

# Disease Risk Prediction Based on EHR

Ningning Li<sup>1</sup>, Qian Lu<sup>1</sup>, Min Liang<sup>1</sup>, Dong Lin<sup>2</sup>, Yuchang Mo<sup>1</sup>

<sup>1</sup>Fujian Province University Key Laboratory of Computational Science, School of Mathematical Sciences, Huaqiao University, Quanzhou Fujian

<sup>2</sup>College of Acupuncture, Fujian University of Traditional Chinese Medicine, Fuzhou Fujian  
Email: yuchangmo@sina.com

Received: Dec. 23<sup>rd</sup>, 2019; accepted: Jan. 1<sup>st</sup>, 2020; published: Jan. 8<sup>th</sup>, 2020

---

## Abstract

Data driven health care, as the use of available large-scale medical data, to provide the best and most personalized care, is becoming one of the main trends of the success of the revolution in the medical industry. Electronic health record is the main carrier to promote the success of this data-driven medical revolution. In this paper, we use the method of deep learning, based on the word embedding model to express the EHR information, and use the characteristics of the long-term memory network model to solve the irregular time of EHR information and the long-term dependence of disease information, so as to achieve the prediction of disease risk. Compared with the convolution neural network model, the results show the effectiveness of this method.

## Keywords

Deep Learning, Electronic Health Records, Word Embedding, Long Short-Term Memory Network

---

# 基于电子病历的疾病风险预测

李宁宁<sup>1</sup>, 陆 迁<sup>1</sup>, 梁 敏<sup>1</sup>, 林 栋<sup>2</sup>, 莫毓昌<sup>1</sup>

<sup>1</sup>华侨大学, 数学科学学院, 计算科学福建省高校重点实验室, 福建 泉州

<sup>2</sup>福建中医药大学, 针灸学院, 福建 福州

Email: yuchangmo@sina.com

收稿日期: 2019年12月23日; 录用日期: 2020年1月1日; 发布日期: 2020年1月8日

---

## 摘 要

数据驱动的医疗保健(data driven healthcare)作为可用的大型医疗数据的使用, 以提供最好的和最个性化的护理, 正成为医疗行业革命成功的主要趋势之一。电子病历是推动这一数据驱动的医疗革命成功的主要载体。本文运用深度学习的方法, 基于词嵌入模型实现对电子病历信息的表示, 利用长短期记忆

(LSTM)网络模型特性,实现对电子病历信息时间上的无规律性和疾病信息长期依赖挑战的解决,实现对疾病风险的预测,通过与卷积神经网络(CNN)模型进行比对,实验结果显示本文方法的有效性。

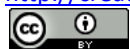
## 关键词

深度学习, 电子病历, 词嵌入, 长短期记忆网络

Copyright © 2020 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

全球医疗保健系统正在迅速采用电子病历(EHR),它是在任何护理交付环境[1]中由一个或多个相遇产生的纵向患者健康信息的系统集成(例如,诊断、药物、实验室测试、程序等)。这将极大地增加有效的电子临床数据。因此,数据驱动的医疗保健作为可用的大型医疗数据的使用,以提供最好的和最个性化的护理,正成为医疗行业革命成功的主要趋势之一[2][3][4]。由于患者 EHR 是进行数据驱动的医疗研究的主要载体,了解 EHR 中包含的信息并对其信息有效提取是至关重要的。

## 2. 相关工作

近年来,病历中表型的提取和表示的现有方法包括:基于向量的表示。这种方法针对每个病人构建了一个向量,它的维度等于在 HER 中出现的不同事件的数量,每个维度上的值是在特定时间段内相应医疗事件的汇总统计(例如,总和、平均值、最大值、最小值等)。使用基于向量的表示,每个表型通常被假定为这些原始医疗事件的线性组合,组合系数是通过一些优化过程获得的,这种表示的局限性在于它忽略了这些事件之间的时间关系。Ho 等人[5]提出了一种基于非负张量因子分解的 EHR 张量表型提取方法。该方法探索了不同医学实体之间的相互作用;其局限性是,他们仍然没有考虑事件-时间关系。基于序列的表示,该方法根据每个事件的时间戳为每个患者构建 EHR 序列,然后,运用频繁模式挖掘方法将时间模式识别为表型,一个问题在于,患者 EHR 之间的高度可变性,这种方法通常会返回大量的模式(也称为“模式爆炸”现象)。基于时间矩阵的表示[6],这种方法表示患者 EHR 作为时间矩阵,其中一维对应于时间,另一维对应于医疗事件。Zhou 等人[7]提出了将相似时态趋势的医疗事件分组在一起的表型分型方法。然而,他们没有考虑不同事件之间的时间关系。Wang 等人[8]提出了一种卷积矩阵分解方法,用于检测患者 EHR 矩阵上的移位不变模式,但不能确定最优模式长度,需要枚举所有可能的值。

为了克服上述电子病历信息的稀疏性、时间上无规律性和疾病信息的长期依赖性挑战,本文将基于自然语言处理[9]中的词嵌入模型将 EHR 中离散的元素嵌入到连续向量空间中,将相同类型的向量合并为单个向量,把向量经过特定池化类型处理表示为一次入院,以这种方式,可变大小的输入被嵌入到连续分布的向量空间中,作为 LSTM 的输入特征,通过 LSTM 对时序信息的有效处理,实现对疾病风险的预测。因为嵌入是从数据中学习的,所以模型不依赖于人工特征工程。实验结果表明,该模型具有有效性。

## 3. 基本概念

### 3.1. 词表示

用“词”来指一个大的上下文中的离散元素(例如,文档中的一个词或一次入院的一种诊断),输入到

许多机器学习模型中通常被表示为固定长度的特征向量。对于文本，通常使用词袋模型。词  $w$  用一个独热向量  $V_w \in R^{|V|}$  表示，其中  $V_w = [v_w^1, \dots, v_w^{|V|}]$ ， $|V|$  是词典中词的数量： $V_w = [0, \dots, 0, 1, 0, \dots, 0]$  (如果  $w = i$ ，则  $v_w^i = 1$ ，表示  $w$  是词典中的第  $i$  个词，否则  $v_w^i = 0$ )。在“词袋”的表示法下，一个句子  $w_0, \dots, w_n$  的向量是其词向量的总和： $u = v_{w_0} + v_{w_1} + \dots + v_{w_n}$ 。然而，“词袋法”无法捕获单词的顺序和语义。

### 3.2. 学习词嵌入矩阵

一种强有力的词袋替代方法是将单词嵌入到  $M$  维的向量空间中，其中  $M \ll |V|$ 。每一个词都映射到一个唯一的向量，该向量是矩阵  $E \in R^{M \times |V|}$  中的一列。词嵌入有几个好处：首先，维数大大减少，不依赖于新词的出现。第二，一个单词的语义是以分布式的方式表示的，也就是说，有多个元素对单词的意思进行编码。第三，使用诸如加法和矩阵乘法等现有代数工具，连续向量的操作要容易得多，这在最近的工作中得到了证明。例如，两个词之间的相似性只是两个向量之间的余弦。更重要的是，嵌入矩阵  $E$  可以从数据中学习。

学习嵌入矩阵  $E$  的方法有很多种，最流行的方法可能是连续的词袋模型。对于词序列中的一个词  $w_i$ ，模型使用  $w_i$  周围的词预测  $w_i$ 。输入上下文大小为  $C$  时， $w_{i-C}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+C}$  被称为  $w_i$  的上下文词。利用嵌入矩阵  $E$  将所有上下文词嵌入向量中，然后求平均值得到平均向量  $h$

$$h = \frac{E^{w_{i-C}} + \dots + E^{w_{i-1}} + E^{w_{i+1}} + \dots + E^{w_{i+C}}}{2C} \tag{1}$$

其中  $E^t$  是矩阵  $E$  的第  $t$  列。然后模型生成输出  $a = \bar{E}h$ ，其中  $\bar{E} \in R^{|V| \times M}$ ，中心词  $w_i$  预测使用 softmax 函数

$$P(w_i | w_{i-C}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+C}) = \text{softmax}(a) \tag{2}$$

参数  $E$  和  $\bar{E}$  通过最小化损失函数来学习

$$L = \frac{1}{T} \sum_{i=1}^T \log P(w_i | w_{i-C}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+C}) \tag{3}$$

通过随机梯度下降的反向传播。

### 3.3. 长短期记忆网络(LSTM)

长短期记忆网络[10]是一种有效解决梯度消失问题的循环神经网络(RNN) [11]。LSTM 的中心是一个线性自循环记忆细胞，允许记忆通过长序列流动。记忆细胞被封闭来调节信息流向或来自细胞的信息。LSTM 在许多应用中都是成功的，例如机器翻译、手写识别和语音识别。

一个 LSTM 单元在时间  $t$  有一个状态  $c_t \in R^k$  的记忆细胞，通过记忆细胞的信息流被 3 个门控制：输入门，遗忘门和输出门。输入门  $i_t \in R^k$  流入到细胞的输入，遗忘门  $f_t \in R^k$  控制记忆细胞的遗忘且输出门  $o_t \in R^k$  调节来自记忆细胞的输出流。在技术细节处理之前，我们用  $\sigma$  表示矢量逐元素的 sigmoid 函数，\* 表示两个矢量的乘积。

三个门都是 sigmoid 单元，它将门的每个元素的值设置为 0 到 1 之间：

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{4}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{5}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{6}$$

其中  $W_{\{i,f,o\}}, U_{\{i,f,o\}}, b_{\{i,f,o\}}$  是参数门控制通过的信息量：当值为 1 时满值，当值为 0 时完全阻塞。在每一个时间步长  $t$ ，通过输入  $x_t \in R^M$  第一次计算输入特征，先前的隐藏层  $h_{t-1} \in R^k$  通过一个挤压 tanh 函数：

$$g_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{7}$$

通过部分遗忘先前的记忆细胞和调节输入特征来更新记忆细胞，如下所示

$$c_t = f_t * c_{t-1} + i_t * g_t \tag{8}$$

记忆细胞序列是相加的，因此梯度也通过链式法则以线性方式更新。这有效地防止梯度消失或爆炸。记忆细胞在通过可学习的遗忘门  $f_t$  记忆过去的经验中起着至关重要的作用。相反， $f_t \rightarrow 1$ ，所有的过去记忆被保存，且随着新的输入新的记忆不断更新。如果  $f_t \rightarrow 0$ ，只有新的经验被更新，系统变得无记忆。

最后，根据记忆  $c_t$  计算隐藏的输出状态  $h_t$ ，由输出门  $o_t$  控制如下：

$$h_t = o_t * \tanh(c_t) \tag{9}$$

注意，由于系统动态是确定性的， $h_t$  是所有先前输入的一个函数： $h_t = \text{LSTM}(x_{1:t})$  然后，输出状态用来产生输出。

### 3.4. 模型复杂性

模型参数个数为  $M \times |V| + M \times K + K \times K + K \times D$ ，由以下部分组成：

LSTM 层的参数：

对于入院嵌入，使用嵌入矩阵  $A \in R^{M \times |V|}$ 。

输入门： $W_i \in R^{M \times |V|}$ ， $U_i \in R^{K \times K}$  和  $b_i \in R^{K \times 1}$ 。

输出门： $W_o \in R^{M \times |V|}$ ， $U_o \in R^{K \times K}$  和  $b_o \in R^{K \times 1}$ 。

遗忘门： $W_f \in R^{M \times |V|}$ ， $U_f \in R^{K \times K}$  和  $b_f \in R^{K \times 1}$ 。

记忆细胞： $W_c \in R^{M \times |V|}$ ， $U_c \in R^{K \times K}$  和  $b_c \in R^{K \times 1}$ 。

## 4. 基于电子病历的风险预测模型

本节我们具体描述对 EHR 信息的处理，利用 LSTM 对序列中的长期依赖性进行建模的能力，解决了可变大小的离散输入和无规律时间的挑战，实现对疾病风险的预测。

### 4.1. 可变大小的入院信息的表示

一次入院包含多个诊断。诊断、手术和药物是使用编码方案编码的。我们观察到 NLP 和 EHR 之间有相当大的相似性，EHR 类似于句子，所以基于 NLP 中的词嵌入模型，我们的方法是将入院嵌入到向量中。一次入院是一个可变大小的代码集(诊断或干预)。设  $D$  是诊断代码集。这个集合索引从 1 到  $D$ 。用  $A \in R^{M \times |D|}$  表示一次诊断嵌入矩阵。设  $A^j$  是第  $j$  列的元素， $A^i$  是矩阵  $A$  的第  $i$  行第  $j$  列元素。每次入院  $t$  包含  $h$  次诊断： $d_1, d_2, \dots, d_h \in \{1, 2, \dots, |D|\}$ 。代码首先被嵌入到向量中，将相同类型的向量合并为单个向

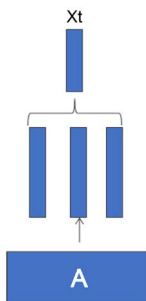


Figure 1. Code embedding  
图 1. 编码嵌入

量, 方式如图 1 所示。诊断与干预代码的嵌入向量是  $A^{d_1}, \dots, A^{d_h}$ 。然后, 我们汇集所有当前诊断向量得到  $x_i \in R^M$ , 但是得到的向量仍是稀疏的, 我们通过引入池化, 进一步处理电子病历的稀疏问题。最后, 多次入院嵌入是一个矩阵, 对其进行随机初始化, 然后通过训练预测任务来学习。

### 4.2. 池化

设  $x_i^j$  是向量  $x_i$  的第  $i$  个元素, 入院采用最大池化、合并池化或平均池化, 合并方式如下:

最大池化 admssion(max adm)。按元素池化如下:

$$x_i^j = \max(A_i^{d_1}, A_i^{d_2}, \dots, A_i^{d_h})$$

$i = 1, 2, \dots, M$ 。这类似于选择性注意诊断和干预之间的影响最大的因素。它也类似于通常的编码实践, 即选择一个诊断作为入院的主要原因。

归一化和池(sum adm)。患有多种疾病(多并发症)的患者比单病症患者更容易发生危险。我们提出了下面的归一化和池:

$$x_i^j = \frac{A_i^{d_1} + A_i^{d_2} + \dots + A_i^{d_h}}{\sqrt{|A_i^{d_1} + A_i^{d_2} + \dots + A_i^{d_h}|}}$$

$i = 1, 2, \dots, M$ 。归一化降低了大量的诊断和干预的影响。

平均池化 admission(mean adm.)。在缺乏主要条件的情况下, 平均池可能是一个合理的选择:

$$x_i^j = \frac{A_i^{d_1} + A_i^{d_2} + \dots + A_i^{d_h}}{h}$$

### 4.3. LSTM 预测模型

深度学习推动基于电子病历特征提取的疾病诊断预测[12] [13] [14], 风险预测[15] [16] [17]不断发展。本文疾病风险预测模型中, 我们使用基本的 LSTM 单元, 其行为由三个门控机制控制: 输入门、输出门和遗忘门。基于门值记忆单元在时间  $t$  累计从输入  $vt$  输入的有用的信息, 并将该信息存储在其内部状态。最终从 LSTM 输出  $Z(p)$  是病人  $P$  的向量表示。一个全连接 softmax 层被用于产生预测概率, 数学表达如下:

$$\hat{y}_p = \text{softmax}(W_y z^{(p)} + b_y) \tag{10}$$

设  $\theta$  为 LSTM 模型中所有参数的集合, 预测概率向量  $\hat{y}_p$  也可用模型后验分布  $P(y_p | X^p; \theta)$  表示, 其中  $y_p$  是真值。利用真值  $y_p$  与预测概率  $\hat{y}_p$  之间的交叉熵来计算损失, 因此, 风险预测的目标函数是交叉熵的平均值:

$$L(\theta) = -\frac{1}{|P|} \sum_{p=1}^{|P|} (y_p^T \log(\hat{y}_p) + (1 - y_p)^T \log(1 - \hat{y}_p)) \tag{11}$$

## 5. 实验

电子病历信息经过词嵌入模型处理, 将离散的元素嵌入到连续向量空间中, 将相同类型的向量合并为单个向量, 然后经过上述特定类型的池化操作将其表示为一次入院。

本文针对疾病风险预测的研究, 因此主要考察模型在慢性病上的性能。实验选取基线为, 传统机

器学习方法随机森林(RF)和深度学习方法的卷积神经网络(CNN), 通过不同模型性能评估方法与本文提出的基于深度学习方法长短期记忆网络(LSTM)形成比对。实验数据选取现有的心脏疾病真实医疗数据集。实验设置医疗代码总数为 $|c|=1016$ , 数据集按 0.75:0.10:0.15 的比例随机划分为训练集、验证集和测试集。将基于 NLP 中词嵌入模型处理的数据作为上述疾病风险预测模型的输入, 通过上述三种模型进行疾病风险的预测, 通过对不同模型的性能评估数据进行比对, 分析本文提出基于 LSTM 模型的优势。

在深度学习模型中, 因为引入了参数矩阵, 可能会导致过拟合。因此, 引入 L2 范数和 Dropout = 0.5 来防止过拟合。L2 范数正则化, 也称为“权重衰减”, 用于防止权重参数出现极值。引入一个常数  $\lambda$  来控制正则化的大小。Dropout 是 DNN 的一种正则化方法。训练期间, 使用预先定义的概率  $1 - p$  (Dropout 率)删除单元, 其余部分像往常一样通过反向传播进行训练。这就阻止了各单元之间的相互适应, 从而防止过拟合。

### 5.1. 实验设置

实验在 windows10 操作系统, python3.7.3, 基于 pytorch 框架, 运用能够处理长时间依赖性的 LSTM 模型, 参数设置如下表 1 所示和参数设置如下表 2 所示的普通卷积神经网络(CNN)模型上运行。

**Table 1.** Parameter settings on the prediction model LSTM

**表 1.** 预测模型 LSTM 上的参数设置

Bitch_size	32	Visit_size	256
L2_reg	0.01	Hidden_size	128
Droupout_rate	0.05	Cuda_id	0
Log_eps	1e-8	Use_gpu	False

其中, LSTM 采用 Hidden\_size 即隐层为 128。

**Table 2.** Parameter settings on prediction model CNN

**表 2.** 预测模型 CNN 上的参数设置

参数名称	参数值	参数名称	参数值
Bitch_size	32	N_filter	16
L2_reg	0.01	Fiter_hs	[2,3,4,5]
Droupout_rate	0.05	Cuda_id	0
Log_eps	1e-8	Use_gpu	False

其中, 对于 CNN 采用 N\_filter 即滤波器的数量为 16, Filter\_hs 即滤波的大小采用大小不等的 2, 3, 4, 5 进行特征提取。

### 5.2. 实验结果与分析

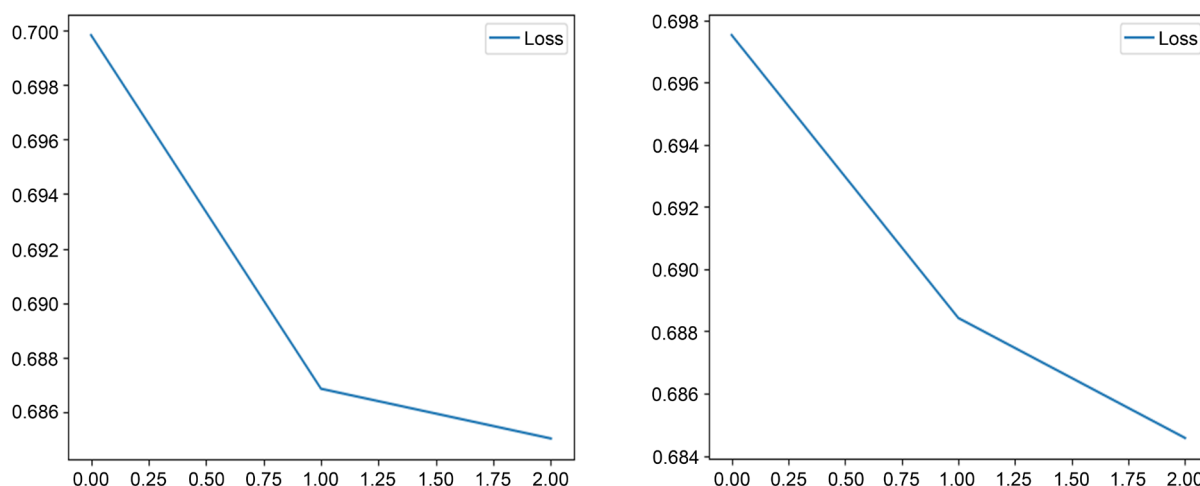
对模型运行 N\_epoch = 10 即迭代 10 次如表 3 所示。



**Table 3.** Experimental results  
**表 3.** 实验结果

n_epoch	mean_cost	test_cnn_loss	mean_cost	test_lstm_loss
0	0.700642	0.664165	0.697237	0.694940
1	0.691573	0.659424	0.693990	0.688363
2	0.686926	0.659424	0.687902	0.688363
3	0.671121	0.659425	0.685048	0.671891
4	0.656447	0.659425	0.683142	0.659692
5	0.655964	0.659957	0.682925	0.649857
6	0.620882	0.659957	0.676765	0.641257
7	0.614401	0.657919	0.674915	0.641257
8	0.567315	0.643577	0.662062	0.627468
9	0.583820	0.648483	0.663815	0.608677

表 3 所示，mean\_cost 分别为 CNN 模型和 LSTM 模型的训练集的平均损失，通过交叉熵函数得到 CNN 和 LSTM 模型的测试损失如上表所示，选取两模型最小损失值进行对比，LSTM 模型明显优于 CNN；此外，CNN 模型随着 epoch 的增大，mean\_cost 减小但测试损失并未明显减小，而 LSTM 模型表现良好。



**Figure 2.** Loss image of CCN and LSTM models  
**图 2.** CNN 和 LSTM 模型的损失图像

图 2 显示了 CNN 和 LSTM 模型的损失变化图像。从图 2 中，我们看到，LSTM 模型的损失相比较于 CNN 模型在前一阶段下降的更快，在后一阶段相比较于自身前一阶段虽然有减小，但是下降的速度还是明显优于 CNN 模型，是表 3 数据的可视化显示，进一步说明本文方法的优势。

表 4 呈现三种模型使用精确度、F1 分数和 AUC 作为模型评估指标的结果。我们可以观察到，与两种基线相比，所提出的方法在所有度量值方面都取得了最佳性能。

**Table 4.** Performance evaluation of the three models**表 4.** 三种模型的性能评估

模型	性能评估指标		
	Accuracy	F1 分数	AUC
RF	0.9137	0.8444	0.8755
CNN	0.8712	0.9260	0.8994
LSTM	0.9339	0.8827	0.9034

## 6. 结论与展望

本文通过对电子病历的处理, 实现了入院信息的表示, 解决了电子病历时间不规律性, 通过 LSTM 模型对时序信息的处理特征实现了对电子病历信息的长时间依赖性的处理, 通过与基线模型, 传统机器学习方法随机森林和深度学习方法 CNN 进行实验比对, 结果显示了本文方法的有效性。

未来, 我们将继续针对疾病风险预测进行进一步的研究, 努力克服 LSTM 在处理长久未来方面的缺陷, 更好的实现对疾病风险的预测。

## 基金项目

国家自然科学基金项目(61972165); 福建省高校创新团队发展计划, 福建省研究生导师团队, 泉州市高层次人才团队项目(2017ZT012)。

## 参考文献

- [1] Jensen, P.B., Jensen, L.J. and Brunak, S. (2012) Mining Electronic Health Records: Towards Better Research Applications and Clinical Care. *Nature Reviews Genetics*, **13**, 395-405. <https://doi.org/10.1038/nrg3208>
- [2] Caroprese, L., Veltri, P., Vocaturo, E. and Zumpano, E. (2018) Deep Learning Techniques for Electronic Health Record Analysis. 2018 *9th International Conference on Information, Intelligence, Systems and Applications*, Zakynthos, Greece, 23-25 July 2018, 1-4. <https://doi.org/10.1109/IISA.2018.8633647>
- [3] Madsen, L.B. (2014) Data-Driven Healthcare: How Analytics and BI Are Transforming the Industry. *MIT Technology Review Business Report*, **117**, 1-19.
- [4] Madsen, L.B. (2014) Data-Driven Healthcare: How Analytics and BI Are Transforming the Industry. Wiley, New York.
- [5] Ho, J.C., Ghosh, J. and Sun, J.M. (2014) Marble: High-Throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*, 115-124.
- [6] Cheng, Y., Wang, F., Zhang, P. and Hu, J.Y. (2016) Risk Prediction with Electronic Health Records: A Deep Learning Approach. *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM'16)*, Miami, FL, 5-7 May 2016, 432-440. <https://doi.org/10.1137/1.9781611974348.49>
- [7] Zhou, J., Wang, F., Hu, J. and Ye, J. (2014) From Micro to Macro: Data Driven Phenotyping by Densification Longitudinal Electronic Medical Records. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, 135-144. <https://doi.org/10.1145/2623330.2623711>
- [8] Wang, F., Zhang, P., Qian, B., Wang, X. and Davidson, I. (2014) Clinical Risk Prediction with Multilinear Sparse Logistic Regression. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, 145-154. <https://doi.org/10.1145/2623330.2623755>
- [9] Chen, L. (2019) Assertion Detection in Clinical Natural Language Processing: A Knowledge-Poor Machine Learning Approach. 2019 *IEEE 2nd International Conference on Information and Computer Technologies*, Kahului, HI, 14-17 March 2019, 37-40. <https://doi.org/10.1109/INFOCT.2019.8710921>
- [10] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [11] Rumelhart, D.E. (1986) Learning Representations by Back-Propagating Errors. *Nature*, **323**, 533-536.



---

<https://doi.org/10.1038/323533a0>

- [12] Choi, E., Bahadori, M.T., Searles, E., *et al.* (2016) Multi-Layer Representation Learning for Medical Concepts. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2016, 1495-1504. <https://doi.org/10.1145/2939672.2939823>
- [13] Choi, E., Bahadori, M., Song, L., *et al.* (2016) GRAM: Graph-Based Attention Model for Healthcare Representation Learning. *Computer Science*, 787-795.
- [14] Ma, F.L., Radha, C. and Zhou, J. (2017) Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2017, 1903-1911. <https://doi.org/10.1145/3097983.3098088>
- [15] Che, Z., Cheng, Y., Zhai, S., Sun, Z. and Liu, Y. (2017) Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records. *2017 IEEE International Conference on Data Mining*, New Orleans, LA, 18-21 November 2017, 787-792. <https://doi.org/10.1109/ICDM.2017.93>
- [16] Choi, E., Bahadori, M.T., Kulas, J.A., Schuetz, A. and Stewart, W.F. (2016) RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism. *Computer Science*, 3504-3512.
- [17] Suo, Q., Ma, F., Yuan, Y., *et al.* (2017) Personalized Disease Prediction Using a CNN-Based Similarity Learning Method. *2017 IEEE International Conference on Bioinformatics and Biomedicine*, Kansas City, MO, 13-16 November 2017, 811-816. <https://doi.org/10.1109/BIBM.2017.8217759>