A Regularized Deep Learning Approach for

Clinical Risk Prediction of Acute Coronary

Syndrome Using Electronic Health Records利用电子病历进行急性冠脉综合征临床风险预测的正规化深度学习方法

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 longterm 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.** **急性冠脉综合征(ACS)是一种常见而严重的心血管疾病，是全球主要的死亡原因和长期严重残疾的主要原因。急性冠脉综合征的临床风险预测对早期干预和治疗具有重要意义。现有的ACS风险评分模型主要基于一小部分精选的风险因素，并且经常将预测变量二分以简化评分计算。 *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. 方法：建立规则化堆叠去噪自动编码器(SDAE)模型，从海量的电子健康记录(EHR)中对ACS患者的临床风险进行分层。为了捕捉相似风险级别患者的特征，并保留不同风险级别之间的区分信息，在SDAE上增加了两个约束，使重建的特征表示包含了更多的患者风险信息，从而有助于更好的临床风险预测结果。*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. 结果：我们在一个包含3464例ACS患者样本的真实临床数据集上验证了我们的方法。我们的方法预测急性冠脉综合征风险的性能仍然稳健，在AUC值和准确度方面分别达到0.868和0.73%。*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.** **结论：所获得的结果表明，与最先进的模型相比，所提出的方法在处理临床风险预测问题上取得了具有竞争力的性能。此外，我们的方法可以通过重构学习策略来提取ACS的信息风险因素。其中一些提取的危险因素不仅与现有的医学领域知识相一致，而且还包含一些可以通过医学领域的进一步调查来验证的提示性假设。**

***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

IEEETRANSACTIONSONBIOMEDICALENGINEERING,VOL.65,NO.5,MAY2018

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

**A**

CUTE 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]. 急性冠状动脉综合征(ACS)包括心肌梗死(MI)和不稳定型心绞痛(UAP)，是最常见的冠心病(CHD)类型。当心肌没有获得足够的富氧血液时，就会发生这种情况，因此在发病后可能会导致显著的死亡率和发病率[1]-[3]。ACS已经成为一个重要的公共卫生问题，因为与治疗干预相关的费用、漫长的慢性病程和高死亡率[4]。据估计，每年美国有140万人受到ACS的影响，中国有250万人[5]、[6]。大约30%的人在有生之年有患ACS的风险[6]。在美国，急性冠脉综合征每年导致40万至50万人死亡，其中约一半人在到达医院前死亡[5]。为此，在诊断和预防方面，利用量化范式减轻不良事件和改善患者结局的能力，可能会给患者和社会带来显著的好处[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 remainadherenttothesestrategies,andfinallytoreducemorbidity and mortality among patients in the high-risk group [12]–[14]. 临床风险预测已被认为是有效管理疾病护理和治疗的关键工具[10]-[12]。特别是，临床风险预测可以帮助临床医生估计发生不利的重大心脏事件(如死亡、心肌梗死和需要重新血运重建等)的可能性。在患者住院期间，[2]、[3]、[42]、[43]、[48]制定针对不良心脏事件高危人群的及时、适当的干预策略，激励患者坚持这些策略，最终降低高危人群[12]-[14]的出血率和死亡率。

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]. 对ACS的临床风险预测进行了广泛的研究。传统上，许多ACS风险评分工具，如GRACE[3]和TIMI[2]等，都是基于对人群样本的长期监测而开发的。它们是从临床试验设计中预先定义的一小组精心挑选的冠状动脉危险因素中衍生出来的。这些模型的临床风险预测的目标是回归患者的风险[2]、[3]、[12]。虽然这些工具很有用，但它们有几个重要的局限性[8]、[12]-[15]，例如，它们基于在昂贵的试验中收集的少量精选患者特征，它们缺乏处理丢失数据的能力，并且只对预测变量进行二分法以简化评分计算，等等。

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 duringtheirhospitalizationwithinthecontextoflargehealthcare 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. 近年来，随着电子健康记录(EHR)可用性的提高，在提高临床风险预测性能方面显示出巨大的潜力。EHR定期记录患者在大型医疗系统中住院期间的各种护理和治疗行为，例如程序、诊断、实验室测试和测量等[46]；这捕捉到了在当前临床环境中接受护理的患者的不同人群的特征[15]、[18]、[19]。这是从EHR开发更准确的风险预测模型的绝佳机会。In this context, many machine learning techniques, such as Logistic Regression (LR), Na¨ıve 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]. 在这种背景下，许多机器学习技术，如Logistic回归(LR)、朴素贝叶斯(NB)和支持向量机(SVM)等已被应用于解决临床风险预测的挑战。从理论的角度来看，这些产生临床风险预测模型的方法是收集一组训练的患者样本，这些样本由临床医生通过参考患者住院期间实际发生的不良事件来人工审查和确认，然后应用有监督的学习算法来构建用于临床风险预测的分类器[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. 通过不同种类的EHR数据有效预测ACS患者的临床风险仍然是一个复杂的问题，并且仍然是医疗保健管理的主要挑战，主要是由于高度的临床复杂性和EHR数据的自然异质性[12]、[21]、[22]。因此，临床风险预测中最重要的任务之一就是开发稳健的预测模型，能够有效地处理高维异构的电子病历数据，并根据获取的电子病历数据准确地对不同的临床风险级别进行分类。

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]. 最近致力于开发深层体系结构的学习算法的研究已经在几个领域展示了令人印象深刻的预测结果[23]-[25]。与传统的分类器(如SVM和LR) 通常只有一层或两层不同，深度学习体系结构具有更多的层，可以潜在地提取抽象和不变的特征，从而提高患者分类的性能[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]. 对大量异构EHR数据的推理能力特别适合我们的目标。因此，本文提出了一种新的基于深度学习的ACS临床风险预测方法。在各种深度学习模型中，堆叠去噪自动编码器(SDAE)具有推理速度快、特征重构能力强等独特优势。临床风险因素)产生良好的分类准确性[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. 因此，本研究采用SDEA来解决临床风险预测问题。具体地说，考虑到手头有复杂的高维EHR数据，我们首先通过训练患者个体的SDAE模型来开发适当的特征表示，以便可以识别ACS危险因素的表示，并且可以从大量的EHR数据中分离出显著的患者特征。为了保证SDAE重建的特征最终能用于临床风险预测问题，在SDAE上增加了两个正则化约束，保留了患者的实际风险信息。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. 一个约束惩罚相同风险级别的特征表示的派生，以使相同风险级别内的患者的重建特征尽可能接近，而另一个约束惩罚不同风险级别的表示的派生，以使不同风险级别的患者的重建特征尽可能地分开。在此特征重建学习阶段之后，我们在得到的重建特征表示层的顶部添加了Softmax回归层，该层是为临床风险预测问题量身定做的。与传统的未将风险信息纳入学习的SDAE和传统的分类算法相比，我们提出的模型可以学习更具区分性的患者特征表示，从而提高了临床风险预测的性能。

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 werecarried 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]. 为了评估所提出的方法，使用从中国人民解放军总医院心内科收集的3464例ACS患者样本的EHR数据集进行了广泛的实验。具体地说，我们从入院记录中提取患者信息来构建ACS患者的临床风险预测模型。如图1所示，入院记录样本包含有价值的患者信息，如人口统计学、病史、体检结果、实验室检查和特殊检查，以及首次诊断和常见合并症等[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. 在临床实践中，医生经常参考ACS患者的入院记录来确定他或她的临床风险[47]。从这个意义上说，我们建议利用入院记录来构建我们的临床风险预测模型，学习信息丰富的ACS危险因素，以帮助医生在ACS患者住院的早期阶段预测其临床风险，从而更好地为个别患者的护理服务。将该方法与没有将临床风险信息纳入学习的常规SDAE和5种最新的分类算法进行了比较：支持向量机(SVM)、Logistic回归(LR)、随机森林(RF)、多层感知器神经网络(MLP)和朴素贝叶斯(NB)。实验结果表明，与基准方法相比，该方法获得了与基准方法相当的性能。

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. 本文的其余部分组织如下。第二部分介绍了本研究的背景。第三节对所提出的方法进行了详细的描述。在真实的临床数据集上设计了实验，并将我们的方法的性能与第四节中的最新方法进行了比较。最后，第五节给出了结论和未来的发展方向。

# II. BACKGROUND

In this section, we briefly introduce the background that is closelyrelatedtotheresearchproposedinthispaper.Oneisclinical risk prediction from EHRs. The other is the methodology research on deep learning and its applications in biomedicine. 在这一部分中，我们简要介绍了与本文提出的研究密切相关的背景。一个是来自EHR的临床风险预测。二是深度学习的方法论研究及其在生物医学中的应用

## 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]. 临床风险预测模型在医疗保健领域的应用越来越广泛。在初级保健中，它们可能被用来通过识别疾病风险较高的患者(如ACS)来有针对性地进行干预。在实践中，最近的一些研究已经开发出经过验证的模型，使临床医生能够可靠地识别ACS的低、中、高风险患者[12]。

|  |
| --- |
| 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].

For example, the GRACE risk 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]. 例如，GRACE风险评分模型，作为ACS中最流行和被广泛接受的风险评分之一，被开发用于预测单个患者的临床风险评分[3]。然而，必须注意的是，这条线上的模型是使用从高度分层的队列中特别选择的一小部分患者特征来估计的。因此，它们只考虑了少数亲手挑选的风险因素，而且是支离破碎的：结论只有在控制良好的条件下才能保留[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].TheEHRtypicallyrecordsadiversesetofclinicalinformation, 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 neuralinspiredalgorithmforriskpredictionbyusingEHR[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. 近年来，随着电子健康记录(EHR)[6]、[13]、[19]、[20]的广泛采用，许多先进的数据挖掘算法被引入临床风险预测。电子病历通常记录一组不同的临床信息，包括患者的人口统计数据、症状、实验室检查结果和治疗行为等，因此，它为临床风险预测提供了一个全面的来源。许多数据挖掘算法，如决策树[11]、贝叶斯网络[12]和模糊推理系统[15]已经被提出以探索EHR数据用于临床风险预测的潜力。例如，Tay等人。利用EHR[29]，提出了一种新的神经启发风险预测算法。Karaolis等人。应用C4.5决策树算法从EHR数据中提取冠心病事件的基本危险因素[14]。在[19]中，开发了一种混合模型，用于自动识别患者EHR中的心脏病危险因素。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]. 在[34]中，采用贝叶斯分析来预测药物开发项目中观察到的人类临床不良事件。在我们之前的工作中，我们开发了一个用于不稳定性心绞痛风险预测的遗传模糊系统，并在从中国一家医院收集的临床数据集上进行了评估[15]。此外，我们以前提出了一种基于概率主题建模的风险分层方法，通过以无监督的方式探索EHR的潜力[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. 在实际临床数据集上的实验结果表明，我们的模型可以生成连贯且信息丰富的患者子特征，以及特定于子特征的风险等级，从而为进一步的任务(如患者队列分析)提供了巨大的潜力。与我们的工作密切相关的工作在[30]中提出，其中描述了从EHR进行临床风险预测的深度学习方法。具体地说，作者将EHR记录建模为纵向事件矩阵，并将卷积神经网络模型应用于该事件矩阵，以建立基于充分的标记的患者样本集的风险预测模型。

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]. 尽管前述方法的适用性已得到证明，但利用不同种类的EHR数据进行临床风险预测仍然是医学信息学和医疗保健管理领域需要解决的关键挑战之一[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 multidimensional 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. 一般来说，风险预测模型的复杂性随着患者特征数量的增加和EHR数据的异质性而增加。值得注意的是，在使用EHR数据的临床风险预测问题中所面临的问题类似于具有高维数据的模式分类问题。在过去的十年中，随着深度机器学习技术的出现，模式分类已经进入了一种新的范式[24]。用于风险预测的多维异构电子病历数据的处理需求，以及深度机器学习技术在处理高维数据方面的优势，推动了学习技术的发展。然而，不断发展的深度机器学习技术在处理电子病历数据方面的优势尚未得到充分利用。基于这些观察结果，本文提出了一种新的基于特定深层网络结构的临床风险预测模型，即SDAE，该模型能够利用深度机器学习过程的优势来处理EHR数据的复杂性。据我们所知，现有的任何其他方法都没有像我们的模型那样具有同样的优点。

## 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]. 深度学习是机器学习中的一组算法，它试图通过使用由多个非线性变换组成的模型体系结构来对数据中的高级抽象进行建模[26]。叠加非线性变换层是深度学习算法的基本思想[26]，[28]。数据在深层架构中经过的层越多，构建的非线性转换就越复杂。由深度学习算法构建的数据的最终表示提供了来自数据的潜在的和基本的信息，这些信息可以用作构建分类器的特征[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. 堆叠去噪自动编码器(SDAE)是目前研究最广泛的深度学习结构之一。SDAE是一种对称神经网络，主要用于以无监督的方式从数据集中学习特征[28]。通常，SDAE中的每个去噪自动编码器都经过训练，从损坏的版本重建干净的“修复”输入。This is done by first corrupting the initial input x ∈ R*D* (*D* is the number of features) into x**˜** by means of a stochastic mapping x**˜** ∼ qD(x**˜**|x).这是通过破坏初始输入x ∈ R*D*（D是特征的数量）到 x˜ 实现的。And then, corrupted input x**˜** is mapped to an internal representation (or code) *f*(x**˜**) through the encoder function *f*, and then maps *f*(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*(x) = *g*(*f*(x**˜**)), and a reconstruction loss function penalizes the error made. 然后，被破坏的输入˜x通过编码器函数f被映射到内部表示(或代码)f(˜x)，然后通过解码函数g以反向方式将f(˜x)映射到输入空间，f和g的组合被称为重构函数r，即r(X)=g(f(˜x))，重构损失函数l惩罚所产生的错误。

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. SDAE对于在贪婪的分层无监督模型[28]中学习特征的层次结构[21]是有用的。学习过程通过利用原始输入数据优化损失函数来开始训练第一个自动编码器，以学习第一个最早的表示层。之后，将学习到的隐含层作为输入数据，用于训练下一个自动编码器以生成更高层的表示，并且该过程重复K次，其中K是隐含层的数目。

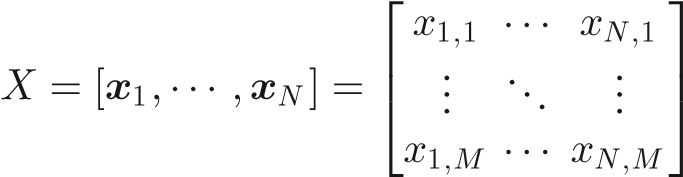
Since deep learning architectures, such as SDAE, deal with data abstraction and representation, it is quite likely to be suitable for analyzing raw data presented in different formats and/or from different sources. In this sense, deep learning has become a huge tide of technology in the field of big data and artificial intelligence. In particular, the deep structure makes significant breakthroughs onimageunderstanding, speechrecognition, natural language processing and many other areas [26], [35], [36]. 由于深度学习体系结构(如SDAE)处理数据抽象和表示，因此它很可能适合于分析以不同格式和/或来自不同来源的原始数据。从这个意义上说，深度学习已经成为大数据和人工智能领域的一股巨大的技术浪潮。特别是，深层结构在图像理解、语音识别、自然语言处理等诸多领域都取得了重大突破[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 multiplatform 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 welltrained 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]. 在医学信息学领域，深度学习逐渐受到人们的关注，并被应用于多个领域。例如，梁等人。提出了一种多模态深度信念网络(DBN)，用于从多平台观察数据中对癌症患者进行聚类[27]。Rahhal等人。描述了一种基于深度学习的心电图信号主动分类方法[28]。Li等人。应用各种训练有素的DBN来区分骨病患者和非骨病患者，目的是选择疾病的信息危险因素[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. 在[32]中，Chen et al.。提出了一种从EHR中进行表型分析的深度学习方法，并将他们的模型应用于慢性病预测建模的具体应用。在[36]中，Tran et al.。提出了一种计算框架，通过一种特定类型的深度神经网络，即受限Boltzmann机器(RBM)，在最少的人类监督下利用EMR。虽然各种深度学习模型在学习好的表示方面显示了良好的结果，但是表示是在无监督的方式下学习的。因此，这些深度学习模型重建的特征表示并不能保证最终对本文所研究的临床风险预测等有监督的任务有用。相反，我们认为，通过在深度学习模型中加入类信息，它可以重构更具区分性的特征表示，这些特征表示对于有监督的任务更有用。因此，在本文中，我们开发了一种用于ACS临床风险预测的正则化SDAE。

# III. METHODS

This section details the proposed clinical risk prediction modelusingdeeplearningforanalyzingalargevolumeofmultidimensional 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.

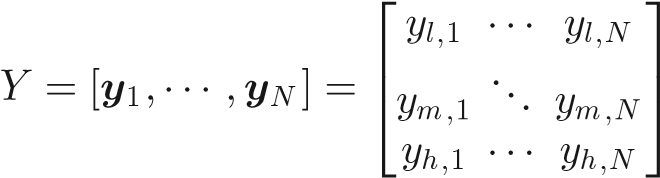
## A. Problem Formulation

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: 该数据集由大量的EHR组成。对应于特定ACS患者样本的每段EHR可以表示为患者特征向量。数据集可以表示为矩阵：

 (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 x, is viewed as a feature vector in R*M*, where the jth coordinate corresponds to the jth patient feature. 其中N是患者样本的数量，M是数据集中的患者特征的数量。对应于特定ACS患者样本x的每列X被视为RM中的特征向量，其中第j个坐标对应于第j个患者特征。

Let *Y* be a set of labels correspond to *N* training patient samples and is denoted as设Y是对应于N个训练患者样本的一组标签，表示为

 (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 R3, where the jth coordinate corresponds to the jth patient risk-level. We set *yj,i* = 1 if x*i* ∈ *j*th patient risk-level, otherwise, *yj,i* = 0. 其中下标l、m和h分别表示ACS患者的低、中和高风险水平。Y的每一列都是R3中的一个向量，其中第j个坐标对应于第j个患者的风险水平。如果Xi∈第j位患者的危险水平，我们设置Yj，i=1，否则，Yj，i=0。

We aim to seek the mapping function *f* : *X* → *Y* using a set of collected patient samples in order to classify each patient sample x to the corresponding patient category. 我们的目标是使用一组收集的患者样本来寻找映射函数f：x→Y，以便将每个患者样本x分类到相应的患者类别。

## 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 x**˜** of the original input data x to obtain the first hidden representation layer h*e*1. 在这一小节中，我们将讨论如何使用我们提出的方法来训练和验证我们的临床风险预测模型。通常，SDAE模型作为对称神经网络，主要用于以无监督的方式学习数据集的特征(参见图2(A))[37]。为了构建具有K个隐含层的深度学习结构，SDAE采用贪婪的分层无监督模式进行训练。具体地说，学习过程开始于以无监督的方式训练第一去噪自动编码器，并且利用原始输入数据x的噪声破坏版本˜x来获得第一隐藏表示层h*e*1。Then, the reconstruction layer of this denoising auto-encoder is moved to the last second layer h*d*1 of the network architecture, and the obtained hidden layer  is used as the input data for training the next auto-encoder to generate higher-level representations h, and so on. 然后，将该去噪自动编码器的重构层移动到网络结构的最后第二层hd1，并且将所获得的隐藏层he1用作用于训练下一个自动编码器的输入数据，以生成更高级的表示he2、……、hek等。After that, the reconstruction process is performed from the higher-level representation  to h and xin a backward manner. Once the greedy layer wise pretraining is complete, one can add on the resulting reconstructed feature layer xa 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)). 之后，以向后的方式执行从较高级表示Hek到HDk、···、HD1和x‘的重建处理。一旦贪婪的层级预训练完成，就可以在所得到的重建特征层x‘上添加Softmax回归层以执行临床风险预测[28]。这产生了一个深度神经网络(DNN)，专为临床风险预测问题量身定做(见图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 x. Given the noise input x**˜** of a patient vector x, a hidden feature vector h ∈ R|h| (|h| indicates the size of h) can be encoded from x**˜** through the nonlinear activation function *f*: 注意，输出节点的数量被设置为与输入节点的数量相同，并且等于患者样本向量x的长度。给定患者向量x的噪声输入˜x，隐藏特征向量h ∈ R|h| (|h|表示h的大小)可以通过非线性激活函数f从˜x编码：

h = *f* (W *e*x**˜** + b*e*)*,* (3)

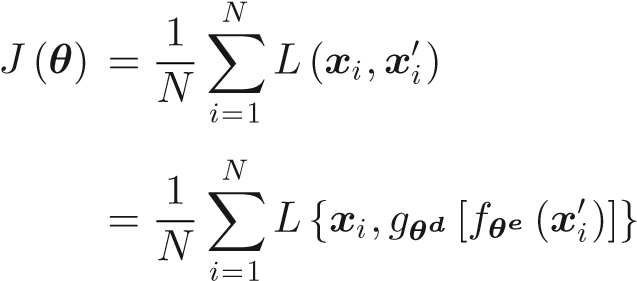
where θ*e* = {W *e,*b*e*} are parameters of the encoder in one DAE, W *e* ∈ R|h|×*M* is the encoder weight matrix, b*e* ∈ R|h| is the encoder bias vector, and *f*(*x*) is the sigmoid function [28].

In the decoding phase, the hidden feature vector h is also mapped through a nonlinear activation function to reconstruct the input vector xas follows: 在解码阶段，还通过非线性激活函数映射隐藏特征向量h，以如下重构输入向量x‘：

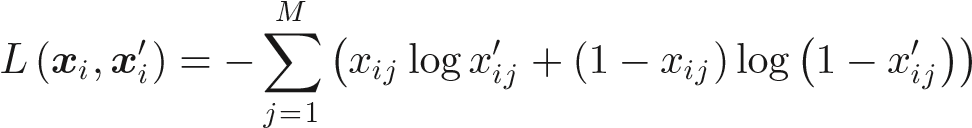
x *,* (4)

where *g*(·) is the decoder function, θ*d* = {W *d,*b*d*} are parameters of the decoder in the DAE, W *d* ∈ R*M* ×|h| is the decoder weight matrix and b*d* ∈ 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., W *d* = W *eT* . 其中g(·)是解码器函数，θd={wd，bd}是DAE中解码器的参数，Wd∈Rm×|h|是解码器权重矩阵，BD∈Rmis是解码器偏置。如文献所示，为了减少参数的数量，为编码层学习的权重简单地绑定到解码层

For each patient sample x*i*, it is corrupted to x**˜***i*, mapped to the hidden vector h*i*, and finally reconstructed as . The parameters are optimized by minimizing the reconstruction error: 对于每个患者样本Xi，它被破坏为˜Xi，映射到隐藏向量hi，最后重建为x‘i。通过最小化重建误差来优化参数：

*,* (5)

where  is the loss function, representing the reconstruction error over all N patient samples. The loss function is measured by cross-entropy as follows: 其中L(xi，xi‘i)是损失函数，表示所有N个患者样本的重建误差。损失函数由交叉熵测量，如下所示：



(6)

To optimize the cost function in (6), we first initialize the parameter vector θ 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. 为了优化(6)中的代价函数，我们首先将参数向量θ初始化为接近零的小值，然后使用称为L-BFGS的二阶优化方法，这是一种基于BFGS更新过程的拟牛顿方法。

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 x**˜** of an ACS patient sample x to obtain the first hidden representation layer h1 = *fθ*0 (x**˜**), which can be further used as the input data for training the next DAE to generate higher-level representations, and so forth. 为了构建具有K个隐含层的深度学习结构，该模型从SDAE的最低层开始，采用梯度下降法逐层训练。具体地，学习过程开始于以无监督的方式训练第一DAE，通过利用ACS患者样本x的噪声输入˜x来优化(6)以获得第一隐藏表示层h1 = *fθ*0 (x**˜**)，该第一隐藏表示层可进一步用作训练下一DAE以生成更高级别表示的输入数据，依此类推。

## C. Preserving Risk Information Into SDAE

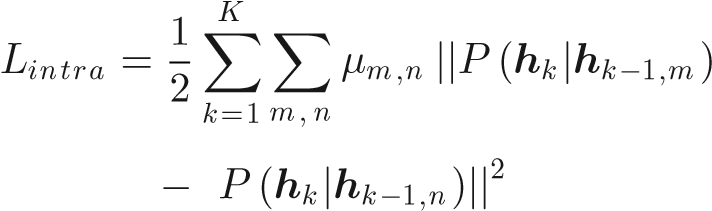
Although SDAE has been used as an extractor to deal with classification problems successfully, the features reconstructed by SDAE in an unsupervised manner. On the other side, risk informationcontainedinthetrainingdatasetcanbeincorporated into learning to make the reconstructed feature representations 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. 虽然SDAE已经成功地用于处理分类问题，但是SDAE重建的特征是无监督的。另一方面，包含在训练数据集中的风险信息可以被合并到学习中，以使重构的特征表示对于进一步的临床风险预测任务有用。为此，我们在SDAE上增加了两个特定的约束，即风险级内亲和力和风险级间排斥力，以适当地保存训练患者样本的临床风险信息。注意，这两个约束可以强制相同风险级别的患者的重构特征表示尽可能接近，而不同风险级别的患者的重构特征表示尽可能地分开。

Suppose two patient samples x*m* and 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.使用堆叠去噪自动编码器网络来学习用于临床风险预测和从异质EHR数据中的区别性特征表示。(A)使用SDAE进行预培训。(B)监督微调。 |

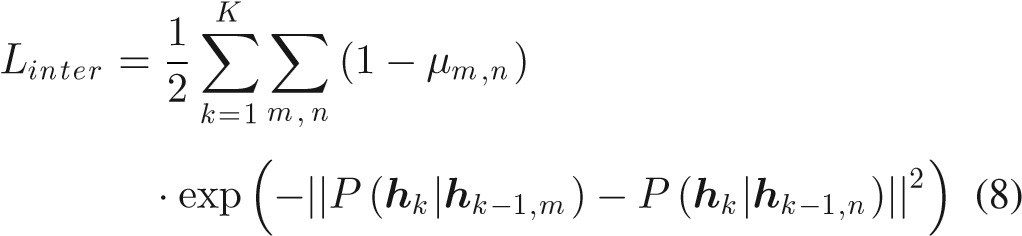
features *P*(h1*,m* |x*m* ) and *P*(h1*,n*|x*n*) are close to each other, and so forth *P*(h2*,m* |h1*,m* ) and *P*(h2*,n*|h1*,n*)*,* ···, and

*P*(h*K,m* |h*k*−1*,m* ) and *P*(h*K,n*|h*k*−1*,n*). Thus, the intra-risklevel constraint for the whole SDAE is proposed to minimize the cost function (7): 因此，提出了整个SDAE的内部风险级别约束，以最小化成本函数(7)：

 (7)

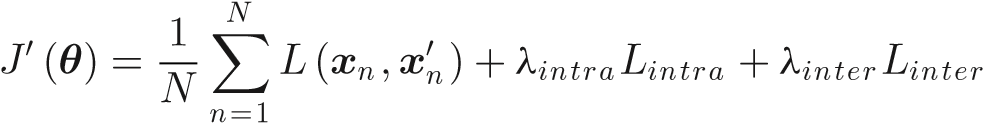
where *μm,n* = 1 if patient samples x*m* and x*n* are in the same risk level, otherwise *μm,n* = 0. h*l*−1*,m* and h*l*−1*,n* correspond to x*m* and x*n*, respectively. 其中，如果患者样本Xm和Xn处于相同的风险水平，则μm,n=1，否则μm,n=0。HK−1,m和HK−1,n分别对应于xm和xn

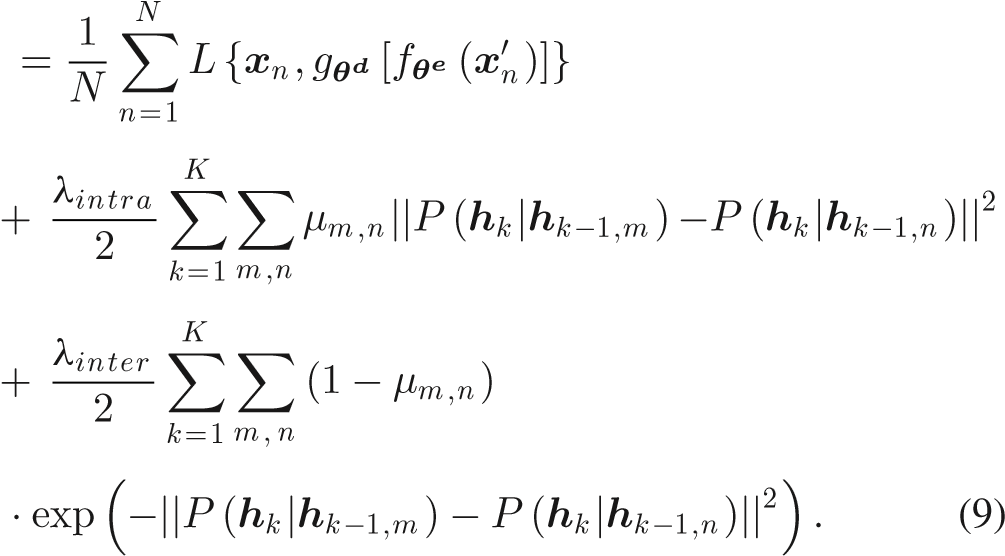
If two patient samples x*m* and x*n* are in the different risk levels, we assume that their corresponding encoding features *P*(h|x*m* ) and *P*(h|x*n*) are separated to each other. Thus, the inter-risk-level constraint for the whole SDAE is proposed to minimize the cost function (8): 如果两个患者样本xm和xn处于不同的风险级别，我们假设它们对应的编码特征P(h|xm)和P(h|xn)彼此分离。因此，提出了整个SDAE的风险级别间约束，以最小化成本函数(8)：



where *μm,n* = 1 if patient samples x*m* and x*n* are in the same risk level, otherwise *μm,n* = 0. h*k*−1*,m* and h*k*−1*,n* correspond to x*m* and x*n*, respectively. 其中，如果患者样本Xm和Xn处于相同的风险水平，则μm，n=1，否则μm，n=0。HK−1，m和HK−1，n分别对应于xm和xn

Incorporating both *Lintra* and *Linter* constraints into learning, we formulate the following objective function (9): 将LINTRAN和LINTERCONSTIONS都纳入到学习中，我们制定了以下目标函数(9)：





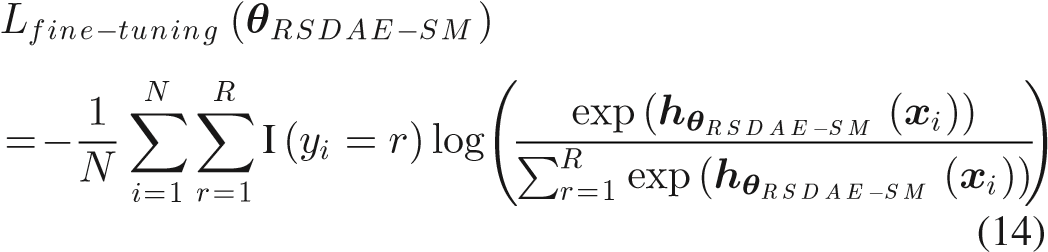
where λ*intra* and λ*inter* are two regularization parameters for the importance of intra-risk-level and inter-risk-level constraints. 其中，λintra和λinter是风险内部和风险间约束重要性的两个正则化参数。

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. 在学习过程中，两个正则项在参数上的梯度可以通过(10)-(13)计算，如下一页底部所示。

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. 提出的正则化SDAE算法在学习过程中保持对患者风险信息特征的记忆，从而使同一风险级别内的重构特征表示尽可能接近，而不同风险级别之间的重构特征表示尽可能地保持距离。我们应用这种正则化的SDAE从EHR数据中预训练一个临床风险预测模型。

## D. Supervised Fine Tuning

Oncethepretrainingiscomplete,weappendasoftmaxregression 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]. 一旦预训练完成，我们在重建的特征表示层的顶部添加一个Softmax回归层，以构建一个深层神经网络，称为正则化堆叠去噪自动编码器与Softmax回归模型(RSDAE-SM)，并使用该RSDAE-SM执行临床风险预测任务(见图2(B))。具体地说，通过对Softmax层使用(14)来最小化交叉熵损失，我们使用反向传播来微调构造的RSDAE-SM[45]。



whereI(·)isanindicationfunctionthattakes1ifthestatementis true, otherwise it takes 0. *R* indicates the number of risk levels, and hθ*RSDAE* −*SM* (x*i*) is the output of RSDAE-SM for an input x{ *i*. 其中，i(·)是一个指示函数，如果语句为真，则取1，否则取0。R表示风险级别的数量，是输入Xi的RSDAE-SM的输出。The estimation of the vector of parameters*e,*

# *e,,*

W *softmax*} of RSDAE-SM starts by initializing the weights 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.参数向量的估计是, RSDAE-SM的Wsoftmax}开始于将Softmax层的权重Wsoftmax初始化为小随机值，而K个隐藏层的权重由在预训练阶段获得的编解码权重初始化。

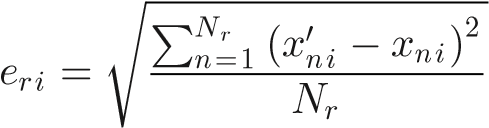
Then, the cost in (14) is minimized with a min-batch gradient descent algorithm. 然后，使用最小批次梯度下降算法最小化(14)中的成本。

## E. Risk Factor Selection

In this subsection, we propose a risk factor selection strategy to identify informative risk factors for ACS patients within differentrisklevels.Ontheonehand,weassumethatpatientfeatures with lower reconstruction errors are more re-constructible, and more re-constructible patient features are more likely to beartheunderlyingcharacteristics.Inthisregard,givenapatient 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. 在这一小节中，我们提出了一种风险因素选择策略，以确定不同风险水平的ACS患者的信息性风险因素。一方面，我们假设具有较低重建误差的患者特征更具可重构性，更具可重构性的患者特征更有可能承载潜在特征。在给定特定风险水平的患者样本的情况下，通过学习的正则化SDAE模型可以获得原始数据和重建数据之间的最小重建误差。另一方面，我们认为，所有不同风险水平的患者样本在其风险因素方面表现不同，从而产生了较大的重建误差。因此，由于输入的患者样本和用于测试的模型不匹配，我们预计在区分学习过程中会产生很大的重建误差。

In this context, Fig. 3 shows our approach to selecting potential risk factors that are more re-constructible as the discriminative features. 在这种情况下，图3显示了我们选择潜在风险因素的方法，这些因素更具可重构性，作为区分特征。

Formally, the reconstructed error of a particular feature *i* (*i* = 1*,*··· *,M*) with respect to a specific risk level *r* ∈ *R*, (*R* = {low − risk*,* medium − risk*,* high − risk}) can be calculated by (15): 形式上，特定特征i(i=1，···，M)相对于特定风险水平r∈R，(R={低−风险，中−风险，高−风险})的重建误差可以通过(15)来计算：

 (15)

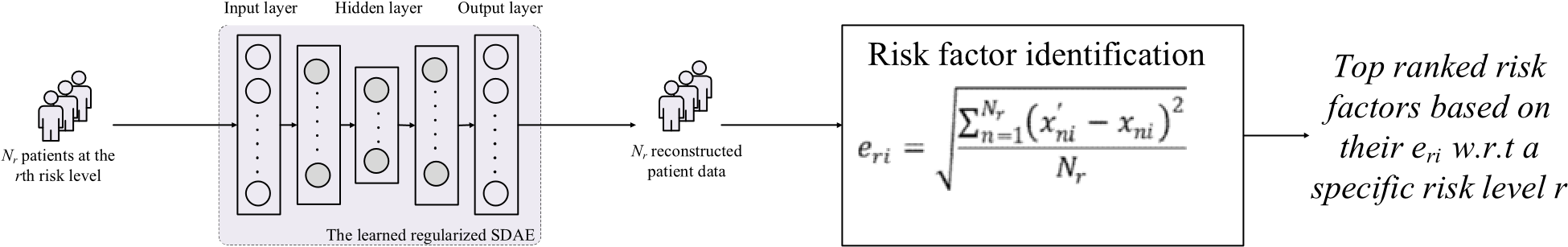


Fig. 3. Using reconstruction error to select informative risk factors for ACS patients at the rth risk level.利用重建误差为第r风险水平的ACS患者选择信息丰富的危险因素。

where *xni* is the ith feature value of the nth patient. (15) calculates the reconstruction error between the reconstructed value of patient feature, and the original value of feature *xni* for total *Nr* patient samples at the *r*th risk level. Specifically, patients with positive large *eri* are likely at the *r*th risk level, and those with negative large *eri* are unlikely otherwise. 其中xni是第n个患者的第i个特征值。(15)对于第r个风险水平的Nr个患者样本，计算患者特征x‘ni的重建值与特征xni的原始值之间的重建误差。具体地说，ERI阳性的患者可能处于第r风险水平，而ERI阴性的患者则不太可能处于第r风险水平。

## 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. 在这一部分中，我们描述了几个实验，这些实验是为了比较所提出的模型和最新方法在临床风险预测任务上的性能。

### 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 ofischemiaalongwithqualifyingelectrocardiographicchanges, positivecardiacenzymes,newdocumentationofcoronaryartery 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. 研究样本包括解放军总医院收治的3463例出院诊断为ACS的患者。ACS定义为出现缺血症状，并伴有合格的心电图改变、心肌酶阳性、有新的冠心病(CAD)病史或既往有CAD病史。ACS患者样本的平均住院时间为8.30天，而一些患者的停留时间非常短，例如只有1天，而另一些患者的住院时间超过3个月。它含蓄地表明了需要住院治疗的患者情况的多样性。根据我们临床合作者的建议，为了在ACS患者住院的早期阶段提供风险预测服务，我们从EHR系统的患者入院记录中提取了实验数据集。

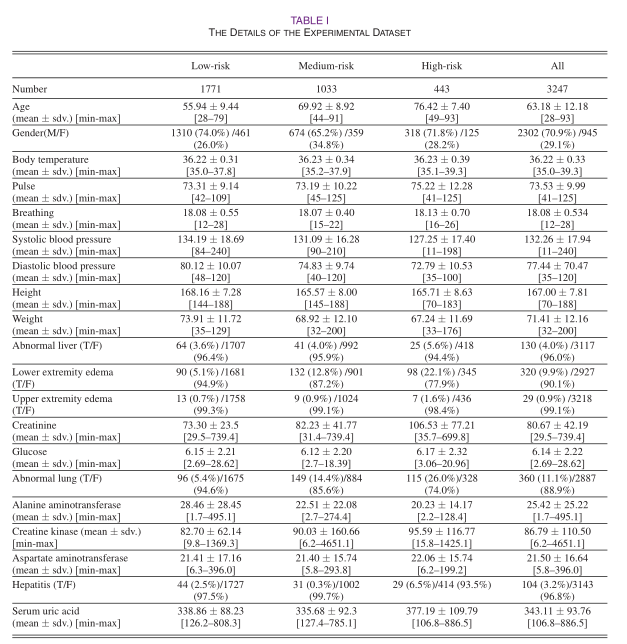
Preprocessing was performed on the experimental dataset. Specifically,forpatientfeatureswithbinaryvalues,wemaptheir binary values into 0 and 1. For example, regarding the feature “Gender”,weassumethat0indicates“male”and1indicates“female”. In addition, for patient features with multiple categories, we 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. 对实验数据集进行了预处理。具体地说，对于具有二进制值的患者特征，我们将它们的二进制值映射为0和1。例如，对于特征“性别”，我们假设0表示“男性”，1表示“女性”。此外，对于具有多个类别的患者特征，我们简单地将它们分成多个分类变量，每个变量都有二进制值来指示患者病例是否具有该特征。例如，如果ACS患者的心功能分级为“III”级，则将特征“心功能-III”设置为1，否则，将“心功能-III”设置为0。此外，对于实值患者特征，我们将其堆叠到范围[0，1]中，其中0和1对应于整个数据集中患者特征的最小值和最大值。因此，在模型学习中保留了真实的患者特征信息。

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. 我们收集了一个数据集，包括565种特征类型的348,758对独特的患者特征-值对。理想情况下，实验数据集应该有3463个∗565=1,956,595个唯一的患者特征-值对。这表明实验数据集非常稀疏，数据集中有很多缺失值。为了弥补这个问题，患者样本和缺失值超过30%的变量都没有包括在分析中。结果，实验数据集中留下了3247个患者样本，200个特征类型。除此之外，在实验中没有进行进一步的显式尝试来处理丢失的数据。请注意，SDAE的一个优点是重建特征表示。这意味着SDAE可以自动隐式地对重建后的缺失数据进行补偿。因此，我们认为我们的模型可以在一定程度上缓解数据稀疏的问题。表I显示了有关实验数据集的更详细信息，包括EHR中频繁记录的排名前20位的患者特征及其与ACS患者风险信息相关的分布。

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 groundtruth. 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. 为了确定事实真相，我们邀请了解放军总医院心内科的三名经验丰富的医生，根据他们的EHR采用多数票将患者的风险分为高、中、低三个级别。然后，我们可以检查我们的模型所建议的可能风险等级与基本事实的一致性。表一显示病人的类别分布。可以看出，高危患者(通常是感兴趣的类别)的数量远远多于临床实践中常见的低风险患者。

### 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. 准确率是常用的预测性能评价方法。然而，对于大多数倾斜的医学数据集，当将整个少数样本错误分类为多数类别时，准确率可能很高。如表一所示，处于低风险水平的患者数量远远多于处于高风险水平的患者数量。为此，在拟议的研究中使用了另一种测量方法，即接收器操作员特性曲线(AUC)下的面积。一般说来，接收者操作员特征曲线绘制了真阳性率(灵敏度)与假阳性率(特异性)的曲线。请注意，AUC是对所用测试的威力的量化汇总，它的范围从0到1。换句话说，接近1的区域表示高威力，而低于0.50的区域表示该方法不能识别ACS患者的临床风险水平。一个好的方法应该能产生较高的AUC值来预测ACS的危险性。在这项研究中，AUC作为预测准确性的典型衡量标准，被绘制成图，以捕捉正确分类的高危病例的数量如何随样本中被错误归类为高危病例的低风险患者的数量而变化。In our experiments, we adopted the source-code from WEKA1to use Mann–Whitney U test to calculate the AUC [39], [40]. 在我们的实验中，我们采用了WEKA1的源代码，使用Mann-Whitney U检验来计算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. 我们在中国人民解放军总医院心内科进行了案例研究，事先获得了医院数据保护委员会和机构审查委员会的伦理批准。必须指出的是，在这项研究和这篇论文中，患者数据都是匿名的。我们在一台与Microsoft Surface Pro 3兼容的PC上进行了所有的实验，该PC的英特尔奔腾IV CPU为2.3 GHz，内存为8G字节，运行在Microsoft Windows 10上。所提出的模型是在JA-VA中实现的。

### 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 highrisk and therefore we implement a mapping {*X, Y* } (where *Y* = {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. 我们比较了RSDAE-SM模型与常规SDAE加软最大层(SDAE-SM)模型以及支持向量机(SVM)、Logistic回归(LR)、随机森林(RF)、多层感知器神经网络(MLP)和朴素贝叶斯(NB)五种最新分类算法作为基准方法的临床风险预测性能。请注意，本研究的目标是将ACS患者直接分类为低风险、中风险和高风险，因此我们为每个模型实施了映射{X，Y}(其中Y={低、中、高})。为了与现有方法进行性能比较，我们将收集到的数据随机分为5组，并使用交叉验证对所有算法进行评估。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.与基准方法相比，(A)AUC和(B)隐含层数不同的模型具有更高的精度。 |

presented to the learned classifier for evaluation. The final results were then averaged over five rotations of the test folds. 在对4个折叠进行完整的学习过程之后，将保留的测试折叠患者数据呈现给学习的分类器进行评估。然后，将最终结果在测试折叠的五次旋转中求平均。

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. 我们首先从AUC和准确性两个方面研究了RSDAE-SM模型的隐含层数对临床风险预测性能的影响。结果表明，对于所提出的RSDAE-SM算法，AUC曲线和整体精度曲线一般随着隐层数从1层增加到4层而略有增加，当K=4时性能最好。随着K的进一步增加，曲线在一定的区间内波动。由于训练数据集的计算规模有限，当使用大量的隐含层时，我们的模型容易出现过拟合。因此，在下面的实验中，我们将隐藏层数设置为4。请注意，隐藏层的数量取决于问题，通常会根据验证数据集的性能进行调整。

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 hiddenlayersofRSDAE-SM,andtheAUCandoverallaccuracy 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 modelevenwithjust4hiddenlayershadover1.8%,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. T与基准方法相比，RSDAE-SM和SDAE-SM在隐含层数大于3的情况下总体表现较好。在RSDAE-SM的任意隐含层数下，深度学习模型的AUC和总体准确率均高于传统的NB、MLP、LR和SVM模型，在深度学习模型中，当隐含层数分别达到1层和3层时，RF的AUC和整体准确率更高，之后深度学习模型的AUC和整体准确率更高。此外，RSDAE-SM算法的性能优于SDAE-SM算法，但SDAE-SM算法随着隐含层数的增加而更接近前者。这表明，通过在SDAE中加入正则化约束，将风险信息保存到学习中，确实可以提高临床风险预测的性能。在具有4个隐含层的模型中，RSDAE-SM比常规SDAE-SM、RF、SVM、LR、MLP和NB分别有0.9%、2.2%、13.6%、12.4%、11.6%和11.3%以上的性能提升。在精度方面，与常规SDAE-SM、RF、SVM、LR、MLP和NB相比，即使只有4个隐含层，该模型的性能也分别提高了1.8%、2.1%、1.8%、14.2%、13.4%和17.2%。因此，我们的模型可以比基准方法更准确地预测患者风险。his could be

#### TABLE II

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | AUC |  |  |
| SVM | LR | RF | NB | MLP | SDAE-SM | RSDAE-SM |
| 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** |
|  |  |  |  | Pr | ecision |  |  |
| SVM | LR | RF | NB | MLP | SDAE-SM | RSDAE-SM |
| 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 |  |  |
| SVM | LR | RF | NB | MLP | SDAE-SM | RSDAE-SM |
| 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** |  |  |
| SVM | LR | RF | NB | MLP | SDAE-SM | RSDAE-SM |
| 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** |

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. 这可以归功于所提出的RSDAE-SM采用深层结构，并保留训练样本的风险信息，以便从EHR数据中学习更多和更具区分性的特征。

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. 表II显示了所提出的模型和基准方法对每个风险水平的预测性能。在七种方法中，所提出的方法的平均AUC值(0.868)是最高的。这表明我们提出的模型对不同风险水平下的患者有很好的预测能力。此外，表II不仅显示了预测模型获得的AUC值，还显示了通常用于说明分类算法性能的精度、召回率和F1值。 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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  | | | | | |
|  |  |  | | | | |
|  |  |  | | | |
|  | |  |  | | |
|  |  | |  |  | |
|  | | |  |  |
|  | |  | |  |

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-SMgenerallyachievesthebestperformance,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. 特别是，我们的模型似乎在F1的高召回率和低精确度之间进行了权衡，这与所有其他方法都不同。请注意，通常不会单独讨论精确度和召回率分数。取而代之的是，两者被组合成单个度量，例如F1，其是两者的调和平均值，即几何平均值除以算术平均值的平方。因此，由于精确度和灵敏度的混合作为一种评价指标，F1分数在特定情况下可能受到限制。在临床风险预测误差方面，在所有比较的方法中，RSDAE-SM总体上取得了最好的性能，特别是在AUC和F1方面，也可以得出严格的排序：RSDAE-SM优于SDAE-SM优于其他基准方法，如表II所示。

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 Pvalues were adjusted for multiple comparisons with the HolmBonferroni 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. 为了评估所提出的模型和基线方法之间的准确性是否存在显著差异，使用配对样本t检验[41]对每对方法之间的配对比较进行了检验，并调整了P值以进行多次与HolmBonferroni校正[42]、[48]的比较。配对样本t检验是用于确定两组观测之间的平均差是否为零的统计过程。转到我们的问题，所有患者样本的临床风险都是使用我们建议的模型和基线方法预测的，结果是在每一对方法之间进行了成对的观察。

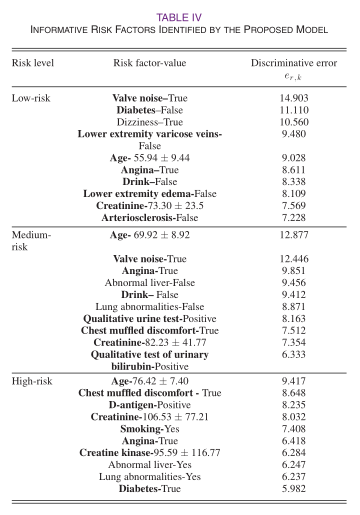
The performance of our approach showed considerable improvements in terms of accuracy in predicting clinical risks of ACS patients and the t-test showed in Table III demonstrated that there are indeed statistically significant differences between our approach and the benchmark methods. As shown in Table III, there are significant differences between the proposed deeplearningmodelsandbenchmarkmodels–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. 我们的方法在预测ACS患者临床风险的准确性方面表现出相当大的改善，表III中的t检验表明，我们的方法与基准方法之间确实存在统计上的显着差异。如表III所示，在实验数据集上，所提出的深度学习模型与基准模型-SVM(P<0.05)、LR(P<0.05)、RF(P<0.05)、MLP(P<0.01)和NB(P<0.01)-在临床风险预测方面存在显著差异。结合表II所示的性能比较结果，这表明与最先进的方法相比，所提出的深度学习方法在临床风险预测方面取得了具有竞争力和统计学意义的性能。

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

#### TABLE IV

INFORMATIVE RISK FACTORS IDENTIFIED BY THE PROPOSED MODEL



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 RSDAESM 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. 必须指出的是，各种模型的风险估计之间存在相关性。例如，建议的RSDAE-SM和LR都可以提取特定风险水平下ACS患者的潜在危险因素，虽然它们之间有统计学上的显著差异(p值<0.05)，但它们之间存在很大的重叠性。事实上，不同的风险模型为这些常见的风险因素赋予了不同的权重。实验结果表明，我们提出的RSDAESM比基准模型具有更好的风险预测性能，表明RSDAE-SM能够提取权重准确的潜在风险因素，与基准模型相比，RSDAESM更接近实际情况(即医生的经验)。

### 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. 该模型可以从电子病历数据中提取信息丰富的风险因素。在这一小节中，我们报告了如何尝试确定特定风险水平内患者组的前10个风险因素，并与临床专家确认其临床有效性，如表IV所示。注意，如上所述，在模型学习过程中，归一化患者特征被分为两类，即0或1(即患者是否具有该特征)。然后，我们恢复了它们，并在表IV中显示了这些特征的实际价值，以确保临床医生可以理解它们。

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

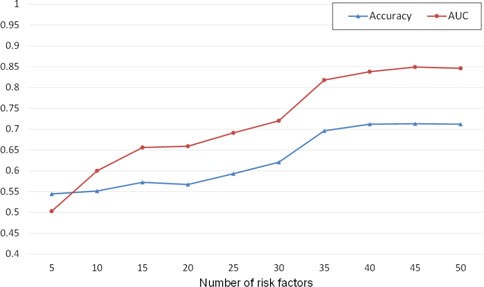


Fig. 5. The impact of the number of risk factors on the performance of ACS risk prediction.

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. 表IV中报告的最具信息量的危险因素

已由我们的临床合作者评审并认可。他们指出，这些选定的患者特征是真正提供信息的风险因素，其中一些因素，如年龄、肌酐、吸烟状况等，已经在临床队列研究中得到验证，并已被确认并用于ACS风险预测[2]、[3]、[9]。我们的模型选择的风险因素在表IV中以粗体显示，如果它们与先前在临床研究文献[2]、[3]、[9]中确定的风险因素一致。此外，临床医生表示，我们的模型可以提供ACS患者不同风险水平的选定危险因素的特定值。请注意，相同风险因素的提取值在用于指示不同的风险级别时是不同的。例如，高危患者的平均年龄为76.42岁，高于中危患者(69.92岁)。第三，我们的模型提供了这些危险因素在患者风险水平上的权重。这有助于理解在不同风险级别下风险因素的重要性。此外，最有趣的是，我们的模型可能会发现一些新的潜在风险因素。例如，对于排名第八的“高危”患者的信息性风险因素“肝脏异常-真”，该院心脏科的临床医生建议，这可能是中国人群特有的ACS(个人沟通)的潜在风险因素。请注意，中国有大量的肝炎患者，在中国人中，ACS和肝炎可能存在潜在的相关性。据我们所知，这在以前的文献中还没有出现过。他们将在临床研究中调查这一发现。

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 thenumber ofriskfactors islarger than35. Itindicates that we only need top 35 informative risk factors so as to increase the AUC and Accuracy. 为了将我们建议的模型应用于临床实践，我们应该考虑实现一个更简单的模型，并研究简单性和准确性之间的折衷。事实上，临床医生感兴趣的是，他们需要收集多少危险因素才能获得良好的预测性能。为此，我们应用我们的风险因素选择策略来提取与每个风险级别相关的最具信息量的风险因素，并将这些风险因素组合中的前50个直接提供给我们提出的ACS风险预测模型。如图所示。5、AUC曲线和准确度曲线均随信息性危险因素个数的增加而增加，当危险因素个数大于35时，AUC曲线和准确度曲线趋于稳定。这表明，我们只需要前35个信息性的危险因素，就可以提高AUC和准确性。

## 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. 本文提出了一种新的学习方法来解决异构EHR数据中ACS的临床风险预测问题。与依赖于少量精选危险因素的传统ACS风险评分方法相比，该方法可以利用大量异构的EHR数据来构建稳健的临床风险预测模型。在一个真实的临床数据集上进行了实验，结果表明，与现有的分类算法相比，该模型在临床风险预测方面取得了与之相当的性能。此外，它在识别不同风险水平的ACS患者的信息危险因素方面有很大的潜力。

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,insteadofusingexpertopinion(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. 未来可能的改进方向包括通过对隐藏表示的某些属性(如稀疏性和熵)施加其他显式惩罚或限制来丰富深度学习体系结构，以避免过度拟合，并提高深度学习网络的泛化能力[32]，[33]。此外，由于丢失数据在EHR中经常发生，我们计划在未来的工作中处理丢失数据的问题，考虑到涉及丢失数据时ACS患者预测的临床风险的不确定性。此外，我们计划进行一项前瞻性研究，观察/收集患者在住院期间和出院后的实际不良心脏事件，然后分析这些事件数据，以对患者的临床风险水平进行分层，而不是使用专家意见(医生的经验)来建立建立所建议模型的“地面事实”，而是进行前瞻性研究，以观察/收集患者在住院期间和出院后的实际不良心脏事件，然后对这些事件数据进行分析，以对患者的临床风险水平进行分层。此外，我们计划通过大规模的实验研究其他最先进的模型，并比较不同方法在临床风险预测上的性能。特别是，我们将从我们合作的医院收集大量的数据集，以测试该模型的有用性和有效性。

## ACKNOWLEDGMENT

TheauthorswouldliketothanktheEditorandtheanonymous 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/nonST 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.

1. 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.
2. 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.
3. 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.
4. P. W. Wilson *et al.*, “Prediction of coronary heart disease using risk factor categories,” *Circulation*, vol. 97, pp. 1837–1847, 1998.
5. 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.
6. 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.
7. M. A. Karaolis *et al.*, “Assessment of the risk factor of coronary heart eventsbasedondataminingwithdecisiontrees,”*IEEETrans.Inf.Technol. Biomed.*, vol. 14, no. 3, pp. 559–567, 2010.
8. 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.
9. S. Paredes *et al.*, “The CardioRisk project: Improvement of cardiovascular risk assessment,” *J. Comput. Sci.*, vol. 9, pp. 39–44, 2015.
10. J. Perk, “European guidelines on cardiovascular disease prevention in clinical practice,” *Eur. Heart J.*, vol. 33, pp. 1635–1701, 2012.
11. 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.
12. 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.
13. 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.
14. M. Ranzato *et al.*, “Unsupervised learning of invariant feature hierarchies with applications to object recognition,” in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2007.
15. 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.
16. 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.
17. 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.
18. 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.
19. G. E. Hinton and R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, pp. 504–507, 2006.
20. 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.

1. M. M. Al Rahhal *et al.*, “Deep learning approach for active classification of electrocardiogram signals,” *Inf. Sci.*, vol. 345, pp. 340–354, 2016.
2. 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.
3. Yu Chen *et al.*, “Risk prediction with electronic health records: A deep learning approach, in *Proc. 16th SIAM Int. Conf. Data Mining*, 2016.
4. 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.
5. A. Ng, “Sparse auto-encoder,” CS294A Lecture Notes, vol. 72, 2011.
6. Z. Huang *et al.*, “Predictive monitoring of clinical pathways,” *Expert Syst. Appl.*, vol. 56, pp. 227–241, 2016.
7. M.Clark,“Predictionofclinicalrisksbyanalysisofpreclinicalandclinical adverse events,” *J. Biomed. Informat.*, vol. 54, pp. 167–173, 2015.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. C. Nadeau and Y. Bengio, “Inference for the generalization error,” *Mach. Learn.*, vol. 52, pp. 239–281, 2003.
15. S. Holm, “A simple sequentially rejective multiple test procedure,” *Scand. J. Statist.*, vol. 6, pp. 65–70, 1979.
16. D. A. Cox *et al.*, “Comparative early and late outcomes after primary percutaneous coronary intervention in ST-segment elevation and nonST-segment elevation acute myocardial infarction (from the CADILLAC Trial),” *Amer. J. Cardiol.*, vol. 98, no. 3, pp. 331–337, 2006.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.