall architecture for the proposed model is presented in Fig. 1. We start with how to model the EHR data and then present three pre-training tasks for different EHR data characteristics. After that, we present how to fine-tune various downstream tasks and finally discuss how to update the model and train the entire network.

3.1 Time-aware Transformer

Here, we embed prenatal care data and learn representations for both pre-training and fine-tuning tasks.

As mentioned in Section 2, prenatal care data can be formed as sequences and employing a Transformer Encoder Layer [32] as representation extractor is straightforward. However, considering irregular time intervals, we propose a new attention mechanism called *Time-aware Multi-head Attention (TMA).*

In the TMA, we add a virtual visit $\boldsymbol{c}_*$ with virtual week $\tau_* = \tau_T + 1$ to each sequence, *i.e.,* expressing the visit sequence of user i as

$C^{(i)} = \left(\boldsymbol{c}_1^{(i)}, \ldots, \boldsymbol{c}_{T_i}^{(i)}, \boldsymbol{c}_*^{(i)}\right)$. The representation of the virtual visit is used to represent the sequence for downstream tasks.

Given a visit vector $\boldsymbol{c_t}$ in $\boldsymbol{C}^{(i)}$, we first encode it to a high-dimensioned space $\boldsymbol{x_t} \in \mathbb{R}^h$ using a fully connected layer as:

$$\boldsymbol{x}_t = \boldsymbol{W}_x \times \boldsymbol{c}_t + \boldsymbol{b}_x, \tag{1}$$

where $\boldsymbol{W}_x \in \mathbb{R}^{h \times N_e}$ and $\boldsymbol{b}_x \in \mathbb{R}^h$ are learnable parameters. Then we can obtain the embedding matrix $\boldsymbol{X}$ by stacking the vectors of each visit: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T, \boldsymbol{x}_*)$.

Then following the Transformer, we employ position encoding to encode position information, but consider the irregular time intervals. Specifically, we use the week index τ to substitute the position in the Transformer, and add the generated position encoding to the input $\boldsymbol{x_t}$ as:

$$\begin{aligned} \boldsymbol{u}_t &= PE\left(\tau_t\right), \\ \boldsymbol{p}_t &= \boldsymbol{x}_t + \boldsymbol{u}_t, \end{aligned} \tag{2}$$

where $PE(\cdot)$ denotes the position encoding in the Transformer. Thus, we can stack the vectors of each visit to form a matrix to represent each pregnant woman as $\boldsymbol{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_T, \boldsymbol{p}_*)$.

To handle irregular time intervals, we propose a *Time-Aware Self-Attention (TSA)* to replace the standard self-attention of the Transformer. If we use $\boldsymbol{P}$ as both query $\boldsymbol{Q}$ and key $\boldsymbol{K}$ in standard self-attention, given two input vectors $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, it can be calculated as:

$$\boldsymbol{A}_{i,j} = \frac{\boldsymbol{q}_i^\top \boldsymbol{k}_j}{\sqrt{h}} = \frac{\boldsymbol{x}_i^\top \boldsymbol{x}_j + \boldsymbol{x}_i^\top \boldsymbol{u}_j + \boldsymbol{u}_i^\top \boldsymbol{x}_j + \boldsymbol{u}_i^\top \boldsymbol{u}_j}{\sqrt{h}}. \tag{3}$$

Since $\boldsymbol{x}_i^\top \boldsymbol{u}_j + \boldsymbol{u}_i^\top \boldsymbol{x}_j + \boldsymbol{u}_i^\top \boldsymbol{u}_j$ cannot reflect the size of the time interval, we directly use $\boldsymbol{X}$ as both query $\boldsymbol{Q}$ and key $\boldsymbol{K}$, and then introduce a fully connected layer to capture the time span information. The time-aware self-attention $\boldsymbol{A}^{(T)}$ is expressed as:

$$\boldsymbol{A}_{i,j}^{(T)} = \frac{\boldsymbol{q}_i^\top \boldsymbol{k}_j + W_\tau \times |\boldsymbol{u}_i - \boldsymbol{u}_j|}{\sqrt{h}}, \tag{4}$$

where $W_\tau \in \mathbb{R}^{1 \times h}$ is learnable parameter, and $|\cdot|$ is and elementwise absolute value. The time interval information is incorporated into the time-aware attention using $|\boldsymbol{u}_i - \boldsymbol{u}_j|$. Therefore, the input of TSA includes not only the input of standard self-attention: query $\boldsymbol{Q}$, key $\boldsymbol{K}$ and value $\boldsymbol{V}$, but also learnable parameter $\boldsymbol{W}_\tau$ and week indexes $\boldsymbol{\tau}$. Then the proposed TSA can be expressed as:

$$TSA\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{W}_\tau, \boldsymbol{\tau}\right) = softmax(\boldsymbol{A}^{(T)}) \times \boldsymbol{V}. \tag{5}$$

Following the multi-head mechanism of Transformer, the proposed *Time-aware Multi-head Attention (TMA)* concatenates multiple individual TSAs as inputs to a fully connected layer:

$$\begin{aligned} G^{(i)} &= TSA\left(\boldsymbol{W}_q^{(i)} \boldsymbol{X}, \boldsymbol{W}_k^{(i)} \boldsymbol{X}, \boldsymbol{W}_v^{(i)} \boldsymbol{X}, \boldsymbol{W}_\tau^{(i)}, \boldsymbol{\tau}\right), \\ \boldsymbol{G} &= \boldsymbol{W}_o \times \left(G^{(1)} \| G^{(2)} \| \ldots \| G^{(n)}\right), \end{aligned} \tag{6}$$

where $\boldsymbol{W}_q^{(i)}, \boldsymbol{W}_k^{(i)}, \boldsymbol{W}_v^{(i)} \in \mathbb{R}^{h \times h}$, $\boldsymbol{W}_\tau^{(i)} \in \mathbb{R}^{1 \times h}$, $\boldsymbol{W}_o \in \mathbb{R}^{h \times nh}$ are learnable parameters, "$\|$" is the concatenating operation and n denotes the number of attention heads. After time-aware multi-head attention, we employ residual connection [18] and layer normalization [1]. Finally, a feed-forward layer processes the hidden state $\boldsymbol{h}_t$ for each visit separately as:

$$\boldsymbol{s}_t = \boldsymbol{W}_2 \times ReLU\left(\boldsymbol{W}_1 \times \boldsymbol{g}_t + \boldsymbol{b}_1\right) + \boldsymbol{b}_2, \tag{7}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{h \times h}$, $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^h$ and $ReLU(x) = \max(0, x)$ is the activation. Then we obtain the final representations of the visit sequence by connecting a residual connection [18] and layer normalization [1] as:

$$S = \left(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T, \boldsymbol{s}_*\right). \tag{8}$$

3.2 Pre-training for Robust Representations

We pre-train RAPT using three pre-training tasks for different EHR data problems, as described in this section.

Pre-training Task #1: Similarity Prediction. We can classify pregnant women according to their health condition, such as healthy pregnant women, pregnant women with high blood pressure, and pregnant women with high body mass index (BMI). If the model can distinguish different health conditions, the representations extracted by the model will be more helpful to downstream tasks. However, there are two problems in enabling the model to distinguish different health conditions. First, it is more difficult to collect enough records with various kinds of abnormal conditions. For example, the incidence rate is approximately 5% for gestational hypertension [38]. Second, we do not have labels of different health conditions. Considering the above two issues, we introduce the similarity prediction task. We first measure the Euclidean distance of all pregnant women's last visits (considering the data miss rate, we only use three records with the lowest miss rate). Then we take the 15% pairs with the smallest distance as the positive samples and the 15% pairs with the largest distance as the negative samples to train the model. In this way, the number of the samples for training is from N to $0.3 \times \frac{N \times (N-1)}{2}$.

We employ Siamese Network [11] in this task, which learns to map samples into a latent space, where the samples with close feature distances have close semantic distances. Given the representations of a sample pair, denoted by $\boldsymbol{s}_*^{(i_1)}$ and $\boldsymbol{s}_*^{(i_2)}$ (see Eq. (8)), we use the Euclidean distance to measure their semantic distance as $d_i = ||\boldsymbol{s}_*^{(i_1)} - \boldsymbol{s}_*^{(i_2)}||_2$. Formally, the loss of the Siamese network over N_p pairs is defined as:

$$\mathcal{L}_s = \frac{1}{N_p} \sum_{i=1}^{N_p} z_i d_i^2 + (1 - z_i) \max\left(m - d_i, 0\right)^2, \tag{9}$$

where z_i indicates whether the pair is same ($z_i = 1$ when the pair is same, otherwise $z_i = 0$), d_i indicates the Euclidean distance of $\boldsymbol{s}_*^{(i_1)}$ and $\boldsymbol{s}_*^{(i_2)}$, and N_p is the number of pairs and m is a preset parameter.

Pre-training Task #2: Masked Prediction. A large number of examination records in the prenatal care data are missing due to the small number of examination items per visit or statistical errors. Although we can use the average of existing data to fill missing items, the absence of some important examination records such as blood pressure will make it difficult to fit downstream tasks. Therefore, the model should have the ability to predict important examination records through examination records of other weeks. By following BERT [14], we introduce the masked prediction task. We randomly mask 30% of all visits by c_*, and then use the corresponding hidden state to predict the important examination records. Our predicted objective set E consists of diastolic pressure, systolic pressure, weight and fundal height.

Let $\boldsymbol{s}^{\dagger}$ denote the corresponding hidden state of a masked visit $\boldsymbol{c}^{\dagger}$. We employ multilayer perceptron (MLP) with ReLU activation to predict the source examination records:

$$\hat{\boldsymbol{c}}^{\dagger} = MLP\left(\boldsymbol{s}^{\dagger}\right). \tag{10}$$

Then, we employ mean squared error (MSE) as the optimized objective:

$$\mathcal{L}_m = \frac{1}{|C^{\dagger}|} \sum_{\boldsymbol{c}^{\dagger} \in \boldsymbol{C}^{\dagger}} ||\hat{\boldsymbol{c}}^{\dagger} - \boldsymbol{c}^{\dagger}||_2^2, \tag{11}$$

where $\boldsymbol{C}^{\dagger}$ is the set that consists of all masked visits, and $|\boldsymbol{C}^{\dagger}|$ denotes the size of $\boldsymbol{C}^{\dagger}$.

Pre-training Task #3: Reasonability Check. During the whole pregnancy, there is a changing trend in specific examination records. If the model can capture the trend, the model can use this trend to predict future examination records and accomplish various downstream tasks. As a result, we introduce the reasonability check task: 50% of the sequences are selected as negative samples, then 50% - 75% visits of these sequences are randomly selected and replaced with visits from other sequences. The other 50% are positive samples, and we do nothing for these sequences. The task allows the model to distinguish normal trends and abnormal trends, thus capturing trends in the examination records.

For all samples, we employ multilayer perceptron (MLP) with ReLU activation to predict the reasonability:

$$\hat{r}_i = Sigmoid\left(MLP\left(\boldsymbol{s}_*^{(i)}\right)\right), \tag{12}$$

where $Sigmoid(x) = \frac{1}{1+e^{-x}}$ is the activation function for mapping the output to $[0,1]$. Let r denote the reasonability label. We employ cross-entropy loss as optimized objective:

$$\mathcal{L}_r = -\frac{1}{N_r} \sum_{i=1}^{N_r} \left(r_i \log\left(\hat{r}_i\right) + \left(1 - r_i\right) \log\left(1 - \hat{r}_i\right)\right), \tag{13}$$

where N_r is the number of samples.

Pre-training Loss. When pre-training the model, we simultaneously train the model on three pre-training tasks, and the final pre-training loss is defined as:

$$\mathcal{L}_p = \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_m + \left(1 - \lambda_1 - \lambda_2\right) \mathcal{L}_r, \tag{14}$$

where $\mathcal{L}_s$ (Eq. (9)), $\mathcal{L}_m$ (Eq. (11)), and $\mathcal{L}_r$ (Eq. (13)) are the loss of similarity prediction, masked prediction and reasonability check, respectively, and λ_1 and λ_2 are hyperparameters to balance the three pre-training tasks.

3.3 Fine-tuning for Downstream Tasks

After pre-training the model with the objective in Eq. (14), we initialize the model with the pre-trained parameters and drop the parameters of MLP for masked prediction and reasonability check.

Classification Task. For the classification task, we employ a simple fully connected layer with Sigmoid activation as follows:

$$\hat{y}_i = Sigmoid\left(\boldsymbol{W}_c \times \boldsymbol{s}_*^{(i)} + \boldsymbol{b}_c\right), \tag{15}$$

where $\boldsymbol{W}_c \in \mathbb{R}^{1 \times h}$ and $\boldsymbol{b}_c \in \mathbb{R}^1$. Let y denote the label, then we can employ cross-entropy loss as the optimized objective:

$$\mathcal{L}_c = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right) \log\left(1 - \hat{y}_i\right)\right), \tag{16}$$

where N is the number of pregnant women.

Regression Task. For the regression task, we also use a simple fully connected layer but without activation function:

$$\hat{\boldsymbol{y}}_i = \boldsymbol{W}_r \times \boldsymbol{s}_*^{(i)} + \boldsymbol{b}_r, \tag{17}$$

where $\boldsymbol{W}_r \in \mathbb{R}^{N_f \times h}$, $\boldsymbol{b}_r \in \mathbb{R}^{N_f}$ and N_f is the number of regression objective. Let $\boldsymbol{y}$ denote the label, then we can employ mean squared error (MSE) as the optimized objective:

$$\mathcal{L}_r = \frac{1}{N} \sum_{i=1}^{N} ||\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i||_2^2, \tag{18}$$

where N is the number of pregnant women.

3.4 Learning and Discussion

Model Parameters. For a specific task, the parameters of the time-aware transformer and the prediction component are the model parameters. Given a visit sequence of pregnant women, the component generates a representation of the pregnant women. Then, a specific prediction component takes the representation as input and outputs the prediction result.

Training process. In the training process, we first pre-train the time-aware transformer with three pre-training tasks. Once our model has been pre-trained, we tune both the time-aware Transformer and prediction component using the task-specific loss to obtain better performance. Compared with traditional EHR data modeling approaches, RAPT has the following merits. First, the proposed model can handle various problems, such as data insufficiency and data incompleteness. Second, the proposed model can handle several problems but does not require additional components. Third, the proposed model is not designed for specific tasks. It provides a general solution for health problems.

Pre-training Tasks. Finally, we compare our pre-training tasks with pre-training tasks in other fields. In natural language processing (NLP), the common pre-training tasks include masked language modeling (MLM) [14], next sentence prediction (NSP) [14], replaced token detection (RTD) [12] and sentence order prediction (SOP) [22]. We follow the MLM task and change it to apply the EHR data. For other tasks, considering that the examination records of healthy pregnant women are similar, NSP and RTD are not applied to EHR data. And for healthy pregnant women, in addition to few examination records such as weight have obvious changing trends, other examination records remain stable throughout pregnancy, so SOP is not applied to EHR data. In computer vision (CV), the mainstream pre-training task is instance discrimination [7, 15, 17]. The task regards each sample as a class. Obviously it does not apply to EHR data, but we propose the similarity prediction task, which has a similar effect to instance discrimination.

Table 1: Datasets statistics. "FV" and "LV" denote the first visit and last visit, respectively.

Dataset	Pre-train	Diab.	Hype.	Outcome	Period
# of samples	63,001	20,160	5,744	8,514	1,556
# of visits	427,369	137,873	38,600	57,081	19,434
Avg. # of visits	6.78	6.84	6.72	6.70	12.49
Avg. week of FV	13.82	14.46	14.51	14.50	14.63
Avg. week of LV	28.18	28.23	28.21	28.20	36.96

4 EXPERIMENTS

In this section, we construct experiments to demonstrate the effectiveness of our model.

4.1 Experimental Setup

4.1.1 Downstream tasks. We use four healthcare-related tasks to test the effectiveness of the above comparison methods.

Gestational Diabetes Prediction and Gestational Hypertension Prediction. The two tasks aim to diagnose two types of pregnancy complications: gestational diabetes and gestational hypertension. Both baselines and the proposed RAPT take prenatal care examination records before 30 weeks as input, with the output sequence representations being taken as input for a sigmoid classifier, which generates the probability of having these pregnancy complications.

Pregnancy Outcome Prediction. Pregnancy outcome prediction aims to predict the final examination records of pregnant women which represent a healthy condition at the time of delivery and can help doctors prescribe appropriate care in pre-pregnancy. We selected four examination records with low miss rates as the target: diastolic pressure, systolic pressure,weight and fundal height. Similar to complications prediction, the models only take prenatal care examination records before 30 weeks as input and then predict the outcome with a fully-connected layer.

Risk Period Prediction. Risk period prediction aims to predict all weeks of risk during pregnancy rather than diagnosing pregnancy complications, which can help doctors prescribe timely treatment. For this setup, if an entire sequence contains n visits, we generate n sequence samples with labels indicating whether the current gestational week is dangerous. Similar to diabetes prediction, we input the representations into a sigmoid classifier to obtain the risk probability. Unlike the above tasks, risk period prediction takes all weeks as input.

Considering prediction timeliness, we only used examination records before 30 weeks in several downstream tasks. Therefore, the same period of data was used in the pre-training process.

4.1.2 Construction of the Datasets. Our data were collected from the prenatal care examination records of a hospital in Beijing spanning from 2008 to 2018. For diabetes prediction and hypertension prediction, we extracted visit sequences of 2,872 pregnant women with gestational hypertension and visit sequences of 10,080 women with gestational diabetes from the original prenatal care records as positive samples. We also randomly selected an equal number of healthy pregnant women and treated their visit sequences as negative samples. For pregnancy outcome prediction, we extracted visit sequences of 2,838 pregnant women with gestational complications whose final examination diastolic pressure, systolic pressure, weight and fundal height were not missing from the dataset. We randomly extracted twice the number of healthy pregnant women from the dataset. For risk period prediction, we extracted visit sequences of 1,556 pregnant women with clear hypertension symptoms as the dataset, *i.e.,* diastolic pressure greater than $90mmHg$ or systolic pressure greater than $140mmHg$. We regarded gestational weeks with symptoms as positive samples and reset them as negative samples. All user identity information was removed or anonymized. All experiments were carried out within the hospital with strict regulations on privacy protection.

For each prenatal care visit, the examination records contained 121 numerical examination items and 2 categorical examination items. We used one-hot encoding to represent the categorical examination items and concatenated them with numerical items as the feature vector $\boldsymbol{c}_t$ for each visit. We summarized the detailed dataset statistics in Table 1.

4.1.3 Comparison Methods. We consider the following methods for comparison:

- LSTM [19]. This is the original long short-term memory neural network with visit sequences as inputs.
- Transformer [32]. This uses an attention mechanism to model sequence data, which deals with long-term dependencies.
- RETAIN [10]. This is the REverse Time AttentIoN model, employing two RNNs to generate attention weights.
- T-LSTM [3]. This is the time-aware LSTM, which adopts a decaying function to handle irregular time intervals between visits.
- Dipole [24]. This is a sequence neural network that is specifically designed for medical visit sequence data. Dipole adopts three attention mechanisms to handle long-term medical code dependencies and provide interpretability.
- HiTANet [23]. This is a hierarchical attention-based model that generates visit representations with local time embeddings and proposes a novel self-attention mechanism to associate timesteps with visits.

4.1.4 Evaluation Metrics. We use *the area under the curve (AUC)*, *F1-score (F1)* and *accuracy (ACC)* as the evaluation metrics for the three classification tasks, and use *root mean square error (RMSE)*, *mean absolute error (MAE)*, *mean absolute percentage error (MAPE)*, *coefficient of determination score (R2)* and *explain variance score (EV)* as the evaluation metrics for pregnancy outcome prediction.

For the four tasks, we split all data into three parts with a ratio of 7:1:2, namely the training set, the validation set and the test set. We trained the model with the training set, tuned the hyperparameters with the validation set, and then reported the performance on the test set.

4.1.5 Parameter Setting. We implemented all baselines and our model with PyTorch 1.7.0. For training models, we used Adam [20] with a batch size of 64. In the experiments, we set the hidden state dimension as $h = 256$ for both baselines and our approach. We set the hyperparameter $m = 3$ in Eq. (9), masked ratio = 0.3 in masked prediction, and $\lambda_1 = 0.2$, $\lambda_2 = 0.3$ in Eq. (14). Finally, we employed

Table 2: Performance comparison for four tasks. Here, "↑" indicates "larger is better" and "↓" indicates "smaller is better".

Task		Diabetes Prediction					Task		Hypertension Prediction				
Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑	Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑
Model	LSTM	0.670	0.559	0.934	0.699	0.738	Model	LSTM	0.735	0.703	0.775	0.743	0.810
	Trans.	0.737	0.643	0.872	0.740	0.811		Trans.	0.733	0.677	0.826	0.744	0.800
	RETAIN	0.644	0.522	**0.971**	0.679	0.708		RETAIN	0.738	0.681	0.812	0.741	0.814
	T-LSTM	0.726	0.631	0.891	0.739	0.795		T-LSTM	0.738	0.625	**0.901**	0.738	0.815
	Dipole	0.724	0.675	0.794	0.730	0.790		Dipole	0.737	**0.730**	0.746	0.738	0.812
	HiTANet	0.747	0.723	0.764	0.743	0.813		HiTANet	0.739	0.718	0.777	0.746	0.811
	RAPT	**0.807**	**0.836**	0.763	**0.798**	**0.867**		RAPT	**0.746**	0.671	0.840	**0.749**	**0.820**
Task		Pregnancy Outcome Prediction					Task		Risk Period Prediction				
Metric		RMSE ↓	MAE ↓	MAPE ↓	R2 ↑	EV ↑	Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑
Model	LSTM	10.661	7.449	0.094	0.000	0.000	Model	LSTM	0.909	0.770	0.838	0.802	0.959
	Trans.	8.620	5.319	0.068	0.338	0.339		Trans.	0.908	0.767	0.808	0.784	0.947
	RETAIN	9.046	5.812	0.081	0.246	0.261		RETAIN	0.848	0.550	0.694	0.613	0.854
	T-LSTM	10.664	7.454	0.104	-0.001	0.000		T-LSTM	0.908	0.772	0.821	0.795	0.960
	Dipole	9.229	6.200	0.079	0.232	0.233		Dipole	0.918	0.807	0.824	0.812	0.965
	HiTANet	8.631	5.377	0.077	0.337	0.337		HiTANet	0.900	0.759	0.775	0.767	0.943
	RAPT	**8.525**	**5.184**	**0.063**	**0.350**	**0.352**		RAPT	**0.976**	**0.964**	**0.925**	**0.944**	**0.985**

dropout [30] (dropout rate = 0.5) for the classification layer of all model on classification tasks. These hyperparameters were selected based on the performance on the validation set.

4.2 Result and Analysis

Table 2 presents the performance comparison of all the methods. From the results in Table 2, we make the following observations.

- Methods using only standard LSTM, such as LSTM and RETAIN, perform worst among all the baselines. These methods do not consider the medical data characteristics. In contrast, Transformer and Dipole can handle long-term dependencies, and T-LSTM and HiTANet can handle irregular time intervals.
- For diabetes prediction, models considering irregular time intervals, such as T-LSTM, HiTANet and RAPT, achieve better performance because some observable incipient symptoms such as obesity, weight gain, and increased age [4], are associated with gestational diabetes, which is a long-term outcome, and the irregular time interval greatly affects the performance.
- For hypertension prediction and risk period prediction tasks, because gestational hypertension does not have clear symptoms in pre-pregnancy and the main symptoms are blood pressure or proteinuria greater than a certain value [21], pregnant women can be diagnosed when these symptoms appear, so the benefit of handling irregular time intervals is small. As a result, these models perform worse than models considering other characteristics.
- For pregnancy outcome prediction, the model using only the last hidden state to represent sequences, such as LSTM and T-LSTM, performed worst among all models. It is clear that the results of all gestational weeks are helpful in predicting outcomes. However, in these models, the examination records of pre-pregnancy were weakened several times.
- Based on experience, Transformer-based models (*i.e.,* HiTANet) perform better than LSTM-based models (*i.e.,* T-LSTM and Dipole). However, because of the size of the hypertension dataset, Transformer-based models suffer from overfitting problems. Risk period prediction divides a whole sequence into shorter test samples, and thus, the ability to handle long-term dependencies is impaired. Therefore, the Transformer-based model performed worse than the LSTM-based model on the two tasks.

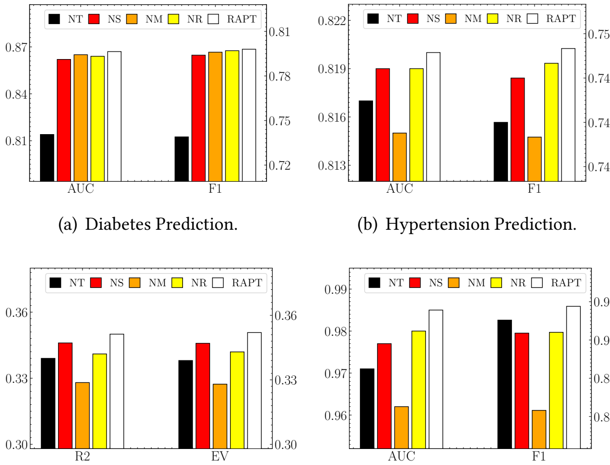
Table 3: Performance comparison of RAPT and human for hypertension prediction task.

Metric	ACC	Precision	Recall	F1	AUC
Human	**0.763**	**1.000**	0.540	0.701	0.770
RAPT	0.746	0.671	**0.840**	**0.749**	**0.820**

- Finally, the proposed RAPT model is consistently better than all of the baselines in all tasks. Our model handles irregular time intervals by time-aware multi-head attention and long-term dependencies by a transformer. In addition to the above two problems, data insufficiency, data incompleteness and short sequence problems are also important issues. Our approach alleviates this problem by pre-training. We further analyze the contribution of each pre-training task in Section 4.3.
- In addition, we tested human performance on the gestational hypertension test set and present the results in Table 3. Human performance is measured with the gold standard, *i.e.,* diastolic pressure greater than 90$mmHg$ or systolic pressure greater than 140$mmHg$. The precision of humans is 100% because when symptoms appear, pregnant women can be diagnosed with hypertension. The gold standard do not have the ability to predict future examination records, so it cannot identify pregnant women who show symptoms after 30 weeks, and the recall is relatively low. As a result, the diagnosis of the model is timelier. Thus, pregnant women can receive treatment early.

4.3 Ablation Study

In our approach, the proposed model consists of time-aware multi-head attention and three pre-training tasks: similarity prediction, masked prediction and reasonability check. Here, we determine how each part actually contributes to the final performance. We



(a) Diabetes Prediction.

(b) Hypertension Prediction.

(c) Pregnancy Outcome Prediction.

(d) Risk Period Prediction.

Figure 2: Ablation study on four tasks.

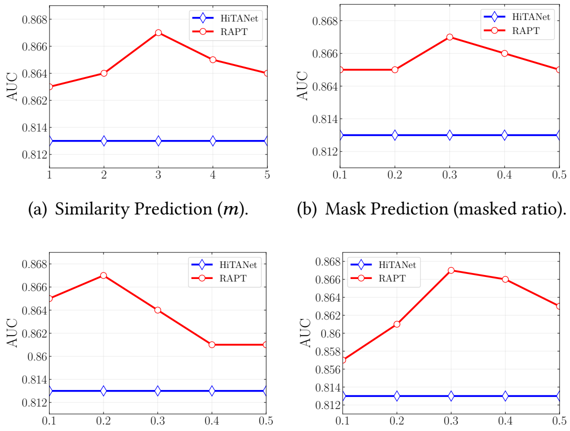
report the result of AUC and F1 scores (R2 and EV scores) for the classification task (regression task). We compare four variants of the proposed RAPT model: (1) NT without time-aware multi-head attention, (2) NS without the similarity prediction pre-training task, (3) NM without the masked prediction pre-training task, (4) NR without the reasonability check pre-training task.

Figure 2 presents all the comparison results of the four variants. First, for diabetes prediction, *NT* performed worst among all models because of the characteristics of diabetes. Second, for other tasks, *NM* performed worst because the masked prediction task enables the model to have the ability to predict missing data and all these tasks are closely related to the prediction of examination records. Third, the similar prediction task allows the model to distinguish similar and dissimilar cases and the reasonability check task enables the model to check determine the data is reasonable. The small difference between *NS* and *NR* indicates that the two pre-training tasks are similar, and the model can predict similar predictions by capturing change trends and predicting final examination records.

4.4 Parameter Sensitivity

In addition to the model components, there are several parameters to tune in our model. Here we incorporate the best baseline HiTANet for comparison. We report the tuning results with AUC scores for diabetes prediction.

We first tune the similarity prediction (m in Eq. (9)) and masked prediction (masked ratio) parameters. We vary the margin (m) of similar prediction in the set $\{1, 2, 3, 4, 5\}$, and the masked ratio of masked prediction in the set $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. As can be seen in Fig 3(a) and Fig 3(b), $m = 3$ and masked ratio= 0.3 lead to the optimal performance. Another parameter to tune is the pre-training loss weights (λ_1, λ_2 in Eq. (14)). We vary both of the weights in the set $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. In Fig 3(c) and Fig 3(d), we can see that $\lambda_1 = 0.2$ and $\lambda_2 = 0.3$ lead to the optimal performance. Overall, our model is relatively stable when varying the four parameters, and consistently better than HiTANet and the other baselines.



(a) Similarity Prediction (m).

(b) Mask Prediction (masked ratio).

(c) Pre-traing Loss (λ_1).

(d) Pre-traing Loss (λ_2).

Figure 3: Parameter sensitivity on Diabetes Prediction.

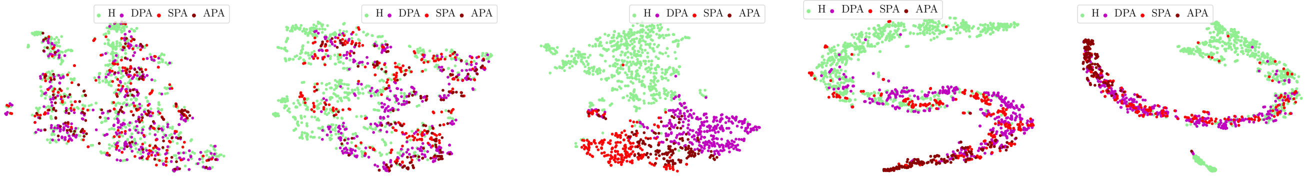
4.5 Qualitative Analysis

Previously, we showed the effectiveness of our model for four tasks. In this part, we qualitatively analyze why the learned representations are useful. We use the low-dimensional embeddings learned by t-SNE [31], which takes representations learned by different models as input. We use models trained on the period risk prediction task for analysis. We select all anomalous weeks and randomly select an equal number of normal weeks, then show the embedding distribution of RAPT in Fig. 4(c). For comparison, we also show the embedding distributions of RAPT without training (Fig. 4(a)), pre-trained RAPT (Fig. 4(b)), RAPT without pre-training (Fig. 4(d)) and the best baseline Dipole (Fig. 4(e). To quantitatively analyze these distributions, we calculated the silhouette score [29] for the five scatters, which indicates cohesion and separation of clustering results.

For RAPT without training, the points belonging to different categories were mixed. For the pre-trained model, healthy pregnant women and pregnant women with high blood pressure were slightly separated. Although the results are still unsatisfactory, considering that there are no monitoring signals, the results are acceptable. Fine-tuned RAPT, RAPT without pre-training and Dipole can all distinguish high blood pressure and healthy pregnant women, but RAPT can further distinguish pregnant women with different symptoms. In addition, the silhouette scores provide the same result. From what has been discussed above, pre-training is helpful for learning robust representations.

5 DIAGNOSIS SYSTEM

Based on our pre-training model, we implemented a system for pregnancy complication diagnosis. The system collects weekly examination records of pregnant women and automatically predicts various anomaly conditions. When anomalous conditions appear, our model first presents a warning to the doctor and then presents prediction results, examination records of pregnant women and interpretation. Figure 5 shows the interface presented to the doctors. Doctors can rely on the results and interpretation of the model to



(a) Without training ($S = -0.08$).

(b) Pre-trained ($S = -0.02$).

(c) Fine-tuned ($S = 0.33$).

(d) Without pre-training ($S = 0.16$).

(e) Dipole ($S = 0.12$).

Figure 4: Scatter plots for embeddings trained by different models. The color of the dots represents the patient condition, *i.e.,* “APA” (all pressure anomaly) indicates that diastolic pressure greater than $90mmHg$ and systolic pressure greater than $140mmHg$, “DPA” (diastolic pressure anomaly) indicates that diastolic pressure greater than $90mmHg$, “SPA” (systolic pressure anomaly) systolic pressure greater than $140mmHg$ and “H” (healthy) indicates healthy. “S” indicates Silhouette Score.

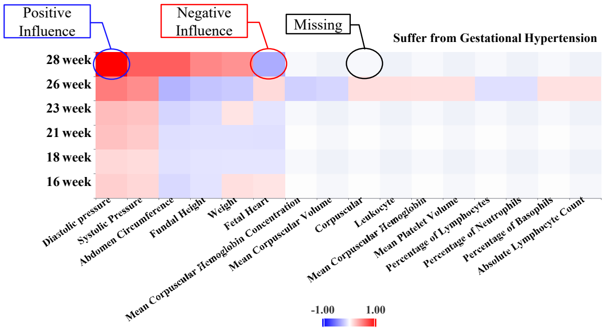


Figure 5: The interface presented to doctors. The vertical denotes the visit timestamp, and the horizontal axis denotes the examination records. For the grid, red indicates positive influence, blue indicates negative influence and the shade indicates the degree of influence.

make a diagnosis, and pregnant women can use the results of the model to receive early treatment and reduce the risk of pregnancy complications.

When the model presents a warning, the model lists the warning in the top right corner, ranks the examination results in order of average importance and lists them from left to right on the x-axis. The doctor can first look at the model predictions, then examine the data from left to right along the x-axis to make a final diagnosis. In the case shown in Fig. 5, the model indicates that the pregnant woman will suffer from gestational hypertension, and the most important examination record is diastolic pressure. For sensitivity, the model assumes that examination records such as diastolic pressure have a positive influence on the probability of suffering from gestational hypertension and that the probability of suffering from gestational hypertension increases as these values increase. In contrast, examination records such as fundal height have a negative influence, and the probability of suffering from gestational hypertension increases as these values decrease.

Following MPCE[33], we introduce sensitivity analysis to calculate interpretation of the interface. By integrating a time-aware Transformer (Eq. (8)) and prediction components (Eq. (16) or Eq. (18)), we obtain the complete RAPT model f for a specific task. The sensitivity of an examination record e of the sequence is given below:

$$\delta_e^y(C, \boldsymbol{\tau}) = \lim_{\Delta e \to 0} \frac{f\left(e + \Delta e, C_{\neg e}, \boldsymbol{\tau}\right) - f\left(C, \boldsymbol{\tau}\right)}{\Delta e} = \frac{\partial f}{\partial e}, \qquad (19)$$

where $C_{\neg e}$ denotes the elements of C except e. A positive value of $\delta_e^y(C, \boldsymbol{\tau})$ indicates that our model tends to regard y as large when e is large, and vice versa.

6 RELATED WORK

Deep Learning for Modeling EHR Data. Since healthcare became an important research domain, various deep learning models have been proposed for modeling EHR data. These models include multilayer perceptron (MLP) based models [6], convolutional neural network (CNN) based models [8, 26], recurrent neural network (RNN) based models [2, 3, 10, 16, 24] and Transformer based models [23, 39]. Since medical data can be formed as sequences, RNN based models and Transformer based models have been widely adopted. To interpret the predicted results, Choi et al. proposed RETAIN [10] to retain the RNN prediction accuracy with better interpretation. The Dipole [24] employed attention mechanisms to obtain the most important visit and solve the long-dependency issue. INPREM [39] designs a linear model for interpretability. To handle irregular time intervals, Baytas et al. proposed T-LSTM [3] to address irregular elapsed times by adding a new gate in LSTM cell. TimeLine [2] introduced a time-aware function to control how much information flows into the RNN. HiTANet [23] is a time-aware Transformer that models irregular time intervals. To address the data insufficiency problem, GRAM [9] employed medical knowledge using graph-based attention and MetaPred [40] introduces meta-learning.

Self-supervised Pre-training. Self-supervised pre-training has gained popularity in recent years because of its ability to utilize large quantities of unlabeled data. In computer vision (CV), instance discrimination is the mainstream approach. Most of earlier works used instance-level classification approach [15, 36]. In recent years, contrastive learning has been successfully applied [37]. To introduce more negative samples, MoCo [17] employs a dictionary queue, PIRL [25] employs a memory bank, and SimCLR [7] uses a large batch rather than techniques for increasing negative samples and obtains better performance. SwAV [5] uses a clustering-based contrastive learning approach and achieves performance close to supervised learning. In the natural language process (NLP), the mainstream method is contextual methods [13, 14, 28]. To introduce contextual information, LM-LSTM [13] trains the model by predicting the next token using previous tokens. ELMo [28] trains the model by predicting a token using tokens from both directions. BERT [14] employs Transformer Encoder Layer and two pre-training task to train the model. Then, there are many variants of BERT, such as ALBERT [22], and StructBERT [34]. All of these models achieves better performance, and BERT-based methods have become the mainstream NLP method.

7 CONCLUSION

In this paper, we studied how to effectively represent EHR data for various downstream tasks. We first designed a novel architecture that is suitable for modeling EHR data, and we proposed pre-training for modeling EHR data. Then, we carefully devised three pre-training tasks to enable the model to handle various characteristics in EHR data, such as insufficiency and incompleteness. Extensive experimental results for four tasks demonstrated the effectiveness and robustness of the proposed model. We also introduced an interpretation method by sensitivity analysis and designed an interface to show the prediction results and interpretation.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2019YFB2102103), the National Natural Science Foundation of China (Grant No. 92046010, 82161148011, 61872369), the Fundamental Research Funds for the Central Universities (Grant No. YWF-21-BJ-J-839), and Beijing Outstanding Young Scientist Program under (Grant No. BJJWZYJH012019100020098).

REFERENCES

[1] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *CoRR* abs/1607.06450 (2016).

[2] Tian Bai, Shanshan Zhang, Brian L. Egleston, and Slobodan Vucetic. 2018. Interpretable Representation Learning for Healthcare via Capturing Disease Progression through Time. In *KDD*. ACM, 43–51.

[3] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient Subtyping via Time-Aware LSTM Networks. In *KDD*. ACM, 65–74.

[4] Thomas A Buchanan, Anny H Xiang, et al. 2005. Gestational diabetes mellitus. *The Journal of clinical investigation* 115, 3 (2005), 485–491.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *NeurIPS*.

[6] Zhengping Che, David C. Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep Computational Phenotyping. In *KDD*. ACM, 507–516.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1597–1607.

[8] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In *SDM*. SIAM, 432–440.

[9] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. GRAM: Graph-based Attention Model for Healthcare Representation Learning. In *KDD*. ACM, 787–795.

[10] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter F. Stewart. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *NIPS*. 3504–3512.

[11] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *CVPR (1)*. IEEE Computer Society, 539–546.

[12] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*. OpenReview.net.

[13] Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised Sequence Learning. In *NIPS*. 3079–3087.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.

[15] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin A. Riedmiller, and Thomas Brox. 2014. Discriminative Unsupervised Feature Learning with Convolutional Neural Networks. In *NIPS*. 766–774.

[16] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M. Glass, and Jimeng Sun. 2020. StageNet: Stage-Aware Neural Networks for Health Risk Prediction. In *WWW*. ACM / IW3C2, 530–540.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*. IEEE, 9726–9735.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[20] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

[21] Evangelia Kintiraki, Sophia Papakatsika, George Kotronis, Dimitrios G Goulis, and Vasilios Kotsis. 2015. Pregnancy-induced hypertension. *Hormones* 14, 2 (2015), 211–223.

[22] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *ICLR*. OpenReview.net.

[23] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. HiTANet: Hierarchical Time-Aware Attention Networks for Risk Prediction on Electronic Health Records. In *KDD*. ACM, 647–656.

[24] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In *KDD*. ACM, 1903–1911.

[25] Ishan Misra and Laurens van der Maaten. 2020. Self-Supervised Learning of pre-training-Invariant Representations. In *CVPR*. IEEE, 6706–6716.

[26] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2017. Deepr: A Convolutional Net for Medical Records. *IEEE J. Biomed. Health Informatics* 21, 1 (2017), 22–30.

[27] World Health Organization. 2017. *World Health Statistics 2017: Monitoring Health for the SDGs, Sustainable Development Goals.* World Health Organization. https://books.google.com/books?id=JVXptAEACAAJ

[28] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *NAACL-HLT*. Association for Computational Linguistics, 2227–2237.

[29] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[30] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1 (2014), 1929–1958.

[31] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.

[33] Jingyuan Wang, Ze Wang, Jianfeng Li, and Junjie Wu. 2018. Multilevel Wavelet Decomposition Network for Interpretable Time Series Analysis. In *KDD*. ACM, 2437–2446.

[34] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. 2020. StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. In *ICLR*. OpenReview.net.

[35] Yan-Ting Wu, Chen-Jie Zhang, Ben Willem Mol, Andrew Kawai, Cheng Li, Lei Chen, Yu Wang, Jian-Zhong Sheng, Jian-Xia Fan, Yi Shi, et al. 2020. Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning. *The Journal of Clinical Endocrinology & Metabolism* (2020).

[36] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. *CoRR* abs/1805.01978 (2018).

[37] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. 2019. Unsupervised Embedding Learning via Invariant and Spreading Instance Feature. In *CVPR*. Computer Vision Foundation / IEEE, 6210–6219.

[38] Jun Zhang, Susan Meikle, and Ann Trumble. 2003. Severe maternal morbidity associated with hypertensive disorders in pregnancy in the United States. *Hypertension in pregnancy* 22, 2 (2003), 203–212.

[39] Xianli Zhang, Buyue Qian, Shilei Cao, Yang Li, Hang Chen, Yefeng Zheng, and Ian Davidson. 2020. INPREM: An Interpretable and Trustworthy Predictive Model for Healthcare. In *KDD*. ACM, 450–460.

[40] Xi Sheryl Zhang, Fengyi Tang, Hiroko H. Dodge, Jiayu Zhou, and Fei Wang. 2019. MetaPred: Meta-Learning for Clinical Risk Prediction with Limited Patient Electronic Health Records. In *KDD*. ACM, 2487–2495.