SCEHR: Supervised Contrastive Learning for Clinical Risk Prediction using Electronic Health Records

Chengxi Zang
Population Health Sciences
Weill Cornell Medicine
New York, NY, USA
chz4001@med.cornell.edu

Fei Wang
Population Health Sciences
Weill Cornell Medicine
New York, NY, USA
few2001@med.cornell.edu

***Abstract*—Contrastive learning has demonstrated promising performance in image and text domains either in a self-supervised or a supervised manner. In this work, we extend the supervised contrastive learning framework to clinical risk prediction problems based on longitudinal electronic health records (EHR). We propose a general supervised contrastive loss $\mathcal{L}_{\text{Contrastive Cross Entropy}} + \lambda \mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ for learning both binary classification (e.g. in-hospital mortality prediction) and multi-label classification (e.g. phenotyping) in a unified framework. Our supervised contrastive loss practices the key idea of contrastive learning, namely, pulling similar samples closer and pushing dissimilar ones apart from each other, simultaneously by its two components: $\mathcal{L}_{\text{Contrastive Cross Entropy}}$ tries to contrast samples with learned anchors which represent positive and negative clusters, and $\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ tries to contrast samples with each other according to their supervised labels. We propose two versions of the above supervised contrastive loss and our experiments on real-world EHR data demonstrate that our proposed loss functions show benefits in improving the performance of strong baselines and even state-of-the-art models on benchmarking tasks for clinical risk predictions. Our loss functions work well with extremely imbalanced data which are common for clinical risk prediction problems. Our loss functions can be easily used to replace (binary or multi-label) cross-entropy loss adopted in existing clinical predictive models. The Pytorch code is released at https://github.com/calvin-zcx/SCEHR.**

***Index Terms*—Supervised contrastive learning; Supervised contrastive loss; Contrastive cross entropy; Supervised contrastive regularizer; Clinical risk predictions; Electronic Health Records; Clinical time series; In-hospital mortality prediction; Phenotyping; Multi-label classification**

I. Introduction

With the accumulation and better availability of electronic health records (EHR) [1], [2], health analytics becomes one of the most important frontiers for data mining and artificial intelligence [3]. Public EHR databases [4] and benchmark suite [5] provide great resource to develop advanced data mining and machine learning algorithms for critical clinical risk prediction problems including in-hospital mortality prediction, disease phenotyping, hospital readmission, etc. [5], [6]. These problems can be formulated as a binary or multi-label classification problem using longitudinal EHR event sequence (by concatenating visits of individual patients over time) and solved by minimizing its corresponding classification loss [e.g. (multi-label or binary) cross-entropy loss] [5]–[7]. Although great endeavors have been devoted to developing complex deep learning models for these clinical risk prediction problems [5], [8]–[17], limited progress has been made over past years on these tasks regarding their performance [17]. In contrast with the majority of current research in designing more advanced predictive models, in this paper, we show that replacing widely adopted cross entropy loss by supervised contrastive loss is a promising way to improve the performance of existing models for clinical risk prediction based on longitudinal EHR data.

Recently, contrastive learning [18], which aims at learning data instance representations by bringing similar instances closer and push dissimilar instances further away from each other, has shown promising results in image classifications [19], [20], medical image understanding [21], and so on [22]. These methods mainly follow a self-supervised strategy [22], [23], which build augmented data with pseudo-labels to deal with the issue of lacking sufficient supervised information. The latest research finds that supervised information can provide additional benefits for contrastive learning in both computer vision [24] and natural language processing tasks [25]. We argue that the general idea of contrastive learning should also be helpful for clinical risk prediction tasks. However, application of contrastive learning in clinical risk prediction scenarios is challenging because: 1) the patient data (such as EHRs) for clinical risk prediction are usually more complex than images or texts in that the clinical events involved are of mixed types, high-dimensional, sparse and noisy; 2) it is challenging to augment EHR with computational methods because of the intrinsic complexity of disease mechanisms; 3) predicted clinical outcomes could also be heterogeneous. Therefore, if contrastive learning strategies can be beneficial to clinical risk prediction problems is still an open question.

In this paper, we propose **SCEHR**, a **S**upervised **C**ontrastive learning framework for clinical risk predictions using longitudinal **E**lectronic **H**ealth **R**ecord data. We illustrate the idea of SCEHR in Figure 1. The key component of SCEHR is **a general supervised contrastive loss** $\mathcal{L}_{\text{Supervised Contrastive}} =$

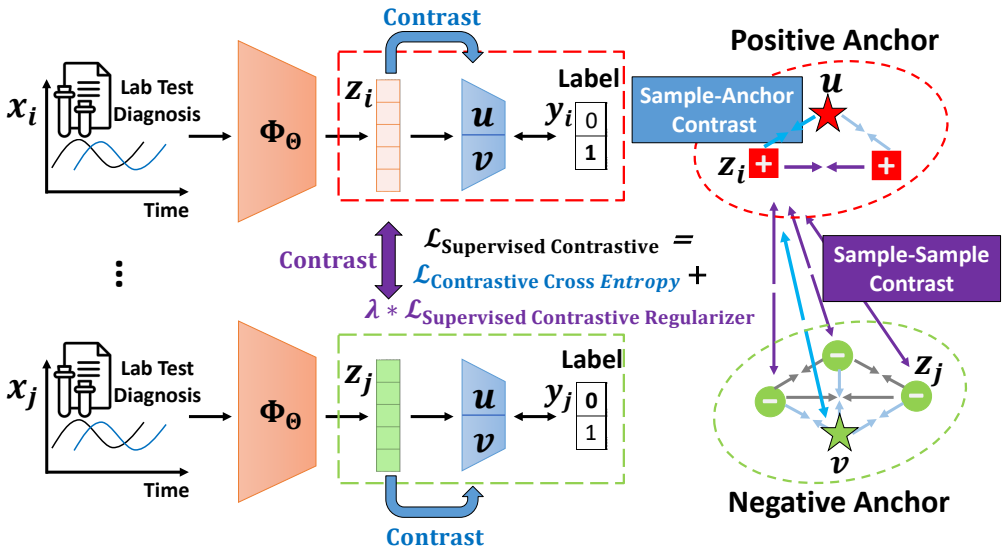


Fig. 1. An illustration of our SCEHR. We propose a general supervised contrastive learning loss $\mathcal{L}_{\text{Contrastive Cross Entropy}} + \lambda\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ for clinical risk prediction problems using longitudinal electronic health records. The overall goal is to improve the performance of binary classification (e.g. in-hospital mortality prediction) and multi-label classification (e.g. phenotyping) by pulling ($\rightarrow\leftarrow$) similar samples closer and pushing ($\leftarrow\rightarrow$) dissimilar samples apart from each other. $\mathcal{L}_{\text{Contrastive Cross Entropy}}$ tries to contrast sample representations with learned positive and negative anchors, and $\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ tries to contrast sample representations with others in a mini-batch according to their labels. For brevity, we only highlight the contrastive pulling and pushing forces associated with sample i in a mini-batch consisting of two positive samples and three negative samples.

$\mathcal{L}_{\text{Contrastive Cross Entropy}} + \lambda\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ for solving binary classification (e.g. in-hospital mortality prediction) and multi-label classification (e.g. phenotyping) in a unified framework. We propose two versions (Eq. 10 and Eq. 11) of the above supervised contrastive loss to implement the key idea of contrastive learning, i.e., pulling similar samples closer and pushing dissimilar ones apart from each other, which can be achieved by minimizing the two components of our $\mathcal{L}_{\text{Supervised Contrastive}}$. Specifically, for an arbitrary neural encoder that maps clinical time series into embedding representations, the $\mathcal{L}_{\text{Contrastive Cross Entropy}}$ learns a positive anchor and a negative anchor (for each class) respectively and tries to contrast the distance between targeted samples and the learned positive anchor versus the distance between the targeted samples and the learned negative anchor, guided by the supervised labels (e.g. positive/dead for in-hospital mortality prediction, or existence of some medical concepts for phenotyping classification). The $\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ tries to contrast every pair of samples with the same labels versus every pair of samples with different labels in a mini-batch. By leveraging supervised information, SCEHR doesn't need data augmentation and pseudo-labels. In addition, we also demonstrate the relationship between $\mathcal{L}_{\text{Supervised Contrastive}}$ and the triplet loss [26].

We validate SCEHR together with two versions of our proposed supervised contrastive losses on benchmarking clinical risk prediction tasks, including in-hospital mortality prediction and phenotyping [5], on a big real-world EHR database (MIMIC-III) [4]. We find that both versions of our proposed loss functions can improve strong baseline models and state-of-the-art models. We further investigate our modeling performance when the level of data imbalance changes. We find that our proposed loss functions work much better than binary cross entropy loss under extreme imbalance situation (say, positive ratio $\leq 1\%$), which is common in prediction problems with rare clinical outcomes. We further visualize our learned embeddings to interpret the effects of our proposed supervised contrastive losses. It is worthwhile to highlight our contributions as follows:

- **Novelty.** We propose a general supervised contrastive loss $\mathcal{L}_{\text{Contrastive Cross Entropy}} + \lambda\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ and its two instances for solving supervised binary classification and multi-label classification in a unified framework. SCEHR is one of the first applying *supervised contrastive learning* to clinical risk predictions with longitudinal EHR data.
- **Effectiveness.** SCEHR can improve both strong baseline models and the state-of-the-art models for clinical risk prediction tasks, including in-hospital mortality prediction and phenotyping. SCEHR does well with extreme data imbalance situation.
- **Flexibility.** Our proposed supervised contrastive loss functions can be easily used to replace (multi-label or binary) cross entropy loss based on existing clinical predictive models. Our PyTorch code is open-sourced at https://github.com/calvin-zcx/SCEHR.

The outline of this paper is: survey (Sec. II), problem definition (Sec. III), proposed method SCEHR (Sec. IV), experiments (Sec. V), and conclusions (Sec. VI).

II. Related Work

Deep predictive models using EHR data. Applying deep models for clinical risk prediction problems (e.g. in-hospital

mortality prediction, phenotyping, decompensation, length-of-stay prediction, readmissions, etc.) based on longitudinal electronic health record (EHR) data [1], [2], [6] show great potentials in improving health care. These tasks are usually formulated as binary or multi-label classification problems by optimizing multi-label or binary cross-entropy loss. Most of research endeavors have been devoted to developing more advanced deep models or trying to incorporate more data to capture the complexity of diseases and the EHR data, including but not limited to RNNs [5], [8], transformers [9], reverse distillation [10], variational inference [11], deep feature selection [12], attentions [13]–[16], an so on. However, despite the fast pace of modeling innovations, much slower progress has been made over past years on these tasks concerning their performance [17]. Instead of designing more complex deep predictive models, here we explore another direction: trying to innovate the default (binary or multi-label) cross entropy loss widely used in existing clinical predictive models. We focus on state-of-the-art models [5], [15], [17] which were benchmarked on public MIMIC-III data [4] considering limitations of using private EHR data.

Contrastive Learning. Contrastive learning [18], [22], aiming at learning good representations by bringing similar samples closer and push dissimilar samples away from each other through constructing contrastive loss functions, has shown promising results in image classifications [19], [20], medical image understanding [21], videos [27], etc. The idea of "contrastive" loss functions can date back to metric learning [28], triplet loss [26], Siamese neural networks [29], and the negative sampling loss of word2vec [30]. The majority of contrastive learning literature adopted self-supervised techniques [22], [23], [31], [32] by building augmented data with pseudo-labels. Recently, by explicitly using supervised labels, supervised contrastive learning has shown better performance for image classification [24] and NLP tasks [25]. To our best knowledge, only one paper [7] tried the contrastive idea for binary classification with EHR data, which adopted the negative sampling loss of word2vec [30] by negatively sampling on built heterogeneous information networks [33]. Different from all the above research, we propose a general supervised contrastive loss (together with its two versions) for solving binary classification and multi-label classification in a unified framework using longitudinal EHR data.

III. Problem Definition

In this section, we define our focused clinical risk prediction problems with longitudinal electronic health records (EHR) data. Let $x_i \in \mathbb{R}^{T_i \times D}$ represent one patient's clinical time series data, which consist of D-dimensional clinical concepts (e.g. individual measurements during his/her stay in ICU) over time T_i. Specifically, $x_{i,t,d} \in \mathbb{R}$ represents the $d^{th} \in \{1, 2, ..., D\}$ clinical concept (e.g. diastolic blood pressure) measured at timestamp $t \in \{1, 2, ..., T_i\}$ for patient i. In total, there are N patients denoted as $X = \{x_1, x_2, ..., x_N\}$ and T_i ($i \in \{1, 2, ..., N\}$) usually varies for different patients according to their length of stay, say, in ICU. Additional static features, e.g. demographic features, are denoted as $S \in \mathbb{R}^{N \times D_S}$ and $s_i \in \mathbb{R}^{1 \times D_S}$ represents patient i's features. For simplicity, we use $X = (X, S)$ to represent all the clinical time series and additional static features (if exist) for modeling. We use $Y \in \{0, 1\}^{N \times D_Y}$ to denote the targeted clinical outcomes, e.g. in-hospital mortality events, the existence of phenotype conditions, etc., which will occur beyond the observational window T_i ($i \in \{1, 2, ..., N\}$) for each patient, and $D_Y \in \mathbb{N}^+$.

Our primary goal is to learn a predictive model $\mathcal{F}_\Theta : X \to Y$, which predicts the probability of the occurrence of clinical outcomes denoted as $\hat{Y}$. The Θ are learnable modeling parameters. Regarding the value of D_Y, the above problem formulation encompasses two special cases:

- **Binary classification problem** ($D_Y = 1$), namely, $\mathcal{F}_\Theta : X \to Y$ where $Y \in \{0, 1\}^{N \times 1}$. Tasks including in-hospital mortality prediction, physiologic decompensation, etc., belong to this category.
- **Multi-label classification problem** ($D_Y > 1$), namely, $\mathcal{F}_\Theta : X \to Y$ where $Y \in \{0, 1\}^{N \times D_S}$, which can be formulated as solving multiple binary classifications simultaneously. The phenotype classification (phenotyping) task belongs to this category.

We will detail the above tasks in the experiment sections. We learn the parameters Θ of $\mathcal{F}_\Theta$ by minimizing the loss function:

$$\arg\min_{\Theta} \ \mathcal{L}(\mathcal{F}_\Theta(X), Y) \tag{1}$$

given supervised information Y, and $\hat{Y} = \mathcal{F}_\Theta(X)$ are the predicted outcomes.

In contrast with the majority of existing efforts in designing $\mathcal{F}_\Theta$, in this paper, we show that the supervised contrastive learning loss $\mathcal{L}_{\text{supervised contrastive}}$ proposed as follows is also an effective way to improve the performance of clinical predictive models.

IV. Supervised Contrastive Learning Framework for EHR

In this section, we introduce our Supervised Contrastive Learning for EHR (SCEHR) model in detail. We show the outline of our SCEHR in Figure 1 as a roadmap for this section and we summarize the overall learning process of our SCEHR in Algorithm 1.

A. General Supervised Contrastive Loss

Let Φ_Θ be any learnable neural encoder for clinical time series X, which maps X into its embedding representation Z by $Z = \Phi_\Theta(X)$. We further define a linear mapping f and a non-linear squeeze function σ (e.g. sigmoid or softmax functions) which maps the learned representations to the predicted probability by $\hat{Y} = \sigma \circ f(Z)$. We propose the following general form of *Supervised Contrastive Loss* for binary or multi-label classification problems:

$$\begin{aligned}\mathcal{L}_{\text{Supervised Contrastive}} = \mathcal{L}_{\text{Contrastive Cross Entropy}} + \\ \lambda \mathcal{L}_{\text{Supervised Contrastive Regularizer}}\end{aligned} \tag{2}$$

Our $\mathcal{L}_{\text{Supervised Contrastive}}(\hat{Y}, Z, Y)$ loss consists of two parts: a (supervised) contrastive cross entropy loss $\mathcal{L}_{\text{contrastive cross entropy}}$

which is a function of predicted labels $\hat{Y}$ against its ground truth labels Y; and a supervised contrastive regularizer $\mathcal{L}_{\text{supervised contrastive regularizer}}$ which regularizes the learned embedding representation Z by the supervised information Y. The regularizer is scaled by a non-negative hyper-parameter λ. We will detail several choices of the above losses for both binary classification and multi-label classification as follows.

B. Contrastive Cross Entropy for Binary Classification

Let $x \in X$, $z \in Z$, $y \in Y$, and $\hat{y} \in \hat{Y}$ represent clinical time series of one patient, its embedding representation, its ground-truth clinical outcomes, and its predicted outcomes respectively. We use u, v to represent the learned anchors of positive or negative clusters respectively, which are modeled as the row vectors of the weight matrix of a linear mapping f.

The **Binary Cross Entropy (BCE)** loss is widely used for clinical risk classification when there are two outcomes coded as 1 or 0, say mortality for positive cases and non-mortality for negative cases. The equation for BCE loss, denoted as $\mathcal{L}_{\text{BCE}}$, is:

$$\begin{aligned}
&\mathcal{L}_{\text{BCE}} \\
&= -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i) \\
&= -\frac{1}{N}\sum_{i=1}^{N} y_i \log \sigma(u^T z_i) + (1-y_i)\log(1-\sigma(u^T z_i)) \\
&= -\frac{1}{N}\sum_{i=1}^{N} y_i \log \sigma(u^T z_i) + (1-y_i)\log \sigma(-u^T z_i) \\
&= -\frac{1}{N}\sum_{i=1}^{N} y_i \log \sigma(\text{sim}(u, z_i)) + (1-y_i)\log \sigma(\text{sim}(-u, z_i))
\end{aligned} \tag{3}$$

where $\sigma(x) = \frac{1}{1+e^{-x}} \in (0,1)$ is the Sigmoid function and $1-\sigma(x) = \frac{1}{1+e^{x}} = \sigma(-x)$. If we define a distance measure $\text{sim}(u, z_i) = u^T z_i$ as the dot product of two data samples, intuitively, minimizing the BCE loss tries to make positive samples z_i ($y_i = 1$) close to the anchor u. Similarly, for negative samples z_i ($y_i = 0$), the BCE loss makes z_i close to $-u$.

Here we propose **Contrastive Binary Cross Entropy (CBCE)** loss, denoted as $\mathcal{L}_{\text{CBCE}}$, as follows:

$$\begin{aligned}
&\mathcal{L}_{\text{CBCE}} \\
&= -\frac{1}{N}\sum_{i=1}^{N} y_i \log \sigma(u^T z_i)\sigma(-v^T z_i) + (1-y_i)\log \sigma(v^T z_i)\sigma(-u^T z_i) \\
&= -\frac{1}{N}\sum_{i=1}^{N} \big\{ y_i \log \sigma(\text{sim}(u, z_i))\sigma(\text{sim}(-v, z_i)) \\
&\qquad + (1-y_i)\log \sigma(\text{sim}(v, z_i))\sigma(\text{sim}(-u, z_i)) \big\}
\end{aligned} \tag{4}$$

which is the first version of our $\mathcal{L}_{\text{contrastive Cross Entropy}}$ term. The above $\mathcal{L}_{\text{CBCE}}$ loss explicitly learns positive anchor u and negative anchor v separately. Minimizing the CBCE loss makes positive sample z_i (when $y_i = 1$) closer to positive anchor u than to the negative anchor v by pulling z_i closer to u and at the same time pushing z_i away from v. Similarly, for a negative sample z_i (when $y_i = 0$), minimizing the loss makes z_i closer to negative anchor v than to the positive anchor u by pulling z_i closer to v and at the same time pushing z_i away from u. Intuitively, two learned anchors u and v represent positive cluster and negative cluster respectively, and the location of each sample representation z is determined by contrasting the force $\text{sim}(u, z)$ with the force $\text{sim}(u, z)$ in a product form. We show the math of these contrastive forces in the following subsection. In all, Equation 4 contrasts each sample with positive and negative anchors in a *product form*.

Following the similar idea of $\mathcal{L}_{\text{CBCE}}$, we can also view a two-dimensional softmax cross entropy as our second instance of the contrastive cross entropy loss $\mathcal{L}_{\text{contrastive Cross Entropy}}$. We denote **Contrastive Softmax Cross Entropy (CSCE)** as $\mathcal{L}_{\text{CSCE}}$, which is defined by the following equation:

$$\begin{aligned}
&\mathcal{L}_{\text{CSCE}} \\
&= -\frac{1}{N}\sum_{i=1}^{N} \big\{ y_i \log \frac{\exp(u^T z_i)}{\exp(u^T z_i) + \exp(v^T z_i)} \\
&\qquad + (1-y_i)\log \frac{\exp(v^T z_i)}{\exp(u^T z_i) + \exp(v^T z_i)} \big\} \\
&= -\frac{1}{N}\sum_{i=1}^{N} \big\{ y_i \log \frac{\exp(\text{sim}(u, z_i))}{\exp(\text{sim}(u, z_i)) + \exp(\text{sim}(v, z_i))} \\
&\qquad + (1-y_i)\log \frac{\exp(\text{sim}(v, z_i))}{\exp(\text{sim}(u, z_i)) + \exp(\text{sim}(v, z_i))} \big\}
\end{aligned} \tag{5}$$

Equation 5 contrasts each sample with positive and negative anchors in a *ratio form*, which is a two-dimensional softmax function followed by a negative likelihood loss. Taking one positive sample z_i (when $y_i = 1$) as an example, minimizing the above loss tries to pull z_i closer to the positive anchor u than to the negative anchor v by pulling z_i to u and at the same time push z_i away from v.

C. Supervised Contrastive Regularizer

Compared with the $\mathcal{L}_{\text{Contrastive Cross Entropy}}$ which compares each sample's distance to the learned positive anchor with its distance to the learned negative anchor, the $\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ tries to explore pair-wise relationships between data samples in a mini-batch. Specifically, the $\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ tries to pull the data pairs with the same labels closer and push data pairs with different labels away from each other. Based on the supervised contrastive loss proposed in [24], we propose a simplified **Supervised Contrastive** loss as the **Regularizer (SCR)**, which is defined by the following equation:

$$\begin{aligned}
&\mathcal{L}_{\text{SCR}}(Z, Y) = \\
&\frac{-1}{N}\sum_{i=1}^{N} \frac{1}{N_{z_i} - 1} \sum_{j=1}^{N} \mathbf{1}_{i \neq j} \mathbf{1}_{y_i = y_j} \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbf{1}_{i \neq k} \exp(\text{sim}(z_i, z_k)/\tau)}
\end{aligned} \tag{6}$$

where N is the number of samples in a mini-batch, N_{z_i} is the number of samples sharing the same label as data z_i,

$\text{sim}(z_i, z_j) = \frac{z_i^T z_j}{||z_i|| ||z_j||}$, and τ is the positive temperature hyper-parameter. Here we do not adopt self-supervised data augmentation strategy [19], [24] and we only use existing supervised information Y. As a result, for each data sample z_i, we consider its distance to other $N-1$ samples and contrast these pair-wise distances according to if two samples share the same label as ratio form as detailed in the Equation 6.

D. Relationship with Triplet Loss

All the above contrastive losses $\mathcal{L}_{\text{CBCE}}$, $\mathcal{L}_{\text{CSCE}}$ and $\mathcal{L}_{\text{SCR}}$ can be approximated by a triplet loss. As for the $\mathcal{L}_{\text{CBCE}}$, the (product form) contrastive term $\log[\sigma(u^T z)\sigma(-v^T z)]$ between sample representation z and two anchors u, v can be approximated as:

$$\begin{aligned}
&\arg\min_{\Theta} -\log\{\sigma(u^T z)\sigma(-v^T z)\} \\
&= \arg\min_{\Theta} -\log\frac{1}{1+\exp(-u^T z)} - \log\frac{1}{1+\exp(v^T z)} \\
&= \arg\min_{\Theta} \log(1+\exp(-u^T z)) + \log(1+\exp(v^T z)) \\
&\approx \arg\min_{\Theta} \exp(-u^T z) + \exp(v^T z) \\
&\approx \arg\min_{\Theta} \{v^T z - u^T z + 2, 0\} \\
&= \arg\min_{\Theta} \{(\alpha v^T z - \alpha u^T z + 2\alpha), 0\}
\end{aligned} \tag{7}$$

where α is a positive scalar, Θ represents learnable parameters of u, v, and $z = \Phi(x)$. The above two approximations are achieved by $u^T z \to +\infty$ and $v^T z \to -\infty$.

As for the $\mathcal{L}_{\text{CSCE}}$, the (ratio form) contrastive term $\log\frac{\exp(u^T z)}{\exp(u^T z)+\exp(v^T z)}$ can be approximated as:

$$\begin{aligned}
&\arg\min_{\Theta} -\log\frac{\exp(u^T z)}{\exp(u^T z)+\exp(v^T z)} \\
&= \arg\min_{\Theta} \log(1+\exp((v-u)^T z)) \\
&\approx \arg\min_{\Theta} \exp((v-u)^T z) \\
&\approx \arg\min_{\Theta} \{v^T z - u^T z + 1, 0\} \\
&= \arg\min_{\Theta} \{(\alpha v^T z - \alpha u^T z + \alpha), 0\}
\end{aligned} \tag{8}$$

where the approximations are achieved by $(v-u)^T z \to -\infty$ and α is a positive scalar.

Though different forms, both contrastive cross entropy losses $\mathcal{L}_{\text{CBCE}}$ and $\mathcal{L}_{\text{CSCE}}$ try to make the distance between z and the targeted anchor u smaller than the distance between z and negative anchor v. Similar argument applies to the $\mathcal{L}_{\text{SCR}}$ as the ratio form contrastive term $\mathcal{L}_{\text{CSCE}}$. This is the major reason why all the above losses are named as **contrastive**.

E. Generalization to Multi-label Classification

We further generalize the above binary classification losses to multi-label classification losses. A typical clinical prediction application is phenotyping which tries to predict the existences of multiple clinical conditions. We model multi-label classification as solving multiple binary classifications simultaneously. Here we define our general multi-label form of $\mathcal{L}_{\text{Supervised Contrastive}}$ as follows:

$$\begin{aligned}
&\mathcal{L}^c_{\text{Supervised Contrastive}} = \\
&\frac{1}{C}\sum_{c=1}^{C} \mathcal{L}^c_{\text{Contrastive Cross Entropy}} + \lambda \mathcal{L}^c_{\text{Supervised Contrastive Regularizer}}
\end{aligned} \tag{9}$$

where C is the number of classes. Equation 2 is a special case of Equation 9 when $C = 1$.

Based on the aforementioned contrastive cross entropy losses $\mathcal{L}_{CBCE}$, $\mathcal{L}_{CSCE}$ (sec. IV-B), and the supervised contrastive regularizer $\mathcal{L}_{SCR}$ (sec. IV-C), here we propose following two versions of our general supervised contrastive loss:

- Our general multi-label form $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$ is:

$$\begin{aligned}
&\frac{1}{C}\sum_{c=1}^{C} \mathcal{L}^c_{CBCE} + \lambda\mathcal{L}^c_{SCR} \\
&= \frac{-1}{CN}\sum_{c=1}^{C}\sum_{i=1}^{N}\Big\{ y_{i,c}\log\sigma(u_c^T z_i)\sigma(-v_c^T z_i) + (1-y_{i,c})\log\sigma(v_c^T z_i)\sigma(-u \\
&+ \frac{\lambda}{N_{y_{i,c}}-1}\sum_{j=1}^{N}\mathbf{1}_{i\neq j}\mathbf{1}_{y_{i,c}=y_{j,c}}\log\frac{\exp(\text{sim}(z_i,z_j)/\tau)}{\sum_{k=1}^{N}\mathbf{1}_{i\neq k}\exp(\text{sim}(z_i,z_k)/\tau)}\Big\}
\end{aligned} \tag{10}$$

- Our general multi-label form $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$ is:

$$\begin{aligned}
&\frac{1}{C}\sum_{c=1}^{C} \mathcal{L}^c_{CSCE} + \lambda\mathcal{L}^c_{SCR} \\
&= \frac{-1}{CN}\sum_{c=1}^{C}\sum_{i=1}^{N}\Big\{ y_{i,c}\log\frac{\exp(u_c^T z_i)}{\exp(u_c^T z_i)+\exp(v_c^T z_i)} \\
&\qquad + (1-y_{i,c})\log\frac{\exp(v_c^T z_i)}{\exp(u_c^T z_i)+\exp(v_c^T z_i)} \\
&+ \frac{\lambda}{N_{y_{i,c}}-1}\sum_{j=1}^{N}\mathbf{1}_{i\neq j}\mathbf{1}_{y_{i,c}=y_{j,c}}\log\frac{\exp(\text{sim}(z_i,z_j)/\tau)}{\sum_{k=1}^{N}\mathbf{1}_{i\neq k}\exp(\text{sim}(z_i,z_k)/\tau)}\Big\}
\end{aligned} \tag{11}$$

It is worthwhile to mention that the above two multi-label classification losses encompass binary-classification losses as special cases when $C = 1$. For simplicity, we use general form $\mathcal{L}_{\text{Supervised Contrastive}} = \mathcal{L}_{\text{Contrastive Cross Entropy}} + \lambda\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ to denote both binary and multi-label cases.

F. Summary

We summarize the overall learning framework of our SCEHR in Algorithm 1. We illustrate the main idea of our SCEHR in Figure 1. The major outputs of algorithms are the targeted neural encoder Φ_Θ for X, the learned positive anchors $U = \{u_c\}_{i=1}^C$ for each of C classes, the learned negatives anchors $V = \{v_c\}_{i=1}^C$ for each of C classes $V = \{v_c\}_{i=1}^C$. The predicted probability of data i belonging to the positive cases of class c (e.g. the predicted risk of in-hospital mortality for mortality prediction task and $c = 1$ represents positive/mortality) are $\sigma(u_c^T z_i)/(\sigma(u_c^T z_i)+\sigma(v_c^T z_i))$ and $\exp(u_c^T z_i)/(\exp(u_c^T z_i) + \exp(v_c^T z_i))$ for Eq. 10 and Eq. 11 respectively. In general, our SCEHR can be used for existing clinical risk prediction models which are used

Algorithm 1: The learning framework of our SCEHR

Input: Data $X = \{x_i\}_{i=1}^{N}$, labels $Y = \{y_i\}_{i=1}^{N}$
Output: Φ_Θ:targeted neural encoder for X, $U = \{u_c\}_{i=1}^{C}, V = \{v_c\}_{i=1}^{C}$: positive and negative anchors for each of C classes
for *each epoch* **do**
  Step 1: Sampling mini-batch $X = \{x_i\}_{i=1}^{n}$
  Step 2: Generating data representations $Z = \{z_i\}_{i=1}^{n} = \Phi(\{x_i\}_{i=1}^{n})$
  Step 3: Computing the supervised contrastive loss $\mathcal{L}_{\text{Supervised Contrastive}} = \mathcal{L}_{\text{Contrastive Cross Entropy}} + \lambda \mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ by Eq. 10 or Eq. 11
  Step 4: Updating Θ, U, V by minimizing above loss.
end
  Return: Φ_Θ, U, V

for binary or multi-label classifications by replacing cross entropy losses with our Eq. 10 and Eq. 11. The PyTorch implementations of our SCEHR are open-sourced at https://github.com/calvin-zcx/SCEHR.

V. Experiments

We validate our SCEHR on a real-world electronic health records (EHR) database, Medical Information Mart for Intensive Care (MIMI-III) [4], which is publicly available. Following benchmarking works [5], we validate our SCEHR by answering the following questions:

- **In-hospital mortality prediction (Sec. V-A)** tries to predict in-hospital mortality states, namely a binary classification task, of ICU patients given their first 48-hour data in ICU. The early-prediction of at-risk patients is the key for patient stratification to improve healthcare results. Our question is: Can our SCEHR improve the performance of benchmarking models for in-hospital mortality prediction task?
- **Phenotyping classification (Sec. V-B)** tries to predict the existence of 25 common clinical conditions (coded by ICD-9 codes in EHR) of patients in ICU, namely a multi-label classification task, given their data in ICU with varying length of time. The phenotyping is key for diagnosis, comorbidity detection, and quality surveillance [34]. Our question is: Can our SCEHR improve the performance of typical benchmarking models for phenotyping task?
- **Data Imbalance Analysis (Sec. V-C)**. Positive cases in the EHR data always make up a smaller proportion than the negative cases. Our question is: How will our SCEHR perform under different levels of data imbalance?
- **Embedding Visualization (Sec. V-D)**. Our SCEHR is supposed to pull similar data embeddings closer and push dissimilar ones apart. Our question is: What will the learned embeddings look like by our SCEHR on the real-world EHR data?

TABLE I
Statistics of datasets. The ratio of positive cases is shown in the round brackets. The mortality data have binary labels, and the phenotyping data have 25-dimensional multi-labels.

	#Train	#Validation	#Test
Mortality	14,681 (13.53%)	3,222 (13.53%)	3,236 (11.56%)
Phenotyping	29,250 (16.54%)	6,371 (16.31%)	6,281 (16.53%)

TABLE II
Statistics of the varying length T_i of each patient in phenotyping dataset.

Phenotyping	#Train	#Validation	#Test
min	1	2	2
max	2804	1843	1993
mean	86.81	88.79	88.75
std.	123.87	125.56	127.66

Datasets. Following the benchmark tasks [5] on the MIMI-III dataset [4], 17 medical concepts (including Capillary refill rate, Diastolic blood pressure, Fraction inspired oxygen, Heart Rate, etc.) observed over time are selected as features, which are further feature-engineered into 76 dimensional medical time series data for predictive models. As for the mortality prediction, the first 48 hour time series are used, leading to $x_i \in \mathbb{R}^{48\times 76}$ medical time series for each patient. Besides, the latest works [15] also included additional 12 dimensional static features based on demographics (e.g. ethnicity, gender, age, height, weight, etc.) to improve the performance. The supervised labels are $\{0,1\}^N$ for N patients. As for the phenotyping classification, the time length T_i of $x_i \in \mathbb{R}^{T_i\times 76}$ varies depends on the length of stay in ICU. The labels for phenotyping multi-label classification are $\{0,1\}^{N\times 25}$. The splitting of the train, validation, and test datasets are summarized in Table I, and the statistics of the varying T_i for phenotyping classification are summarized in Table II.

We implemented our codes by Python 3.9.1, Pytorch-1.7.1, Cuda 10.1 and trained all the models on 1 GeForce RTX 2080 Ti GPU and 16 CPU cores in Linux server with Ubuntu 18.04.2 LTS. We open-source our codes at https://github.com/**/SCL-EHR and refer to [4] for the public MIMIC-III dataset and [5] for the data pre-processing and benchmarking codes.

A. In-hospital Mortality Prediction

Setup. The in-hospital mortality prediction, which is formulated as a binary classification problem, is always learned by optimizing binary cross entropy (BCE) loss in existing works [5], [15]. In this task, we evaluate our SCEHR 's capability of improving benchmark models for mortality prediction by replacing the BCE loss.

To be comparable with benchmark models, we adopt the most widely used: a) LSTM-based models (a 2-layerd LSTM model with $7,697$ learnable parameters) [5] ; and b) the state-of-the-art attention-based model Concare (a complex channel-wise GRU model with attention layers and using additional static demographic features, leading to $322,706$

TABLE III

IN-HOSPITAL MORTALITY PREDICTION RESULTS BY BENCHMARKING LSTM MODEL [5] UNDER DIFFERENT LOSSES. BCE: BINARY CROSS ENTROPY; CBCE: CONTRASTIVE BINARY CROSS ENTROPY; CSCE: CONTRASTIVE SOFTMAX CROSS ENTROPY; SCR: SUPERVISED CONTRASTIVE REGULARIZER. WE HIGHLIGHT THE BEST PERFORMANCE W.R.T DIFFERENT METRICS. WE ALSO REPORT THE STANDARD DEVIATION (STD.) OF BOOTSTRAPPED RESULTS BY RE-SAMPLING THE TEST SET 100 TIMES WITH REPLACEMENT IN ROUND BRACKETS FOR REFERENCE.

	AUROC	**AUPRC**	**Accuracy**	**min(Se, P+)**
$\mathcal{L}_{\text{BCE}}$	0.854(0.010)	0.483(0.031)	0.896(0.005)	0.487(0.026)
$\mathcal{L}_{\text{BCE}} + \lambda\mathcal{L}_{\text{SCR}}$	0.858(0.009)	0.489(0.028)	0.892 (0.005)	0.487 (0.023)
$\mathcal{L}_{\textbf{CBCE}} + \lambda\mathcal{L}_{\textbf{SCR}}$	**0.860(0.009)**	**0.504(0.031)**	**0.897(0.005)**	0.482 (0.025)
$\mathcal{L}_{\textbf{CSCE}} + \lambda\mathcal{L}_{\textbf{SCR}}$	**0.860(0.010)**	0.501 (0.030)	0.893 (0.005)	**0.505(0.024)**

TABLE IV

IN-HOSPITAL MORTALITY PREDICTION RESULTS BY BENCHMARKING CONCARE [15] MODEL UNDER DIFFERENT LOSSES. ADDITIONAL STATIC DEMOGRAPHIC FEATURES ARE USED IN THIS EXPERIMENT.

	AUROC	**AUPRC**	**Accuracy**	**min(Se, P+)**
$\mathcal{L}_{\text{BCE}}$	0.864(0.010)	0.500(0.027)	0.899(0.005)	0.484(0.022)
$\mathcal{L}_{\text{BCE}} + \lambda\mathcal{L}_{\text{SCR}}$	0.864(0.009)	0.494(0.027)	0.901(0.005)	**0.500(0.022)**
$\mathcal{L}_{\textbf{CBCE}} + \lambda\mathcal{L}_{\textbf{SCR}}$	**0.868(0.008)**	0.507(0.027)	**0.903(0.005)**	0.484(0.021)
$\mathcal{L}_{\textbf{CSCE}} + \lambda\mathcal{L}_{\textbf{SCR}}$	**0.868(0.009)**	**0.508(0.027)**	0.902 (0.005)	0.497(0.022)

learnable parameters in total) [15], and compare these models with a) their original binary cross entropy loss $\mathcal{L}_{\text{BCE}}$; b) binary cross entropy loss with supervised contrastive regularizer $\mathcal{L}_{\text{BCE}} + \lambda\mathcal{L}_{\text{SCR}}$; c) our contrastive binary cross entropy loss with supervised contrastive regularizer $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$; d) our contrastive softmax cross entropy loss with supervised contrastive reularizer $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$. To be consistent with baseline implementations, we control for the same learning settings, including Adam optimizer [35] with learning rate 0.001, dropout 0.3, weight decay 0, and only grid search for best AUROC performance among two varying hyper-parameters, namely, batch size $\{128, 256, 512, 1024\}$ and $\lambda \in [0, 0.01]$. The hidden dimensions of Z, namely the penultimate layer for contrastive learning regularizer are 16 for LSTM and 32 for Concare. We set the maximum epochs of training for LSTM and Concare are 100 and 150 respectively. We set the temperature $\tau = 0.1$ for all the following experiments.

We evaluate the performance of this binary classification by the widely-adopted benchmark metrics, including *AUROC* which is the area under the receiver operating characteristic curve; *AUPRC* which is the area under the precision and recall (also known as sensitivity) curve; *Accuracy* which is the ratio of correctly predicted cases to the total cases; and *min(Se, P+)* which is the upper bound of the minimum of different sensitivity and precision pairs.

Results. Table III and Table IV show that our SCEHR improves the best performance of both the benchmark LSTM model and the state-of-the-art Concare model with respect to all the four metrics for the in-hospital mortality prediction task on the MIMIC-III dataset. More specifically, we find both two contrastive losses $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$ and $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$ outperforms $\mathcal{L}_{\text{BCE}}$ w.r.t all the metrics. The $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$ achieved the best AUROC, AUPRC, Accuracy, while the $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$ achieved similar AUROC and the best min(Se, P+) for both models, regardless of the different complexity of two benchmark models. Besides, simply applying the regularizer $\lambda\mathcal{L}_{\text{SCR}}$ to $\mathcal{L}_{\text{BCE}}$ also improves the best AUROC performance of using bare $\mathcal{L}_{\text{BCE}}$ for LSTM.

We observe similar empirical running times for different losses under the same predictive model. All the above loss functions finish 100 epochs with 256 batch size within 3 minutes for the LSTM-based model and 45 minutes for the Concare model.

In conclusion, $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$ or $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$ improves the performance of strong benchmarking model LSTM and the state-of-the-art Concare model by replacing BCE loss. Both two supervised contrastive terms, namely $\mathcal{L}_{\text{Contrastive Cross Entropy}}$ and $\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ can introduce additional performance improvement.

B. Phenotyping Classification

TABLE V

PREDICTION RESULTS OF 25 PHENOTYPES BY BENCHMARKING LSTM [5] MODEL UNDER DIFFERENT LOSSES. BCE: MULTI-LABEL BINARY CROSS ENTROPY; CBCE: MULTI-LABEL CONTRASTIVE BINARY CROSS ENTROPY; CSCE: MULTI-LABEL CONTRASTIVE SOFTMAX CROSS ENTROPY; SCR: MULTI-LABEL SUPERVISED CONTRASTIVE REGULARIZER. WE HIGHLIGHT THE BEST PERFORMANCE W.R.T DIFFERENT METRICS.

	Micro AUROC	**Macro AUROC**	**Weighted AUROC**
$\mathcal{L}_{\text{BCE}}$	0.822	0.772	0.758
$\mathcal{L}_{\text{BCE}} + \lambda\mathcal{L}_{\text{SCR}}$	**0.824**	**0.775**	**0.761**
$\mathcal{L}_{\textbf{CBCE}} + \lambda\mathcal{L}_{\textbf{SCR}}$	0.823	0.774	**0.761**
$\mathcal{L}_{\textbf{CSCE}} + \lambda\mathcal{L}_{\textbf{SCR}}$	**0.824**	0.774	**0.761**

Setup. The phenotyping, which is formulated as a multi-label classification problem, is learned by optimizing the mean of multiple binary cross entropy losses (BCE) in existing benchmarking models [5]. In this task, we evaluate our SCEHR 's ability to improve the benchmarking phenotyping models by replacing the BCE loss.

We examined the LSTM-based model (a 1-layerd LSTM model with $348,441$ learnable parameters) [1] [5] under different losses, including a) *multi-label* cross entropy loss $\mathcal{L}_{\text{BCE}}$; b) *multi-label* cross entropy loss with *multi-label* supervised contrastive regularizer $\mathcal{L}_{\text{BCE}} + \lambda\mathcal{L}_{\text{SCR}}$; c) our *multi-label* contrastive binary cross entropy loss with *multi-label* supervised contrastive regularizer $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$; d) our *multi-label* contrastive softmax cross entropy loss with *multi-label* supervised contrastive reularizer $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$. We evaluate multi-label classification performance by standard metrics including *Micro-AUROC*, *Macro-AUROC*, and *weighted-AUROC* [36]. We adopt the same setting for consistency, including Adam optimizer with learning rate 0.001, dropout 0.3, weight decay 0, and we grid search for best micro-AUROC performance among two varying hyper-parameters, namely, batch size $\{128, 256, 512, 1024\}$ and $\lambda \in [0, 0.01]$. The hidden dimension of Z, namely the penultimate layer for contrastive learning regularizer is 256.

Results. Table V reports different AUROC scores, we find that our SCEHR improves benchmarking LSTM models w.r.t all the metrics. More specifically, our $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$ and applying $\mathcal{L}_{\text{SCR}}$ directly to BCE loss achieved the best performance, indicating the benefits of introducing supervised contrastive terms.

C. Data Imbalance Analysis

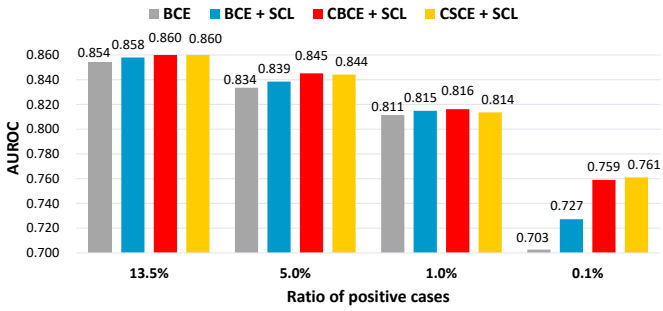
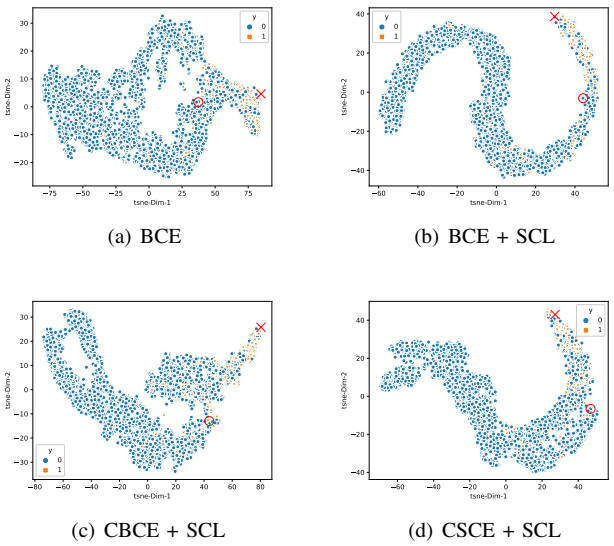


Fig. 2. In-hospital mortality prediction under different data imbalance levels.

Setup. We further investigate the performance of our loss functions when the number of positive cases in the training data is imbalanced at different levels. We studied the in-hospital mortality prediction by the benchmarking LSTM model. As shown in Table I, the original ratio of positive cases in the training dataset is 13.53%. We downsample the training data with different levels of positive cases, namely, 5%, 1%, and 0.1%, and keep the test data the same. The number (with the ratio of positive cases in the round brackets) of patients in down-sampled training datasets are $13,374$ (5%) , $12,825$ (1%), $12,708$ (0.1%), respectively. Follow the same experimental setting as section V-A, we search the best AUROC performance on the hyper-parameter space spanned by batch size $\{128, 256, 512, 1024\}$ and $\lambda \in [0, 0.01]$.

[1]We choose standard LSTM benchmarking model because different LSTM benchmarks in [5] have similar auroc performance, and the state-of-the-art Concare [15] can not be applied to time series with varying length.



(a) BCE (b) BCE + SCL

(c) CBCE + SCL (d) CSCE + SCL

Fig. 3. t-SNE plots of patient's embedding representations learned by the same LSTM-based mortality predictive model under BCE and different supervised contrastive losses on the test dataset. Orange crosses and blue dots represent the positive and negative cases respectively. The positive cases account for 11.56% of the total population. We highlight the learned positive anchor by a red cross and the negative anchor by a red dot.

Results. We report the AUROC achieved by different losses under different data imbalance levels (the ratio of positive cases) in Figure 2. We find consistent improvements of our $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$ and $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$ over the BCE loss under different imbalance levels. Besides, introducing the self-supervised regularizer to BCE also improves, but not as significant as $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$ and $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$. When the prevalence of positive cases is very rare, say 0.1%, we find that our $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$ and $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$ outperforms BCE a lot.

In conclusion, our experimental result implies that when the focused clinical outcome is rare (e.g. rare diseases) in EHR datasets, namely, a very small fraction of positive cases among the total population, replacing the BCE loss by our $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$ and $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$ can improve binary classification performance.

D. Embedding Visualization

Setup. We here try to visualize embedding representations of each patient in the test dataset learned by different losses to illustrate the effect of supervised contrastive terms. All the representations are learned by the same LSTM-based mortality predictive model as discussed in Section V-A under different losses, including a) the BCE loss $\mathcal{L}_{\text{BCE}}$; b) BCE loss with supervised contrastive regularizer $\mathcal{L}_{\text{BCE}} + \lambda\mathcal{L}_{\text{SCR}}$; c) contrastive binary cross entropy loss with supervised contrastive regularizer $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$; d) contrastive softmax cross entropy loss with supervised contrastive reularizer $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$. We control for batch size 256 for all the learning processes. We plot the 16-dimensional hidden representations Z by t-SNE [37] with

50 perplexity under 1000 iterations. The t-SNE is initialized by PCA as suggested in [38].

Results. We show embedding visualizations in Figure 3. Compared with the BCE plot (Figure 3a), we find that all the loss functions with supervised contrastive terms (Figure 3b-d) better squeeze positive samples near the red cross and negative samples near the red circle, implying their ability to pull representations with the same label closer and push representations with different labels apart. What's more, compared with $\mathcal{L}_{\text{BCE}} + \lambda\mathcal{L}_{\text{SCR}}$, our $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$ and $\mathcal{L}_{\text{CSCE}} + \lambda\mathcal{L}_{\text{SCR}}$ show more complex structures and at the same time a relatively good gap between classes, which are possible reasons accounting for their better performance. Visual inspection implies best class separation by our $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$ in Figure 3c among others, which is consistent with the best AUROC achieved by $\mathcal{L}_{\text{CBCE}} + \lambda\mathcal{L}_{\text{SCR}}$. Besides, we can also find many points that are located among data clusters with different labels, indicating the intrinsic difficulty in clinical risk predictions with longitudinal EHR data [17].

VI. Conclusion

In this paper, we propose a general supervised contrastive loss form $\mathcal{L}_{\text{Contrastive Cross Entropy}} + \lambda\mathcal{L}_{\text{Supervised Contrastive Regularizer}}$ for solving both binary classification and multi-label classification in a unified framework for clinical risk prediction using EHR data. Our proposed loss improves the performance of strong baselines and even state-of-the-art models on benchmarking clinical risk prediction using real-world longitudinal EHR data, works well with extremely imbalanced data, and can be easily used to existing clinical risk predictive models by replacing their (binary or multi-label) cross entropy loss. Our Pytorch code is released at https://github.com/calvin-zcx/SCEHR. For future work, more instances of the above supervised contrastive loss can be proposed. More clinical risk predictive models, EHR datasets, and self-supervised data augmentation techniques for longitudinal EHR data need further investigation.

Acknowledgement

This work was supported by NSF 1750326, ONR N00014-18-1-2585 and NIH RF1AG072449. The authors would also like to acknowledge the support from Google Faculty Research Award and Amazon Web Services Machine Learning for Research Award.

References

[1] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[2] J. R. A. Solares, F. E. D. Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, A. C. P. Gomes, A. H. Payberah, M. Zottoli, M. Nazarzadeh *et al.*, "Deep learning for electronic health records: A comparative review of multiple deep neural architectures," *Journal of biomedical informatics*, vol. 101, p. 103337, 2020.

[3] F. Wang and A. Preininger, "Ai in health: state of the art, challenges, and future directions," *Yearbook of medical informatics*, vol. 28, no. 1, p. 16, 2019.

[4] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[5] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[6] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1–10, 2018.

[7] T. Wanyan, H. Honarvar, S. K. Jaladanki, C. Zang, N. Naik, S. Somani, J. K. De Freitas, I. Paranjpe, A. Vaid, R. Miotto *et al.*, "Contrastive learning improves critical event prediction in covid-19 patients," *arXiv preprint arXiv:2101.04013*, 2021.

[8] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun, "Stagenet: Stage-aware neural networks for health risk prediction," in *Proceedings of The Web Conference 2020*, 2020, pp. 530–540.

[9] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi, "Behrt: transformer for electronic health records," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[10] R. S. Kodialam, R. Boiarsky, and D. Sontag, "Deep contextual clinical prediction with reverse distillation," *arXiv preprint arXiv:2007.05611*, 2020.

[11] C. Chen, J. Liang, F. Ma, L. M. Glass, J. Sun, and C. Xiao, "Unite: Uncertainty-based health risk prediction leveraging multi-sourced data," *arXiv preprint arXiv:2010.11389*, 2020.

[12] L. Ma, J. Gao, Y. Wang, C. Zhang, J. Wang, W. Ruan, W. Tang, X. Gao, and X. Ma, "Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 825–832.

[13] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," *arXiv preprint arXiv:1608.05745*, 2016.

[14] H. Song, D. Rajan, J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[15] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, and J. Gao, "Concare: Personalized clinical feature embedding via capturing the healthcare context," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 833–840.

[16] J. Luo, M. Ye, C. Xiao, and F. Ma, "Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 647–656.

[17] D. Bellamy, L. Celi, and A. L. Beam, "Evaluating progress on machine learning for longitudinal electronic healthcare data," *arXiv preprint arXiv:2010.01149*, 2020.

[18] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, 2020.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[20] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *arXiv preprint arXiv:2006.09882*, 2020.

[21] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," *arXiv preprint arXiv:2010.00747*, 2020.

[22] X. Liu, F. Zhang, Z. Hou, Z. Wang, L. Mian, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *arXiv preprint arXiv:2006.08218*, vol. 1, no. 2, 2020.

[23] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2021.

[24] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.

[25] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," *arXiv preprint arXiv:2011.01403*, 2020.

[26] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," 2010.

[27] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," *arXiv preprint arXiv:2010.09709*, 2020.

[28] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification." *Journal of machine learning research*, vol. 10, no. 2, 2009.

[29] D. Chicco, "Siamese neural networks: An overview," *Artificial Neural Networks*, pp. 73–94, 2021.

[30] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[31] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," *arXiv preprint arXiv:2010.01028*, 2020.

[32] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," 2020.

[33] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 787–795.

[34] A. Oellrich, N. Collier, T. Groza, D. Rebholz-Schuhmann, N. Shah, O. Bodenreider, M. R. Boland, I. Georgiev, H. Liu, K. Livingston *et al.*, "The digital revolution in phenotyping," *Briefings in bioinformatics*, vol. 17, no. 5, pp. 819–830, 2016.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] scikit learn.org, *Compute Area Under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores.*, 2021 (accessed January 29, 2021), https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

[37] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[38] D. Kobak and G. C. Linderman, "Initialization is critical for preserving global data structure in both t -sne and umap," *Nature Biotechnology*, p. 1–2, Feb 2021.