A Regularized Deep Learning Approach for Clinical Risk Prediction of Acute Coronary Syndrome Using Electronic Health Records

Zhengxing Huang, Wei Dong, Huilong Duan, and Jiquan Liu

***Abstract*—*Objective:* Acute coronary syndrome (ACS), as a common and severe cardiovascular disease, is a leading cause of death and the principal cause of serious long-term disability globally. Clinical risk prediction of ACS is important for early intervention and treatment. Existing ACS risk scoring models are based mainly on a small set of hand-picked risk factors and often dichotomize predictive variables to simplify the score calculation. *Methods:* This study develops a regularized stacked denoising autoencoder (SDAE) model to stratify clinical risks of ACS patients from a large volume of electronic health records (EHR). To capture characteristics of patients at similar risk levels, and preserve the discriminating information across different risk levels, two constraints are added on SDAE to make the reconstructed feature representations contain more risk information of patients, which contribute to a better clinical risk prediction result. *Results:* We validate our approach on a real clinical dataset consisting of 3464 ACS patient samples. The performance of our approach for predicting ACS risk remains robust and reaches 0.868 and 0.73 in terms of both AUC and accuracy, respectively. *Conclusions:* The obtained results show that the proposed approach achieves a competitive performance compared to state-of-the-art models in dealing with the clinical risk prediction problem. In addition, our approach can extract informative risk factors of ACS via a reconstructive learning strategy. Some of these extracted risk factors are not only consistent with existing medical domain knowledge, but also contain suggestive hypotheses that could be validated by further investigations in the medical domain.**

***Index Terms*—Acute coronary syndrome, clinical risk prediction, deep learning, electronic health record, stacked denoising auto-encoder.**

Manuscript received April 24, 2017; revised June 11, 2017 and July 14, 2017; accepted July 20, 2017. Date of publication July 24, 2017; date of current version April 19, 2018. This work was supported in part by the National Nature Science Foundation of China under Grant 61672450 and in part by the National Key Research and Development Program of China under Grant 2016YFC1300303. *(Corresponding author: Jiquan Liu.)*

Z. Huang and H. Duan are with the College of Biomedical Engineering and Instrument Science, Zhejiang University.

W. Dong is with the Department of Cardiology, Chinese PLA General Hospital.

J. Liu is with the College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310027, China (e-mail: liujq@zju.edu.cn).

Digital Object Identifier 10.1109/TBME.2017.2731158

I. Introduction

ACUTE coronary syndrome (ACS), which includes both myocardial infarction (MI) and unstable angina, is the most common type of coronary heart diseases (CHD). It occurs when the heart muscle does not receive sufficient oxygen-rich blood, and thus may cause significant mortality and morbidity after the onset [1]–[3]. ACS has become an important public health issue because of the costs associated with treatment interventions, the prolonged chronic course, and high mortality rates [4]. Every year, ACS is estimated to affect 1.4 million people in the United States, and 2.5 million people in China [5], [6]. Approximately 30% people are at risk of having ACS during their lifetime [6]. In the United States, ACS results in deaths of 400,000 to 500,000 people each year, with about half of those dying before they reach the hospital [5]. To this end, the ability to leverage a quantitative paradigm to alleviate adverse events and improve patient outcomes, both in terms of diagnosis and prevention, could potentially confer significant benefits both to patients and to society [7]–[9].

Clinical risk prediction has been recognized as a critical tool for efficiently managing disease care and treatments [10]–[12]. In particular, clinical risk prediction can help clinicians to estimate the chance of a unfavorable major cardiac event (such as death, myocardial infarction, and requiring revascularization, etc.) during patients' hospitalizations [2], [3], [42], [43], [48] to develop timely and appropriate intervention strategies to those at high risk of adverse cardiac events, to motivate patients to remain adherent to these strategies, and finally to reduce morbidity and mortality among patients in the high-risk group [12]–[14].

Extensive research efforts have been devoted to clinical risk prediction of ACS. Traditionally, many ACS risk-scoring tools, e.g., GRACE [3], and TIMI [2], etc., have been developed based on monitoring population samples over a long period of time. They have been derived from a small set of handpicked coronary risk factors predefined in clinical trial designs. The goal of clinical risk prediction for these models was to regress the risk of patients [2], [3], [12]. Although being useful, these tools have several important limitations [8], [12]–[15], e.g., they are based on a small set of handpicked patient features collected in costly trials, they lack the ability to deal with missing data, and only dichotomize predictive variables to simplify the score calculation, etc. [17].

Recently, the increased availability of electronic health records (EHR) has demonstrated great potential to improve the performance of clinical risk prediction. EHRs regularly record various care and treatment behaviors, e.g., procedures, diagnoses, and laboratory tests and measurements, etc., of patients during their hospitalization within the context of large healthcare systems [46]; this captures the characteristics of heterogeneous populations of patients receiving care in their current clinical setting [15], [18], [19]. There is an excellent opportunity to develop more accurate risk prediction models from EHRs. In this context, many machine learning techniques, such as Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM), etc., have been applied to address the challenge of clinical risk prediction. From a theoretical perspective, these approaches to producing a clinical risk prediction model are to collect a set of training patient samples that are manually reviewed and confirmed by clinicians by referring to actual occurring adverse events during the patients' hospitalizations, and then applying a supervised learning algorithm to building a classifier for clinical risk prediction [16].

Effective prediction of clinical risks of ACS patients via their heterogeneous EHR data is still an intricate problem and remains a major challenge for healthcare management, mainly due to high clinical complexity and the natural heterogeneity in EHR data [12], [21], [22]. Therefore, one of the most important tasks in clinical risk prediction is to develop robust prediction models that can effectively handle high dimensional heterogeneous EHR data and accurately classify different clinical risks levels based on the acquired EHR data.

Recent research devoted to developing learning algorithms for deep architectures has demonstrated impressive prediction results in several areas [23]–[25]. Unlike the traditional classifiers, such as SVM and LR that typically have effectively one or two layers, deep learning architectures with a greater number of layers, can potentially extract abstract and invariant features for better performance of patient classification [25]. The ability of inference on a large volume of heterogeneous EHR data is particularly suitable for our aim. Therefore, this paper proposes a novel approach for clinical risk prediction of ACS based on deep learning. Among various deep learning models, the Stacked Denoising Auto-encoder (SDAE) has particular advantages such as rapid inference and the ability to reconstruct features (a.k.a. clinical risk factors) yielding good classification accuracy [26]. Thus, SDEA is adopted to address the problem of clinical risk prediction in this study. Specifically, given the complex and high-dimensional EHR data available at hand, we first developed an appropriate feature representation by training a SDAE model for patient individuals, such that a representation of ACS risk factors can be identified and the salient patient features can be disentangled from a large volume of EHR data. In order to ensure that features reconstructed by SDAE can finally be useful to the clinical risk prediction problem, two regularization constraints that preserve actual risk information of patients are added on SDAE. One constraint penalizes a derivation of feature representations in the same risk level, to make the reconstructed features of patients within the same risk level as close as possible, and the other penalizes a derivation of representations in different risk levels, to make the reconstructed features of patients at the different risk levels as separated as possible. After this feature reconstruction learning phase, we append a softmax regression layer on the top of the resulting reconstructed feature representation layer, which is tailored to the clinical risk prediction problem. Comparing with both traditional SDAE without incorporating risk information into learning and conventional classification algorithms, our proposed model can learn more discriminative patient feature representations and thus improve the performance of clinical risk prediction.

To evaluate the proposed approach, extensive experiments using an EHR dataset consisting of 3,464 ACS patient samples and collected from the Cardiology department of the Chinese PLA General Hospital were carried out. Specifically, we extracted patient's information from admission records to construct clinical risk prediction models for ACS patients. As shown in Fig. 1, an admission record sample contains valuable patient information such as demographics, medical history, physical examination results, lab test and specific inspection, and the first primary diagnosis and common comorbidities, etc [44]. In clinical practice, physicians often refer to the admission record of the ACS patient to determine his or her clinical risk [47]. In this sense, we propose to utilize admission records to construct our clinical risk prediction model and learn informative ACS risk factors for helping physicians predict clinical risks for ACS patients at the early stage of their hospitalizations, so as to make the practice better for the care of individual patients. The approach was compared with the regular SDAE without efforts on incorporating clinical risk information into learning, and five state-of-the-art classification algorithms: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Multi-Layer Perceptron neural networks (MLP), and Naive Bayes (NB). The experimental results show that our approach obtains a competitive performance in comparison with benchmark methods.

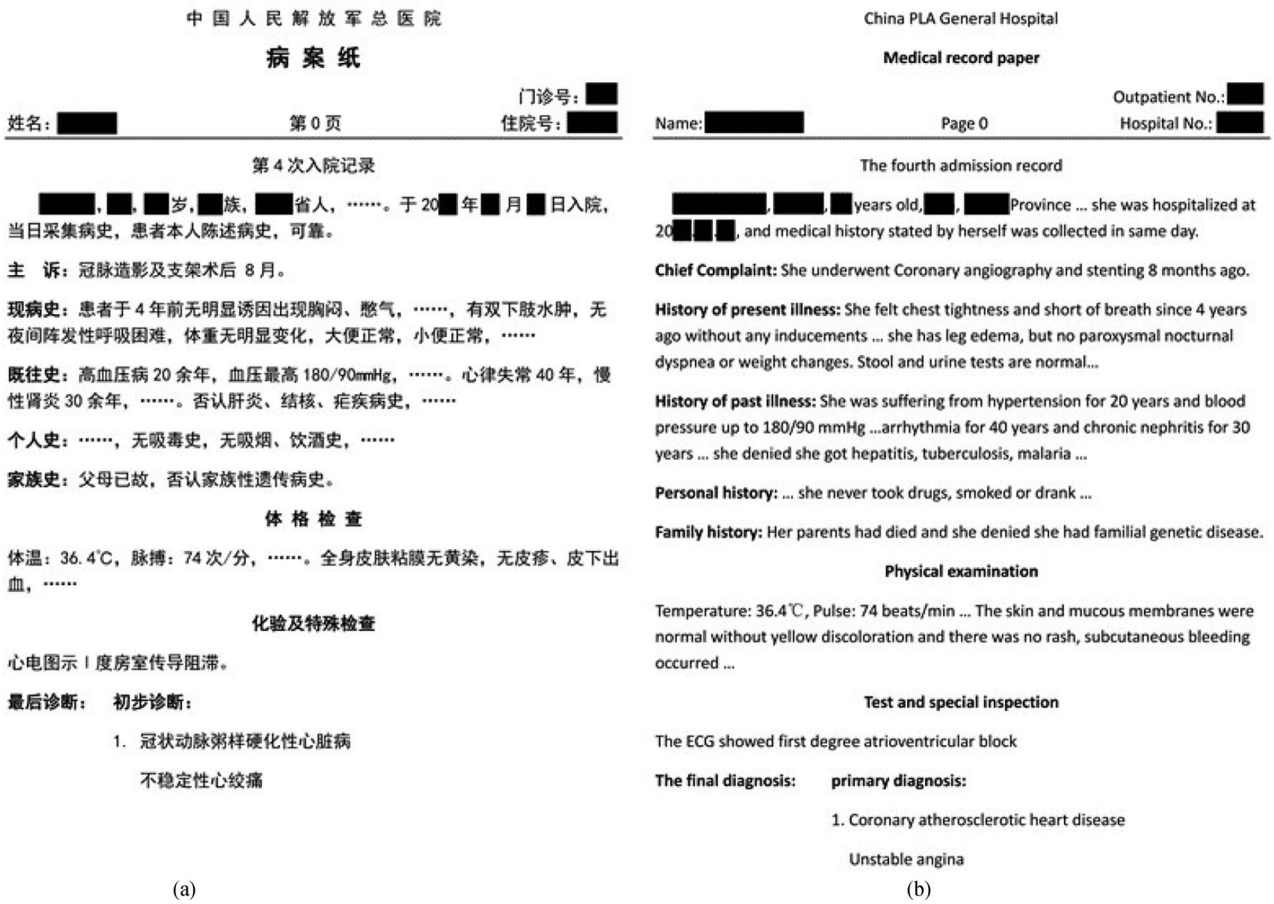
The remainder of this paper is organized as follows. Section II presents the background of this study. Detailed descriptions of the proposed approach are presented in Section III. Experiments on a real clinical dataset are designed, and the performance of our method is compared with the state-of-the-art methods in Section IV. Finally, conclusions and future directions are drawn in Section V.

II. Background

In this section, we briefly introduce the background that is closely related to the research proposed in this paper. One is clinical risk prediction from EHRs. The other is the methodology research on deep learning and its applications in biomedicine.

A. Clinical Risk Prediction

Clinical risk prediction models are increasingly used in healthcare for a wide range of applications. In primary care, they may be used to target interventions by identifying patients with higher risk of diseases such as ACS. In practice, a number of recent studies have been developed with validated models to enable clinicians to reliably identify patients at low, medium, and high risk for ACS [12]. For example, the GRACE risk



中国人民解放军总医院

病案纸

门诊号：[redacted]

姓名：[redacted] 第 0 页 住院号：[redacted]

第 4 次入院记录

[redacted]，[redacted]，[redacted]岁，[redacted]族，[redacted]省人，……。于 20[redacted]年[redacted]月[redacted]日入院，当日采集病史，患者本人陈述病史，可靠。

主　诉：冠脉造影及支架术后 8 月。

现病史：患者于 4 年前无明显诱因出现胸闷、憋气，……，有双下肢水肿，无夜间阵发性呼吸困难，体重无明显变化，大便正常，小便正常，……

既往史：高血压病 20 余年，血压最高 180/90mmHg，……。心律失常 40 年，慢性肾炎 30 余年，……。否认肝炎、结核、疟疾病史，……

个人史：……，无吸毒史，无吸烟、饮酒史，……

家族史：父母已故，否认家族性遗传病史。

体 格 检 查

体温：36.4℃，脉搏：74 次/分，……。全身皮肤粘膜无黄染，无皮疹、皮下出血，……

化验及特殊检查

心电图示Ⅰ度房室传导阻滞。

最后诊断：　初步诊断：

1. 冠状动脉粥样硬化性心脏病

   不稳定性心绞痛

(a)

China PLA General Hospital

Medical record paper

Outpatient No.:[redacted]

Name:[redacted] Page 0 Hospital No.:[redacted]

The fourth admission record

[redacted], [redacted], [redacted] years old, [redacted], [redacted] Province ... she was hospitalized at 20[redacted], and medical history stated by herself was collected in same day.

Chief Complaint: She underwent Coronary angiography and stenting 8 months ago.

History of present illness: She felt chest tightness and short of breath since 4 years ago without any inducements ... she has leg edema, but no paroxysmal nocturnal dyspnea or weight changes. Stool and urine tests are normal...

History of past illness: She was suffering from hypertension for 20 years and blood pressure up to 180/90 mmHg ...arrhythmia for 40 years and chronic nephritis for 30 years ... she denied she got hepatitis, tuberculosis, malaria ...

Personal history: ... she never took drugs, smoked or drank ...

Family history: Her parents had died and she denied she had familial genetic disease.

Physical examination

Temperature: 36.4℃, Pulse: 74 beats/min ... The skin and mucous membranes were normal without yellow discoloration and there was no rash, subcutaneous bleeding occurred ...

Test and special inspection

The ECG showed first degree atrioventricular block

The final diagnosis: **primary diagnosis:**

1. Coronary atherosclerotic heart disease

   Unstable angina

(b)

Fig. 1. An example of part of an admission record extracted from a Chinese hospital. (a) Original admission record in Chinese, and (b) its' translation version in English [44].

scoring model, as one of the most popular and well-accepted risk scores of ACS, was developed to predict clinical risk score of individual patients [3]. However, it must be noted that the models along this line have been estimated using a small set of specially chosen patient features from highly-stratified cohorts. In consequence of this, they just account for a small number of hand-picked risk factors and are fragmented: the conclusions only retain under well-controlled conditions [13].

Recently, many advanced data mining algorithms have been introduced for clinical risk prediction, with the wide spread adoption of electronic health records (EHR) [6], [13], [19], [20]. The EHR typically records a diverse set of clinical information, including patient demographics, symptoms, laboratory test results, and treatment behaviors, etc. It, therefore, provides a comprehensive source for clinical risk prediction. Many data mining algorithms, such as decision trees [11], Bayesian networks [12], and fuzzy inference systems [15], etc., have been proposed to explore the potential of EHR data for clinical risk prediction. For example, Tay *et al.* presented a novel neural-inspired algorithm for risk prediction by using EHR [29]. Karaolis *et al*. applied the C4.5 decision tree algorithm to extract essential risk factors of coronary heart events from EHR data [14]. In [19], a hybrid model was developed for automatically identifying risk factors of heart disease in patient EHR. In [34], Bayesian analyses were undertaken to predict human clinical adverse events observed in drug development programs. In our previous work, a genetic fuzzy system was developed for unstable angina risk prediction, and evaluated on a clinical dataset collected from a Chinese hospital [15]. In addition, we previously proposed a probabilistic topic modeling based approach to risk stratification by exploring the potential of EHRs in an unsupervised fashion [12]. The experimental results on an actual clinical dataset demonstrate that our model can generate coherent and informative patient sub-profiles, and sub-profile-specific risk tiers, and thus provide significant potential to be explored for the further tasks, such as patient cohort analysis. Work that is closely related to ours was presented in [30], in which a deep learning approach for clinical risk prediction from EHR was described. Specifically, the authors modeled the EHR record as a longitudinal event matrix, and applied a Convolutional Neural Network model to this event matrix to build a risk prediction model on the sufficient and labeled set of patient samples.

Despite the demonstrated applicability of the aforementioned methods, clinical risk prediction utilizing heterogeneous EHR data remains one of the key challenges to be addressed in the field of medical informatics and healthcare management [29], [31]. In general, the complexity of a risk prediction model increases with the increase in number of patient features and the

heterogeneity of EHR data. It is important to note that the problems faced in clinical risk prediction problem with EHR data are similar to pattern classification problems that have high dimensional data. In the last decade, pattern classification has advanced into a new paradigm with emerging techniques of deep machine learning [24]. The requirements for handling multi-dimensional heterogeneous EHR data for risk prediction and the advantages of deep machine learning techniques for handling high dimensional data have motivated the development of learning techniques. However, advantages of evolving deep machine learning techniques have not yet been fully utilized for processing EHR data. Motivated by these observations, this paper presents a novel clinical risk prediction model based on a specific deep network structure, i.e., SDAE, which is capable of utilizing the benefit of deep machine learning process to handle the complexity of EHR data. To the best of our knowledge, no other existing approaches present the same merits as our model.

B. Deep Learning

Deep learning is a set of algorithms in machine learning that attempt to model high-level abstractions in the data by using model architectures composed of multiple non-linear transformations [26]. Stacking up the nonlinear transformation layers is the general basic idea in deep learning algorithms [26], [28]. The more layers the data goes through within the deep architecture, the more complicated are the nonlinear transformations that are constructed. The final representation of data constructed by the deep learning algorithm provides latent and essential information from the data that can be used as features in building classifiers [24].

Stacked denoising auto-encoder (SDAE), one of the most extensively investigated deep learning architectures, is employed in this study. SDAE is a symmetrical neural network, and mainly used for learning the features from dataset in an unsupervised fashion [28]. Typically, each denoising auto-encoder in SDAE is trained to reconstruct a clean "repaired" input from a corrupted version of it. This is done by first corrupting the initial input $\boldsymbol{x} \in \mathbb{R}^D$ (D is the number of features) into $\tilde{\boldsymbol{x}}$ by means of a stochastic mapping $\tilde{\boldsymbol{x}} \sim \boldsymbol{q_D}(\tilde{\boldsymbol{x}}|\boldsymbol{x})$. And then, corrupted input $\tilde{\boldsymbol{x}}$ is mapped to an internal representation (or code) $f(\tilde{\boldsymbol{x}})$ through the encoder function f, and then maps $f(\tilde{\boldsymbol{x}})$, in a backward manner, to the input space through a decoding function g. The composition of both f and g is called the reconstruction function r, i.e., $r(\boldsymbol{x}) = g(f(\tilde{\boldsymbol{x}}))$, and a reconstruction loss function ℓ penalizes the error made.

The SDAE is useful to learn a hierarchy of features [21] in a greedy layer-wise unsupervised model [28]. The learning process starts to train the first auto-encoder by optimizing the loss function with the original input data to learn the first hidden representation layer. After that, the learned hidden layer is used as the input data for training the next auto-encoder to generate higher-level representations, and this process is repeated with K times, where K is the number of hidden layers.

Since deep learning architectures, such as SDAE, deal with data abstraction and representation, it is quite likely to be suitable for analyzing raw data presented in different formats and/or from different sources. In this sense, deep learning has become a huge tide of technology in the field of big data and artificial intelligence. In particular, the deep structure makes significant breakthroughs on image understanding, speech recognition, natural language processing and many other areas [26], [35], [36]. In the medical informatics domain, deep learning has gradually attracted more attentions and has been adopted in several applications. For example, Liang *et al*. presented a multimodal deep belief network (DBN) to cluster cancer patients from multi-platform observation data [27]. Rahhal *et al*. described a deep learning based approach for the active classification of electrocardiogram signals [28]. Li *et al*. applied a variety of well-trained DBN to discriminate between patients suffering from bone disease and patients without the disease for the purpose of selecting the informative risk factors of the disease [24]. In [32], Chen *et al*. proposed a deep learning approach for phenotyping from EHR, and they applied their model on a specific application of predictive modeling of chronic diseases. In [36], Tran *et al*. present a computational framework to harness EMR with minimal human supervision via a specific type of deep neural network, i.e., restricted Boltzmann machine (RBM). Although various deep learning models have demonstrated promising results in learning good representations, the representations are learned in an unsupervised manner. Consequently, it is not ensure that the reconstructed feature representations by these deep learning models can finally be useful to the supervised tasks, such as clinical risk prediction studied in this paper. On the contrary, we argue that, by incorporating class information into the deep learning models, it can reconstruct more discriminative feature representations which are more useful for the supervised task. Thus, we develop a regularized SDAE for clinical risk prediction of ACS in this paper.

III. Methods

This section details the proposed clinical risk prediction model using deep learning for analyzing a large volume of multi-dimensional heterogeneous EHR data. We firstly formulate the proposed problem, and then propose our learning schema for clinical risk prediction. Moreover, we present the approach to extract potential informative risk factors via the reconstruction learning strategy.

A. Problem Formulation

The dataset is compounded by a substantial amount of EHR. Each piece of EHR, corresponding to a particular ACS patient sample, can be represented as a patient feature vector. The dataset can be represented as a matrix:

$$X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] = \begin{bmatrix} x_{1,1} & \cdots & x_{N,1} \\ \vdots & \ddots & \vdots \\ x_{1,M} & \cdots & x_{N,M} \end{bmatrix} \tag{1}$$

where N is the number of patient samples, and M is the number of patient features in the dataset. Each column of X, corresponding to a particular ACS patient sample $\boldsymbol{x}$, is viewed as a

feature vector in $\mathbb{R}^M$, where the jth coordinate corresponds to the jth patient feature.

Let Y be a set of labels correspond to N training patient samples and is denoted as

$$Y = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N] = \begin{bmatrix} y_{l,1} & \cdots & y_{l,N} \\ y_{m,1} & \ddots & y_{m,N} \\ y_{h,1} & \cdots & y_{h,N} \end{bmatrix} \tag{2}$$

where subscripts l, m and h indicate the low-, medium-, and high-risk levels of ACS patients, respectively. Each column of Y is a vector in $\mathbb{R}^3$, where the jth coordinate corresponds to the jth patient risk-level. We set $y_{j,i} = 1$ if $\boldsymbol{x}_i \in j$th patient risk-level, otherwise, $y_{j,i} = 0$.

We aim to seek the mapping function $f : X \rightarrow Y$ using a set of collected patient samples in order to classify each patient sample $\boldsymbol{x}$ to the corresponding patient category.

B. Pretraining Using SDAE

In this subsection, we discuss the training and validation of our clinical risk prediction model using our proposed approach. Typically, a SDAE model, as a symmetrical neural network, is mainly used for learning the features of a dataset in an unsupervised manner (see Fig. 2(a)) [37]. To build a deep learning architecture with K hidden layers, SDAE is trained in a greedy layer-wise unsupervised mode. Specifically, the learning process starts by training the first denoising auto-encoder in an unsupervised way and with the noisy corrupted version $\tilde{\boldsymbol{x}}$ of the original input data $\boldsymbol{x}$ to obtain the first hidden representation layer $\boldsymbol{h}_1^e$. Then, the reconstruction layer of this denoising auto-encoder is moved to the last second layer $\boldsymbol{h}_1^d$ of the network architecture, and the obtained hidden layer $\boldsymbol{h}_1^e$ is used as the input data for training the next auto-encoder to generate higher-level representations $\boldsymbol{h}_2^e, \cdots, \boldsymbol{h}_k^e$, and so on. After that, the reconstruction process is performed from the higher-level representation $\boldsymbol{h}_K^e$ to $\boldsymbol{h}_K^d, \cdots, \boldsymbol{h}_1^d$ and $\boldsymbol{x'}$ in a backward manner. Once the greedy layer wise pretraining is complete, one can add on the resulting reconstructed feature layer $\boldsymbol{x'}$ a softmax regression layer to perform clinical risk prediction [28]. This yields a deep neural network (DNN), tailored to clinical risk prediction problem (see Fig. 2(b)).

Note that the number of output nodes are set to be the same as the number of input nodes and is equal to the length of the patient sample vector $\boldsymbol{x}$. Given the noise input $\tilde{\boldsymbol{x}}$ of a patient vector $\boldsymbol{x}$, a hidden feature vector $\boldsymbol{h} \in \mathbb{R}^{|\boldsymbol{h}|}$ ($|\boldsymbol{h}|$ indicates the size of $\boldsymbol{h}$) can be encoded from $\tilde{\boldsymbol{x}}$ through the nonlinear activation function f:

$$\boldsymbol{h} = f\left(\boldsymbol{W}^e \tilde{\boldsymbol{x}} + \boldsymbol{b}^e\right), \tag{3}$$

where $\boldsymbol{\theta}^e = \{\boldsymbol{W}^e, \boldsymbol{b}^e\}$ are parameters of the encoder in one DAE, $\boldsymbol{W}^e \in \mathbb{R}^{|\boldsymbol{h}| \times M}$ is the encoder weight matrix, $\boldsymbol{b}^e \in \mathbb{R}^{|\boldsymbol{h}|}$ is the encoder bias vector, and $f(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function [28].

In the decoding phase, the hidden feature vector $\boldsymbol{h}$ is also mapped through a nonlinear activation function to reconstruct the input vector $\boldsymbol{x'}$ as follows:

$$\boldsymbol{x'} = g\left(\boldsymbol{W}^d \boldsymbol{h} + \boldsymbol{b}^d\right), \tag{4}$$

where $g(\cdot)$ is the decoder function, $\boldsymbol{\theta}^d = \{\boldsymbol{W}^d, \boldsymbol{b}^d\}$ are parameters of the decoder in the DAE, $\boldsymbol{W}^d \in \mathbb{R}^{M \times |\boldsymbol{h}|}$ is the decoder weight matrix and $\boldsymbol{b}^d \in \mathbb{R}^M$ is the decoder bias. To reduce the number of parameters, the weights learned for the coding layer are simply tied to the decoding layer, as indicated in literature [24], [38], i.e., $\boldsymbol{W}^d = \boldsymbol{W}^{eT}$.

For each patient sample $\boldsymbol{x}_i$, it is corrupted to $\tilde{\boldsymbol{x}}_i$, mapped to the hidden vector $\boldsymbol{h}_i$, and finally reconstructed as $\boldsymbol{x}_i'$. The parameters are optimized by minimizing the reconstruction error:

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{i=1}^{N} L\left(\boldsymbol{x}_i, \boldsymbol{x}_i'\right) \\ &= \frac{1}{N} \sum_{i=1}^{N} L\left\{\boldsymbol{x}_i, g_{\boldsymbol{\theta}^d}\left[f_{\boldsymbol{\theta}^e}\left(\boldsymbol{x}_i'\right)\right]\right\}, \end{aligned} \tag{5}$$

where $L(\boldsymbol{x}_i, \boldsymbol{x}_i')$ is the loss function, representing the reconstruction error over all N patient samples. The loss function is measured by cross-entropy as follows:

$$L\left(\boldsymbol{x}_i, \boldsymbol{x}_i'\right) = -\sum_{j=1}^{M} \left(x_{ij} \log x_{ij}' + (1 - x_{ij}) \log\left(1 - x_{ij}'\right)\right) \tag{6}$$

To optimize the cost function in (6), we first initialize the parameter vector $\boldsymbol{\theta}$ to small values near zero then we use the second order optimization method called L-BFGS which is a quasi-Newton method based on the BFGS update procedure.

In order to build a deep learning architecture with K hidden layers, the proposed model is trained by gradient descent layer by layer, beginning with the lowest layer of SDAE. Specifically, the learning process starts by training the first DAE in an unsupervised way by optimizing (6) with the noise input $\tilde{\boldsymbol{x}}$ of an ACS patient sample $\boldsymbol{x}$ to obtain the first hidden representation layer $\boldsymbol{h}_1 = f_{\theta^0}(\tilde{\boldsymbol{x}})$, which can be further used as the input data for training the next DAE to generate higher-level representations, and so forth.

C. Preserving Risk Information Into SDAE

Although SDAE has been used as an extractor to deal with classification problems successfully, the features reconstructed by SDAE in an unsupervised manner. On the other side, risk information contained in the training dataset can be incorporated into learning to make the reconstructed feature representations useful for the further task of clinical risk prediction. To this end, we add two specific constraints, i.e., intra-risk-level affinity and inter-risk-level repulsion, on SDAE to properly preserve clinical risk information of training patient samples. Note that these twofold constraints can enforce the reconstructed feature representations of patients in the same risk level to be as close as possible and ones in different risk levels to be separated as much as possible.

Suppose two patient samples $\boldsymbol{x}_m$ and $\boldsymbol{x}_n$ are in the same risk level, we assume that their corresponding encoding

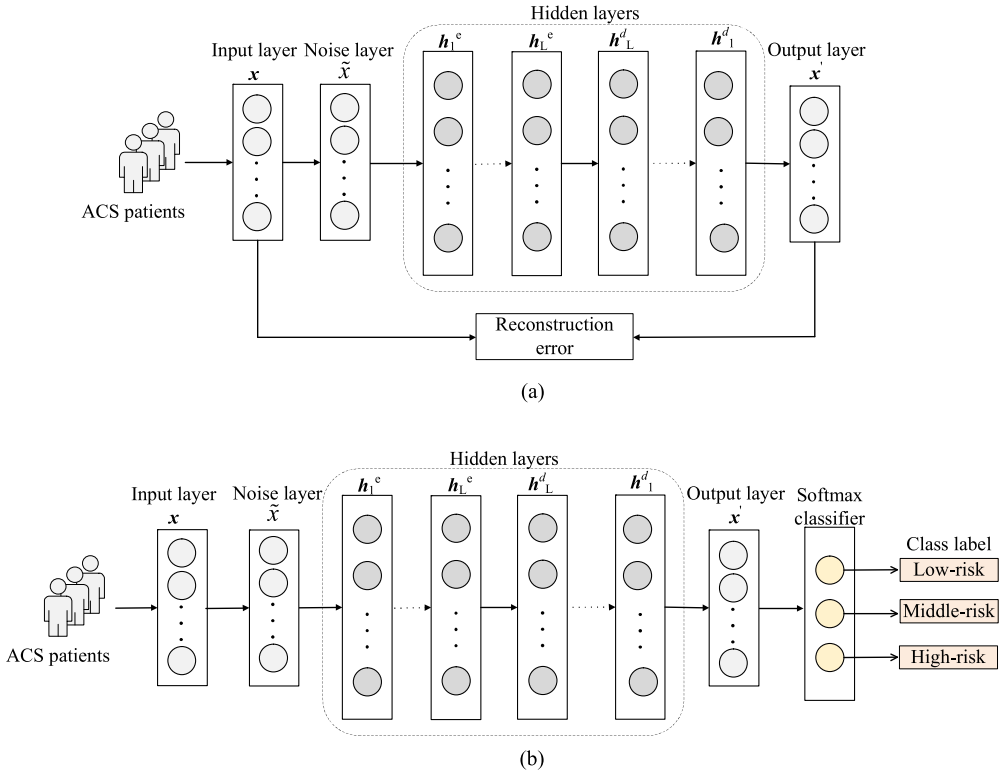


Fig. 2. Using stacked denoising auto-encoder network to learn discriminative feature representations for clinical risk prediction and from heterogeneous EHR data. (a) pre-training using SDAE. (b) Supervised fine-tuning.

features $P(\boldsymbol{h}_{1,m}|\boldsymbol{x}_m)$ and $P(\boldsymbol{h}_{1,n}|\boldsymbol{x}_n)$ are close to each other, and so forth $P(\boldsymbol{h}_{2,m}|\boldsymbol{h}_{1,m})$ and $P(\boldsymbol{h}_{2,n}|\boldsymbol{h}_{1,n})$, $\cdots$, and $P(\boldsymbol{h}_{K,m}|\boldsymbol{h}_{k-1,m})$ and $P(\boldsymbol{h}_{K,n}|\boldsymbol{h}_{k-1,n})$. Thus, the intra-risk-level constraint for the whole SDAE is proposed to minimize the cost function (7):

$$L_{intra} = \frac{1}{2}\sum_{k=1}^{K}\sum_{m,\,n}\mu_{m,n}\,||P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,m}\right) - P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,n}\right)||^2 \tag{7}$$

where $\mu_{m,n} = 1$ if patient samples $\boldsymbol{x}_m$ and $\boldsymbol{x}_n$ are in the same risk level, otherwise $\mu_{m,n} = 0$. $\boldsymbol{h}_{l-1,m}$ and $\boldsymbol{h}_{l-1,n}$ correspond to $\boldsymbol{x}_m$ and $\boldsymbol{x}_n$, respectively.

If two patient samples $\boldsymbol{x}_m$ and $\boldsymbol{x}_n$ are in the different risk levels, we assume that their corresponding encoding features $P(\boldsymbol{h}|\boldsymbol{x}_m)$ and $P(\boldsymbol{h}|\boldsymbol{x}_n)$ are separated to each other. Thus, the inter-risk-level constraint for the whole SDAE is proposed to minimize the cost function (8):

$$L_{inter} = \frac{1}{2}\sum_{k=1}^{K}\sum_{m,\,n}(1-\mu_{m,n}) \cdot \exp\left(-||P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,m}\right) - P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,n}\right)||^2\right) \tag{8}$$

where $\mu_{m,n} = 1$ if patient samples $\boldsymbol{x}_m$ and $\boldsymbol{x}_n$ are in the same risk level, otherwise $\mu_{m,n} = 0$. $\boldsymbol{h}_{k-1,m}$ and $\boldsymbol{h}_{k-1,n}$ correspond to $\boldsymbol{x}_m$ and $\boldsymbol{x}_n$, respectively.

Incorporating both L_{intra} and L_{inter} constraints into learning, we formulate the following objective function (9):

$$\begin{aligned} J'(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{n=1}^{N} L\left(\boldsymbol{x}_n, \boldsymbol{x}'_n\right) + \lambda_{intra}L_{intra} + \lambda_{inter}L_{inter} \\ &= \frac{1}{N}\sum_{n=1}^{N} L\left\{\boldsymbol{x}_n, g_{\boldsymbol{\theta^d}}\left[f_{\boldsymbol{\theta^e}}\left(\boldsymbol{x}'_n\right)\right]\right\} \\ &+ \frac{\lambda_{intra}}{2}\sum_{k=1}^{K}\sum_{m,n}\mu_{m,n}||P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,m}\right) - P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,n}\right)||^2 \\ &+ \frac{\lambda_{inter}}{2}\sum_{k=1}^{K}\sum_{m,\,n}(1-\mu_{m,n}) \\ &\cdot \exp\left(-||P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,m}\right) - P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,n}\right)||^2\right). \end{aligned} \tag{9}$$

where λ_{intra} and λ_{inter} are two regularization parameters for the importance of intra-risk-level and inter-risk-level constraints.

During learning, the gradients of the two regularization terms over the parameters can be calculated by (10)–(13): shown at bottom of the next page.

The proposed regularized SDAE keeps in memory of the characteristics of patient risk information during learning, and thus it has the ability to enforce the reconstructed feature representations within the same risk level to be as close as possible and the reconstructed feature representations between different

risk levels to be kept distant as much as possible. We apply this regularized SDAE to pre-train a clinical risk prediction model from EHR data.

D. Supervised Fine Tuning

Once the pretraining is complete, we append a softmax regression layer on the top of the reconstructed feature representation layer to construct a deep neural network, named regularized stacked denoising auto-encoder with softmax regression model (RSDAE-SM), and the use this RSDAE-SM to perform clinical risk prediction task (see Fig. 2(b)). Specifically, we fine tune the constructed RSDAE-SM using backpropagation by minimizing the cross entropy loss using (14) for the softmax layer [45].

$$L_{fine-tuning}\left(\boldsymbol{\theta}_{RSDAE-SM}\right) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}\mathrm{I}\left(y_i = r\right)\log\left(\frac{\exp\left(\boldsymbol{h}_{\boldsymbol{\theta}_{RSDAE-SM}}\left(\boldsymbol{x}_i\right)\right)}{\sum_{r=1}^{R}\exp\left(\boldsymbol{h}_{\boldsymbol{\theta}_{RSDAE-SM}}\left(\boldsymbol{x}_i\right)\right)}\right) \tag{14}$$

where $\mathrm{I}(\cdot)$ is an indication function that takes 1 if the statement is true, otherwise it takes 0. R indicates the number of risk levels, and $\boldsymbol{h}_{\boldsymbol{\theta}_{RSDAE-SM}}\left(\boldsymbol{x}_i\right)$ is the output of RSDAE-SM for an input $\boldsymbol{x}_i$. The estimation of the vector of parameters $\boldsymbol{\theta}_{\mathrm{RSDAE-SM}} = \{\boldsymbol{W}_1^e, \cdots, \boldsymbol{W}_L^e, \boldsymbol{W}_L^d, \cdots, \boldsymbol{W}_1^d, \boldsymbol{b}_1^e, \cdots, \boldsymbol{b}_L^e, \boldsymbol{b}_L^d, \cdots, \boldsymbol{b}_1^d, \boldsymbol{W}_{softmax}\}$ of RSDAE-SM starts by initializing the weights $\boldsymbol{W}_{softmax}$ of the softmax layer to small random values whereas the weights of the K hidden layers are initialized by the encoding/decoding weights obtained in the pretraining phase. Then, the cost in (14) is minimized with a min-batch gradient descent algorithm.

E. Risk Factor Selection

In this subsection, we propose a risk factor selection strategy to identify informative risk factors for ACS patients within different risk levels. On the one hand, we assume that patient features with lower reconstruction errors are more re-constructible, and more re-constructible patient features are more likely to bear the underlying characteristics. In this regard, given a patient sample within a specific risk level, we expect a small reconstruction error between the original data and the reconstructed one can be obtained by the learned regularized SDAE model. On the other hand, we argue that all different risk-level patient samples behave differently with respect to their risk factors so as to produce a large reconstructed error. Thus, we expect a large reconstruction error during the discriminative learning due to the mismatch of an input patient sample and the model used for testing.

In this context, Fig. 3 shows our approach to selecting potential risk factors that are more re-constructible as the discriminative features.

Formally, the reconstructed error of a particular feature i $(i = 1, \cdots, M)$ with respect to a specific risk level $r \in R$, $(R = \{\mathrm{low-risk},\ \mathrm{medium-risk},\ \mathrm{high-risk}\})$ can be calculated by (15):

$$e_{ri} = \sqrt{\frac{\sum_{n=1}^{N_r}\left(x'_{ni} - x_{ni}\right)^2}{N_r}} \tag{15}$$

$$\frac{\partial L_{intra}}{\partial \boldsymbol{w}} = \frac{\partial \sum_{k=1}^{K}\sum_{m,n}\mu_{m,n}\left|\left|P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,m}\right) - P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,n}\right)\right|\right|^2}{\partial \boldsymbol{w}}$$
$$= \sum_{k=1}^{K}\sum_{m,n}\mu_{m,n}\cdot\left(h_{k,j}^m - h_{k,j}^n\right)\cdot\left(p_{k,j}^m\cdot\left(1-p_{k,j}^m\right)\cdot h_{k-1,i}^m - p_{k,j}^n\cdot\left(1-p_{k,j}^n\right)\cdot h_{k-1,i}^n\right) \tag{10}$$

$$\frac{\partial L_{intra}}{\partial \boldsymbol{b}} = \frac{\partial \sum_{k=1}^{K}\sum_{m,n}\mu_{m,n}\left|\left|P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,m}\right) - P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,n}\right)\right|\right|^2}{\partial \boldsymbol{b}}$$
$$= \sum_{k=1}^{K}\sum_{m,n}\mu_{m,n}\cdot\left(h_{k,j}^m - h_{k,j}^n\right)\cdot\left(p_{k,j}^m\cdot\left(1-p_{k,j}^m\right) - p_{k,j}^n\cdot\left(1-p_{k,j}^n\right)\right) \tag{11}$$

$$\frac{\partial L_{inter}}{\partial \boldsymbol{w}} = \frac{\partial \frac{1}{2}\sum_{k=1}^{K}\sum_{m,n}\left(1-\mu_{m,n}\right)\cdot\exp\left(-\left|\left|P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,m}\right) - P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,n}\right)\right|\right|^2\right)}{\partial \boldsymbol{w}}$$
$$= -\sum_{k=1}^{K}\sum_{m,n}\left(1-\mu_{m,n}\right)\cdot\exp\left(-\sum\nolimits_j\left(h_{k,j}^m - h_{k,j}^n\right)^2\right)\cdot\left(h_{k,j}^m - h_{k,j}^n\right)$$
$$\cdot\left(p_{k,j}^m\cdot\left(1-p_{k,j}^m\right)\cdot h_{k-1,i}^m - p_{k,j}^n\cdot\left(1-p_{k,j}^n\right)\cdot h_{k-1,i}^n\right) \tag{12}$$

$$\frac{\partial L_{inter}}{\partial \boldsymbol{b}} = \frac{\partial \sum_{k=1}^{K}\frac{1}{2}\sum_{m,n}\left(1-\mu_{m,n}\right)\cdot\exp\left(-\left|\left|P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,m}\right) - P\left(\boldsymbol{h}_k|\boldsymbol{h}_{k-1,n}\right)\right|\right|^2\right)}{\partial \boldsymbol{b}}$$
$$= -\sum_{k=1}^{K}\sum_{m,n}\left(1-\mu_{m,n}\right)\cdot\exp\left(-\sum\nolimits_j\left(h_{k,j}^m - h_{k,j}^n\right)^2\right)\cdot\left(h_{k,j}^m - h_{k,j}^n\right)\cdot\left(p_{k,j}^m\cdot\left(1-p_{k,j}^m\right) - p_{k,j}^n\cdot\left(1-p_{k,j}^n\right)\right) \tag{13}$$

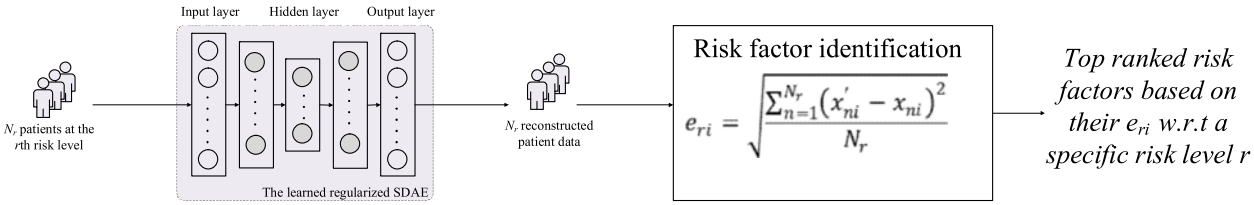


Fig. 3. Using reconstruction error to select informative risk factors for ACS patients at the rth risk level.

where x_{ni} is the ith feature value of the nth patient. (15) calculates the reconstruction error between the reconstructed value of patient feature x'_{ni}, and the original value of feature x_{ni} for total N_r patient samples at the rth risk level. Specifically, patients with positive large e_{ri} are likely at the rth risk level, and those with negative large e_{ri} are unlikely otherwise.

IV. Experiments

In this section, we describe several experiments that were conducted to compare the performance of the proposed model with start-of-the-art methods on the clinical risk prediction task.

A. Experimental Data

The study sample consisted of 3463 patients admitted to the Chinese PLA General Hospital with a discharge diagnosis of ACS. ACS was defined as presentation with symptoms of ischemia along with qualifying electrocardiographic changes, positive cardiac enzymes, new documentation of coronary artery disease (CAD) or prior existence of CAD. The mean length of stay of ACS patient samples was 8.30 days, while some patients staying a very short time, e.g., only 1 day in the hospital, and others take more than 3 months in the hospital. It implicitly indicates the diversity of patient conditions requiring their hospitalizations. As suggested by our clinical collaborators and to provide risk prediction service in an early stage of hospitalizations of ACS patients, we extracted the experimental dataset from admission records of patients from the EHR system.

Preprocessing was performed on the experimental dataset. Specifically, for patient features with binary values, we map their binary values into 0 and 1. For example, regarding the feature "Gender", we assume that 0 indicates "male" and 1 indicates "female". In addition, for patient features with multiple categories, we simply spitted them into multiple categorical variables, and each variable has binary values to indicate if the patient case has this feature or not. For instance, if the cardiac function rating of an ACS patient is in "III"-level, we set the feature "Cardiac function–III" as 1, otherwise, "Cardiac function–III" is set as 0. Moreover, for real-valued patient features, we stacked them into the scope [0, 1], where 0 and 1 correspond to the minimum and the maximum value of a patient feature in the whole dataset. Thus, the information of real-valued patient features is retained in model learning.

We collected a data-set included 348,758 unique patient feature-value pairs of 565 feature types. Ideally, the experimental dataset should have $3463 * 565 = 1{,}956{,}595$ unique patient feature-value pairs. It indicates that the experimental dataset is very sparse, and there are a lot of missing values in the dataset. To remedy it, both patient samples and variables with more than 30% missing values were not included in the analysis. As a result, 3247 patient samples with 200 feature types were left in the experimental dataset. Except this, no further explicit attempts have been performed to handle the missing-data in the experiments. Note that one merit of SDAE is to reconstruct feature representations. It means SDAE can automatically and implicitly impute the missing-data after reconstruction. Thus, we argue that our model can ease the problem of data sparsity to some extent. More detailed information about the experimental dataset, including top-20 ranked patient features that were frequently recorded in EHRs, and their distributions with regard to risk information of ACS patients, are shown in Table I.

In order to establish the ground-truth, we asked three experienced physicians in the Cardiology department of Chinese PLA General Hospital to stratify the risks of patients into three levels, i.e., high-, medium-, and low-risk, based on their EHRs adopting a majority voting. Then, we can check the consistency of the possible risk tiers suggested by our models with the ground-truth. Table I shows the class distribution of patient cases. It can be seen that high-risk patient cases (typically the class of interest) are highly outnumbered by the low-risk patient cases that are commonly found in clinical practice.

B. Evaluation Metrics

Accuracy is commonly used as the evaluation method for the prediction performance. However, for most skewed medical data-sets, the accuracy can be high when misclassifying the entire minority samples to the class of majority. As shown in Table I, the number of patients at low-risk levels is much greater than that of high-risk levels. To this end, an alternative measurement, i.e., the area under the receiver operator characteristic curve (AUC) is used in the proposed research. In general terms, the receiver operator characteristic curve plots the true positive rate (sensitivity) against the false positive rate (specificity). Note that AUC is a quantitative summary of the power of the employed test and it varies from 0 to 1. In other words, an area close to 1 denotes high power, whereas an area below 0.50 means that the method is not able to identify clinical risk levels of ACS patients. A good method should produce high AUC values for ACS risk prediction. In this study, AUC, as a typical measure of prediction accuracy, was plotted to capture how the number of correctly classified high-risk cases varied with the number of low risk patients in the sample that were incorrectly classified as high-risk cases. In our experiments, we adopted

TABLE I
THE DETAILS OF THE EXPERIMENTAL DATASET

	Low-risk	Medium-risk	High-risk	All
Number	1771	1033	443	3247
Age (mean ± sdv.) [min-max]	55.94 ± 9.44 [28–79]	69.92 ± 8.92 [44–91]	76.42 ± 7.40 [49–93]	63.18 ± 12.18 [28–93]
Gender(M/F)	1310 (74.0%) /461 (26.0%)	674 (65.2%) /359 (34.8%)	318 (71.8%) /125 (28.2%)	2302 (70.9%) /945 (29.1%)
Body temperature (mean ± sdv.) [min-max]	36.22 ± 0.31 [35.0–37.8]	36.23 ± 0.34 [35.2–37.9]	36.23 ± 0.39 [35.1–39.3]	36.22 ± 0.33 [35.0–39.3]
Pulse (mean ± sdv.) [min-max]	73.31 ± 9.14 [42–109]	73.19 ± 10.22 [45–125]	75.22 ± 12.28 [41–125]	73.53 ± 9.99 [41–125]
Breathing (mean ± sdv.) [min-max]	18.08 ± 0.55 [12–28]	18.07 ± 0.40 [15–22]	18.13 ± 0.70 [16–26]	18.08 ± 0.534 [12–28]
Systolic blood pressure (mean ± sdv.) [min-max]	134.19 ± 18.69 [84–240]	131.09 ± 16.28 [90–210]	127.25 ± 17.40 [11–198]	132.26 ± 17.94 [11–240]
Diastolic blood pressure (mean ± sdv.) [min-max]	80.12 ± 10.07 [48–120]	74.83 ± 9.74 [40–120]	72.79 ± 10.53 [35–100]	77.44 ± 70.47 [35–120]
Height (mean ± sdv.) [min-max]	168.16 ± 7.28 [144–188]	165.57 ± 8.00 [145–188]	165.71 ± 8.63 [70–183]	167.00 ± 7.81 [70–188]
Weight (mean ± sdv.) [min-max]	73.91 ± 11.72 [35–129]	68.92 ± 12.10 [32–200]	67.24 ± 11.69 [33–176]	71.41 ± 12.16 [32–200]
Abnormal liver (T/F)	64 (3.6%) /1707 (96.4%)	41 (4.0%) /992 (95.9%)	25 (5.6%) /418 (94.4%)	130 (4.0%) /3117 (96.0%)
Lower extremity edema (T/F)	90 (5.1%) /1681 (94.9%)	132 (12.8%) /901 (87.2%)	98 (22.1%) /345 (77.9%)	320 (9.9%) /2927 (90.1%)
Upper extremity edema (T/F)	13 (0.7%) /1758 (99.3%)	9 (0.9%) /1024 (99.1%)	7 (1.6%) /436 (98.4%)	29 (0.9%) /3218 (99.1%)
Creatinine (mean ± sdv.) [min-max]	73.30 ± 23.5 [29.5–739.4]	82.23 ± 41.77 [31.4–739.4]	106.53 ± 77.21 [35.7–699.8]	80.67 ± 42.19 [29.5–739.4]
Glucose (mean ± sdv.) [min-max]	6.15 ± 2.21 [2.69–28.62]	6.12 ± 2.20 [2.7–18.39]	6.17 ± 2.32 [3.06–20.96]	6.14 ± 2.22 [2.69–28.62]
Abnormal lung (T/F)	96 (5.4%)/1675 (94.6%)	149 (14.4%)/884 (85.6%)	115 (26.0%)/328 (74.0%)	360 (11.1%)/2887 (88.9%)
Alanine aminotransferase (mean ± sdv.) [min-max]	28.46 ± 28.45 [1.7–495.1]	22.51 ± 22.08 [2.7–274.4]	20.23 ± 14.17 [2.2–128.4]	25.42 ± 25.22 [1.7–495.1]
Creatine kinase (mean ± sdv.) [min-max]	82.70 ± 62.14 [9.8–1369.3]	90.03 ± 160.66 [6.2–4651.1]	95.59 ± 116.77 [15.8–1425.1]	86.79 ± 110.50 [6.2–4651.1]
Aspartate aminotransferase (mean ± sdv.) [min-max]	21.41 ± 17.16 [6.3–396.0]	21.40 ± 15.74 [5.8–293.8]	22.06 ± 15.74 [6.2–199.2]	21.50 ± 16.64 [5.8–396.0]
Hepatitis (T/F)	44 (2.5%)/1727 (97.5%)	31 (0.3%)/1002 (99.7%)	29 (6.5%)/414 (93.5%)	104 (3.2%)/3143 (96.8%)
Serum uric acid (mean ± sdv.) [min-max]	338.86 ± 88.23 [126.2–808.3]	335.68 ± 92.3 [127.4–785.1]	377.19 ± 109.79 [106.8–886.5]	343.11 ± 93.76 [106.8–886.5]

the source-code from WEKA[1] to use Mann–Whitney U test to calculate the AUC [39], [40].

C. Experimental Setup

We performed our case study in the Cardiology Department at the Chinese PLA General Hospital by obtaining a prior ethics approval from the data protection committee and the institutional review board of the hospital. It must mention that the patient data was anonymized in this study and in this paper. We conducted all experiments on a Microsoft Surface Pro 3 Compatible PC with an Intel Pentium IV CPU 2.3 GHz, and 8G byte main memory running on Microsoft Windows 10. The proposed model was implemented in JAVA.

[1]http://weka.wikispaces.com/Area+under+the+curve

D. Comparison With State-of-the-Art Supervised Learning Algorithms

We compared the clinical risk prediction performance of the proposed RSDAE-SM model with the regular SDAE appending a softmax layer (SDAE-SM), and five state-of-the-art classification algorithms, i.e., Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Multi-layer perceptron neural networks (MLP), and naive Bayes (NB), as benchmark methods. Note that the objective of this study is to classify directly an ACS patient as low-, medium-, and high-risk and therefore we implement a mapping $\{X, Y\}$ (where $Y = \{\text{low, medium, high}\}$) for each model. To compare the performance of the proposed model with existing approaches, we randomly divided the collected dataset into 5 folds and evaluated all the algorithms using cross-validation. After a full learning process on 4 folds, the withheld test fold patient data was

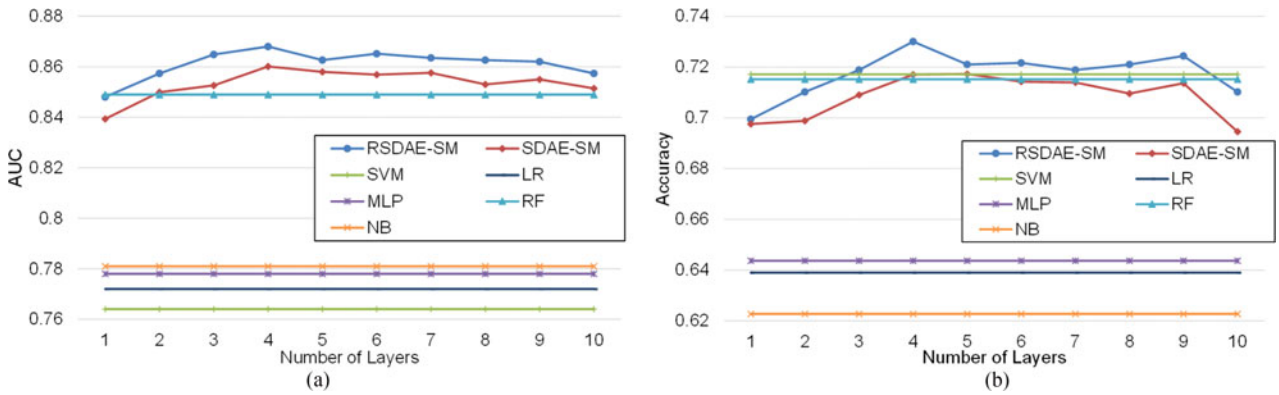


Fig. 4. The (a) AUC and (b) accuracy of the proposed model with varying numbers of hidden layers in comparison with benchmark methods.

presented to the learned classifier for evaluation. The final results were then averaged over five rotations of the test folds.

We first investigated the impact of the number of hidden layers of the proposed RSDAE-SM model on the performance of clinical risk prediction in terms of both AUC and accuracy. The number of hidden layers was varied from 1 to 10 and the performance was computed for each of these, as illustrated in Fig. 4. The results show that, for the proposed RSDAE-SM, the curves of both the AUC and overall accuracy generally increase slightly with the increase in the number of hidden layers from one to 4 hidden layers, and achieve the best performance when $K = 4$. With further increases of K, the curves fluctuate within a certain interval. Because the computation scale of the training dataset is limited, our model is prone to over-fitting when using large numbers of hidden layers. Thus, we set the number of hidden layers to 4 in the following experiments. Note that the number of hidden layers is problem dependent and generally tuned based on the performance on the validation dataset.

In comparison with benchmark approaches, both the proposed RSDAE-SM and SDAE-SM generally performed better when the number of hidden layers is larger than 3. The AUC and overall accuracy for deep learning models were higher than conventional NB, MLP, LR and SVM models for any number of hidden layers of RSDAE-SM, and the AUC and overall accuracy of the RF was higher for up to 1 and 3 hidden layers in the deep learning models, respectively, after which the AUC and overall accuracy of deep learning models were higher. In addition, the proposed RSDAE-SM outperforms SDAE-SM although the latter closely approaches the former with the increases of the number of hidden layers. It indicates that the preservation of risk information into learning by adding regularized constraints into SDAE can indeed improve the performance of clinical risk prediction. In the model with 4 hidden layers, RSDAE-SM had over 0.9%, 2.2%, 13.6%, 12.4%, 11.6% and 11.3% performance gain in comparison with the regular SDAE-SM, RF, SVM, LR, MLP and NB, respectively. And, in terms of accuracy, the proposed model even with just 4 hidden layers had over 1.8%, 2.1%, 1.8%, 14.2%, 13.4% and 17.2% performance gain in comparison with the regular SDAE-SM, RF, SVM, LR, MLP and NB, respectively. Our model could therefore clearly predict patient risks more accurately than the benchmark approaches. This could be contributed by the proposed RSDAE-SM using a deep architecture and preserving risk information of training samples to learn much more and discriminative features from EHR data.

TABLE II
THE PERFORMANCE OF CLINICAL RISK PREDICTION USING 5-FOLD CROSS VALIDATION

	SVM	LR	RF	NB	MLP	SDAE-SM	RSDAE-SM
AUC							
Low	.821	.809	.889	.818	.831	.899	**.909**
Medium	.711	.687	.770	.701	.687	.778	**.790**
High	.662	.820	.876	.823	.777	**.896**	.885
Average	.764	.772	.849	.781	.778	.860	**.868**
Precision							
Low	.817	.719	.773	.669	.758	.778	**.838**
Medium	.586	.492	.594	.502	.507	**.618**	.611
High	**.664**	.495	.657	.481	.420	.580	.606
Average	.723	.616	.700	.590	.632	.700	**.734**
Recall							
Low	**.879**	.829	.934	.905	.819	.778	.876
Medium	**.628**	.430	.539	.232	.449	.618	.613
High	.352	.336	.251	.406	.397	**.580**	.492
Average	.727	.635	.715	.623	.644	.717	**.740**
F1							
Low	.847	.770	.846	.769	.787	.843	**.856**
Medium	.607	.459	.565	.318	.476	.527	**.612**
High	.460	.401	.363	.441	.408	.536	**.543**
Average	.717	.621	.691	.581	.637	.701	**.736**

The prediction performance of both the proposed model and benchmark methods for each risk level is shown in Table II. The average AUC value for the proposed approach (0.868) was the highest of the seven methods. It indicates that our proposed model has good ability to correctly predict patients within the different risk levels. In addition, Table II shows not only the AUC values obtained by the prediction models, but also the precision, recall, and F1 values that are commonly used to illustrate the performance of a classification algorithm. In particular, it appears that our model trades-off high recall with low preci-

TABLE III
STATISTICAL DIFFERENCES BETWEEN BENCHMARK MODELS AND THE PROPOSED MODELS

Model	SVM	LR	RF	MLP	NB	SDAE-SM	RSDAE-SM
SVM	•	•	•	*	*	*	*
LR		•	•	*	*	*	*
RF			•	*	*	*	*
MLP				•	•	**	**
NB					•	**	**
SDAE-SM						•	•
RSDAE-SM							•

**: p-value < 0.01; *: p-value < 0.05; •: p-value > 0.05

sion for F1, which is different from all the other methods. Note that, usually, precision and recall scores are not discussed in isolation. Instead, both are combined into a single measure, e.g., F1, which is the harmonic mean of the two, i.e., the square of the geometric mean divided by the arithmetic mean. Thus, F1 score may be limited in particular circumstances due to the mixing of precision and sensitivity as one evaluation metric. With respect to the clinical risk prediction error, the proposed RSDAE-SM generally achieves the best performance, especially in terms of both AUC and F1, among all of the compared methods, and a strict ordering can also be figured out: RSDAE-SM being better than SDAE-SM being better than other benchmark methods, as shown in Table II.

In order to assess whether there were significant differences in terms of accuracy between the proposed model and baseline methods, pairwise comparisons between each pair of methods were tested using the paired sample t-test [41], and P-values were adjusted for multiple comparisons with the Holm-Bonferroni correction [42], [48]. The paired sample t-test is a statistical procedure used to determine whether the mean difference between two sets of observations is zero. Shifting to our problem, clinical risks of all patient samples were predicted using both our proposed model and baseline methods, resulting in pairs of observations between each pair of methods.

The performance of our approach showed considerable improvements in terms of accuracy in predicting clinical risks of ACS patients and the t-test showed in Table III demonstrated that there are indeed statistically significant differences between our approach and the benchmark methods. As shown in Table III, there are significant differences between the proposed deep learning models and benchmark models – SVM ($P < 0.05$), LR ($P < 0.05$), RF ($P < 0.05$), MLP ($P < 0.01$), and NB ($P < 0.01$) – in terms of clinical risk prediction on the experimental dataset. Combined with the performance comparison results shown in Tables II, this suggests that the proposed deep learning approach achieves a competitive and statistically significant performance in clinical risk prediction in comparison with the state-of-the-art methods.

It must mention that there are correlations between the risk estimates from the various models. For example, both the proposed RSDAE-SM and LR can extract potential risk factors with a large overlap for ACS patients at a specific risk level, although there is a statistically significant difference between them (p-value < 0.05). In fact, these common risk factors are assigned different weights by the different risk models. The experimental results demonstrate that our proposed RSDAE-SM has yielded the better risk prediction performance than the benchmark models, indicating RSDAE-SM can extract potential risk factors with accurate weights, which are more close to the ground truth (i.e., physicians' experiences) in comparison with the baseline models.

TABLE IV
INFORMATIVE RISK FACTORS IDENTIFIED BY THE PROPOSED MODEL

Risk level	Risk factor-value	Discriminative error $e_{r,k}$
Low-risk	**Valve noise–**True	14.903
	Diabetes–False	11.110
	Dizziness–True	10.560
	Lower extremity varicose veins-False	9.480
	Age- 55.94 ± 9.44	9.028
	Angina–True	8.611
	Drink–False	8.338
	Lower extremity edema-False	8.109
	Creatinine-73.30 ± 23.5	7.569
	Arteriosclerosis-False	7.228
Medium-risk	**Age-** 69.92 ± 8.92	12.877
	Valve noise-True	12.446
	Angina-True	9.851
	Abnormal liver-False	9.456
	Drink– False	9.412
	Lung abnormalities-False	8.871
	Qualitative urine test-Positive	8.163
	Chest muffled discomfort-True	7.512
	Creatinine-82.23 ± 41.77	7.354
	Qualitative test of urinary bilirubin-Positive	6.333
High-risk	**Age-**76.42 ± 7.40	9.417
	Chest muffled discomfort - True	8.648
	D-antigen-Positive	8.235
	Creatinine-106.53 ± 77.21	8.032
	Smoking-Yes	7.408
	Angina-True	6.418
	Creatine kinase-95.59 ± 116.77	6.284
	Abnormal liver-Yes	6.247
	Lung abnormalities-Yes	6.237
	Diabetes-True	5.982

E. Informative Risk Factor Selection

The proposed model can extract informative risk factors from EHR data. In this subsection, we report how tried to identify the top 10 risk factors for the patient group within a particular risk level, and confirm their clinical validity with the clinical experts, as shown in Table IV. Note that normalized patient features were categorized into two classes, i.e., 0 or 1 (i.e., the patient case has this feature or not), during the model learning process, as mentioned above. And then, we recovered them and show the actual values of these features in Table IV, to make sure that they are understandable for clinicians.

The most informative risk factors reported in Table IV have been reviewed and endorsed by our clinical collaborators. They

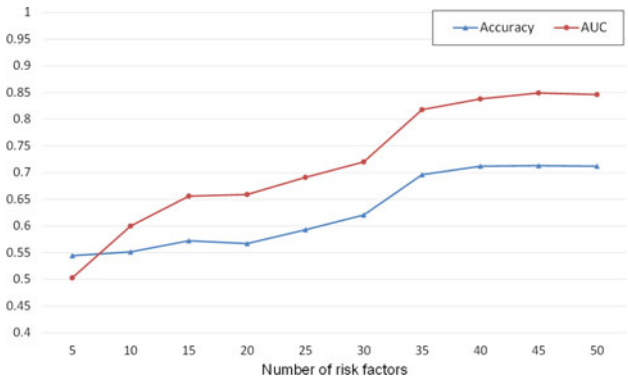


Fig. 5. The impact of the number of risk factors on the performance of ACS risk prediction.

pointed out that these selected patient features are truly informative risk factors, and some of them, such as age, creatinine, smoking status, etc., had already been validated within a clinical cohort study and had been recognized and adopted for ACS risk prediction [2], [3], [9]. The risk factors selected by our model are in bold in Table IV, if they are in accordance with risk factors identified previously in clinical research literature [2], [3], [9]. In addition, clinicians stated that our model can provide specific values of selected risk factors with different risk levels of ACS patients. Note that the extracted values of the same risk factor are different when they are used to indicate different risk levels. For example, the average age of high risk patients is 76.42, which is higher than one of middle risk patients (69.92 years old). Thirdly, our model provides weights of these risk factors at risk levels of patients. This helps to understand the significances of risk factors given different risk levels. Moreover and most interesting, our model may find some new potential risk factors. For example, with regard to the eighth ranked informative risk factor "Abnormal liver - True" of 'high risk' patients, clinicians from the cardiology department of the hospital suggested that this might be a potential risk factor of ACS (personal communication) specific to the Chinese population. Note that China has a large population with hepatitis, there could be potential correlation between ACS and Hepatitis for Chinese people. To the best of our knowledge, this has not previously been indicated in the literature. They will investigate this finding in their clinical study.

To apply our proposed model in clinical practice, we should consider the implementation of a simpler model and investigate the compromise between simplicity and accuracy. In fact, clinicians are interested to know that how many risk factors they need to collect to obtain a good predictive performance. To this end, we apply our risk factor selection strategy to extract the most informative risk factors with regard to each risk level and feed the top-50 in the union of these risk factors directly to our proposed model for ACS risk prediction. As shown in Fig. 5, the curves of both AUC and Accuracy increase with the increase of the number of informative risk factors and then remain stable when the number of risk factors is larger than 35. It indicates that we only need top 35 informative risk factors so as to increase the AUC and Accuracy.

V. Conclusions

This paper proposes a novel learning approach to address the clinical risk prediction problem of ACS from heterogeneous EHR data. In comparison with traditional ACS risk scoring methodologies that relied on a small set of handpicked risk factors, the proposed approach can utilize a large volume of heterogeneous EHR data to construct a robust clinical risk prediction model. The experiments were conducted on a real clinical dataset and results demonstrate that the proposed model can achieve a competitive performance in clinical risk prediction, compared with state-of-the-art classification algorithms. In addition, it has great potential to identify informative risk factors for ACS patients to different risk levels.

Possible future directions to improve our model include enriching the deep learning architecture by having other explicit penalties or constraints on certain properties of the hidden representations such as sparsity and entropy, to avoid over-fitting and to improve the generalization ability of the deep learning network [32], [33]. In addition, and as missing data frequently occur in EHR, we plan to handle the missing-data problem in our future work, taking into account the uncertainty of the predicted clinical risks of ACS patients when missing data are involved. Moreover, instead of using expert opinion (physicians' experiences) to establish the "ground truth" on which to build the proposed model, we plan to conduct a prospective study to observe/collect actual adverse cardiac events of patients during their hospitalizations and after discharge, and then analyze these event data to stratify clinical risk levels of patients. Furthermore, we plan to investigate other state-of-the-art models with a large scale of experiments and compare the performances of different methods on clinical risk prediction. In particular, a large volume of dataset will be collected from our collaborative hospital to be tested for the usefulness and effectiveness of the model.

Acknowledgment

The authors would like to thank the Editor and the anonymous reviewers for their constructive comments on an earlier draft of this paper.

References

[1] J. L. Mega *et al.*, "Rivaroxaban in patients with a recent acute coronary syndrome," *New Engl. J. Med.*, vol. 366, no. 1, pp. 9–19, 2012.
[2] E. M. Antman *et al.*, "The TIMI risk score for unstable angina/non-ST elevation MI: A method for prognostication and therapeutic decision making," *J. Amer. Med. Assoc.*, vol. 284, no. 7, pp. 835–842, 2000.
[3] S. G. Goodman *et al.*, "The expanded global registry of acute coronary events: Baseline characteristics, management practices, and hospital outcomes of patients with acute coronary syndromes," *Amer. Heart J.*, vol. 158, no. 2, pp. 193–201, 2009.
[4] C. J. L. Murray and A. D. Lope, "Global mortality, disability, and the contribution of risk factors: Global burden of disease study," *Lancet*, vol. 349, no. 9063, pp. 1436–1442, 1997.
[5] D. Mozaffarian *et al.*, "American Heart Association Statistics Committee and Stroke Statistics—2015 update: A report from the American Heart Association," *Circulation*, vol. 131, no. 4, pp. e29–322, 2015.
[6] W. Chen *et al.*, *Report on Cardiovascular Disease in China 2014*. Beijing, China: Encyclopedia China Publ. House, 2015.
[7] C. C. Miller *et al.*, *Risk Stratification: A Practical Guide for Clinicians*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[8] P. M. Brindle *et al.*, "The accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: A systematic review," *Heart*, vol. 92, no. 12, pp. 1752–1759, 2006.
[9] D. C. Goff *et al.*, "2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, *Circulation*, vol. 129, pp. S49–S73, 2014.
[10] E. Boersma *et al.*, "Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation. Results from an international trial of 9461 patients," *Circulation*, vol. 101, no. 22, pp. 2557–2567, 2000.
[11] P. W. Wilson *et al.*, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, pp. 1837–1847, 1998.
[12] Z. Huang *et al.*, "A probabilistic topic model for clinical risk stratification from electronic health records," *J. Biomed. Informat.*, vol. 58, pp. 28–36, 2015.
[13] M. Matheny *et al.*, "Systematic review of cardiovascular disease risk assessment tools," Agency for Healthcare Research and Quality, Rockville, MD, USA, Tech. Rep. 11-05155-EF-1, 2011.
[14] M. A. Karaolis *et al.*, "Assessment of the risk factor of coronary heart events based on data mining with decision trees," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 559–567, 2010.
[15] S. Bandyopadhyay *et al.*, "Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data," *Data Mining Knowl. Discovery*, vol. 29, no. 4, pp. 1033–1069, 2015.
[16] S. Paredes *et al.*, "The CardioRisk project: Improvement of cardiovascular risk assessment," *J. Comput. Sci.*, vol. 9, pp. 39–44, 2015.
[17] J. Perk, "European guidelines on cardiovascular disease prevention in clinical practice," *Eur. Heart J.*, vol. 33, pp. 1635–1701, 2012.
[18] Z. Huang *et al.*, "On mining latent treatment patterns from electronic medical records," *Data Mining Knowl. Discovery*, vol. 29, no. 4, pp. 914–949, 2015.
[19] H. Yang and J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," *J. Biomed. Informat.*, vol. 58, pp. S171–S182, 2015.
[20] V. Garla *et al.*, "Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management," *J. Biomed. Informat.*, vol. 46, no. 5, pp. 869–875, 2013.
[21] M. Ranzato *et al.*, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2007.
[22] M. Torii *et al.*, "Risk factor detection for heart disease by applying text analytics in electronic medical records," *J. Biomed. Informat.*, vol. 58, pp. S164–S170, 2015.
[23] G. W. Hruby *et al.*, "Facilitating biomedical researchers' interrogation of electronic health record data: Ideas from outside of biomedical informatics," *J. Biomed. Informat.*, vol. 60, pp. 376–384, 2016.
[24] H. Li *et al.*, "Identifying informative risk factors and predicting bone disease progression via deep belief networks," *Methods*, vol. 69, no. 3, pp. 257–265, 2014.
[25] R. Salakhutdinov and G. E. Hinton, "Learning a nonlinear embedding by preserving class neighborhood structure," *J. Mach. Learn. Res.*, vol. 2, pp. 412–419, 2007.
[26] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.
[27] M. Liang *et al.*, "Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 12, no. 4, pp. 928–939, 2015.
[28] M. M. Al Rahhal *et al.*, "Deep learning approach for active classification of electrocardiogram signals," *Inf. Sci.*, vol. 345, pp. 340–354, 2016.
[29] D. Tay *et al.*, "A novel neural-inspired learning algorithm with application to clinical risk prediction," *J. Biomed. Informat.*, vol. 54, pp. 305–314, 2015.
[30] Yu Chen *et al.*, "Risk prediction with electronic health records: A deep learning approach, in *Proc. 16th SIAM Int. Conf. Data Mining*, 2016.
[31] M. Ranzato *et al.*, "Efficient learning of sparse representations with an energy-based model," in *Advances in Neural Information Processing Systems 19*, B. Scholkopf *et al.*, Eds. Cambridge, MA, USA: MIT Press, 2007, pp. 1137–1144.
[32] A. Ng, "Sparse auto-encoder," CS294A Lecture Notes, vol. 72, 2011.
[33] Z. Huang *et al.*, "Predictive monitoring of clinical pathways," *Expert Syst. Appl.*, vol. 56, pp. 227–241, 2016.
[34] M. Clark, "Prediction of clinical risks by analysis of preclinical and clinical adverse events," *J. Biomed. Informat.*, vol. 54, pp. 167–173, 2015.
[35] W. Xu *et al.*, "SD-MSAEs: Promoter recognition in human genome based on deep feature extraction," *J. Biomed. Informat.*, vol. 61, pp. 55–62, 2016.
[36] T. Tran *et al.*, "Learning vector representation of medical objectives via EMR-driven nonnegative restricted Boltzmann machines (eNRBM)," *J. Biomed. Informat.*, vol. 54, pp. 96–105, 2015.
[37] J. Deng *et al.*, "Autoencoder-based unsupervised domain adaption for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 2535–2543, Sep. 2014.
[38] P. Vincent *et al.*, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, 2010.
[39] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating (ROC) curve characteristic," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[40] S. J. Mason and N. E. Graham, "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation," *Quart. J. Roy. Meteorol. Soc.*, vol. 128, pp. 2145–2166, 2002.
[41] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, pp. 239–281, 2003.
[42] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, pp. 65–70, 1979.
[43] D. A. Cox *et al.*, "Comparative early and late outcomes after primary percutaneous coronary intervention in ST-segment elevation and non-ST-segment elevation acute myocardial infarction (from the CADILLAC Trial)," *Amer. J. Cardiol.*, vol. 98, no. 3, pp. 331–337, 2006.
[44] D. Hu *et al.*, "Utilizing Chinese admission records for MACE prediction of acute coronary syndrome," *Int. J. Environ. Res. Public Health*, vol. 13, no. 9, 2016, Paper 912.
[45] Y. Bengio *et al.*, "Representation Learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[46] P. Y. Wu *et al.*, "Omic and electronic health record big data analytics for precision medicine," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 263–273, Feb. 2017.
[47] Z. Huang *et al.*, "MACE prediction of acute coronary syndrome via boosted resampling classification using electronic medical records," *J. Biomed. Informat.*, vol. 66, pp. 161–170, 2017.
[48] M. T. Cooney *et al.*, "Value and limitations of existing scores for the assessment of cardiovascular risk: A review for clinicians," *J. Amer. Coll. Cardiol.*, vol. 54, no. 14, pp. 1209–1227, 2009.