

Bootstrapping Your Own Positive Sample: Contrastive Learning With Electronic Health Record Data

Tingyi Wanyan

*Indiana University
Bloomington, IN, USA*

TIWANYAN@IU.EDU

Jing Zhang

Renmin University of China

ZHANG-JING@RUC.EDU.CN

Ying Ding

*University of Texas at Austin
Austin, Texas, USA*

YING.DING@ISCHOOL.UTEXAS.EDU

Ariful Azad

*Indiana University
Bloomington, IN, USA*

AZAD@IU.EDU

Zhangyang Wang

*University of Texas at Austin
Austin, Texas, USA*

ATLASWANG@UTEXAS.EDU

Benjamin S Glicksberg

*Icahn School of Medicine at Mount Sinai
New York, NY, USA*

BENJAMIN.GLICKSBERG@MSSM.EDU

Abstract

Electronic Health Record (EHR) data has been of tremendous utility in Artificial Intelligence (AI) for healthcare such as predicting future clinical events. These tasks, however, often come with many challenges when using classical machine learning models due to a myriad of factors including class imbalance and data heterogeneity (i.e., the complex intra-class variances). To address some of these research gaps, this paper leverages the exciting contrastive learning framework and proposes a novel contrastive regularized clinical classification model. The contrastive loss is found to substantially augment EHR-based prediction: it effectively characterizes the similar/dissimilar patterns (by its "push-and-pull" form), meanwhile mitigating the highly skewed class distribution by learning more balanced feature spaces (as also echoed by recent findings). In particular, when naively exporting the contrastive learning to the EHR data, one hurdle is in generating positive samples, since EHR data is not as amendable to data augmentation as image data. To this end, we have introduced two unique positive sampling strategies specifically tailored for EHR data: a *feature-based* positive sampling that exploits the feature space neighborhood structure to reinforce the feature learning; and an *attribute-based* positive sampling that incorporates pre-generated patient similarity metrics to define the sample proximity. Both sampling approaches are designed with an awareness of unique high intra-class variance in EHR data. Our overall framework yields highly competitive experimental results in predicting the mortality risk on real-world COVID-19 EHR data with a total of 5,712 patients admitted to a large, urban health system. Specifically, our method reaches a

high AUROC prediction score of 0.959, which outperforms other baselines and alternatives: cross-entropy(0.873) and focal loss(0.931).

1. Introduction

The use and adoption of electronic health record (EHR) systems in hospitals have rapidly grown in the past decade, and the massive amount of EHR data accumulated from routine care has naturally facilitated a surge of research from data-driven clinical informatics applications, such as medical concept extraction (Jiang et al., 2011), patient trajectory modeling (Ebadollahi et al., 2010), disease inference (Austin et al., 2013), and clinical decision support systems (Kuperman et al., 2007). While primarily designed for operation processes, EHR systems electronically store data associated with each patient encounter with the health system, including disease diagnoses, laboratory test results, vital signs, and more. In recent years, many machine learning techniques, including deep learning, have been leveraged to derive insights from EHR data (Shickel et al., 2017).

One of the challenges in learning from EHR data is the heterogeneous nature by which it is represented – in terms of not only data types involved, but also the various types of contributing factors to disease phenotypes as well as noise, bias, and confounding variables. More specifically, patients with the same disease or outcome can deviate in terms of phenotype representation leading to a high amount of **intra-class variance** in terms of etiology and presentation, especially for complex diseases. Another technical barrier to successful machine learning implementation of EHR data arises from the severe **data imbalance** as commonly seen in real-world biomedical data. Many important healthcare events are rare, and the class distribution in EHR data is often highly skewed in having a much higher proportion of the majority, background class (i.e., healthy) than those the outcome of interest, such as rare clinical diseases. These aspects can strongly bias the classifier to miss the rare but critical classes. Additionally, many features have high missingness such as lab test results which are sporadic and only performed in certain clinical scenarios.

Recently, contrastive learning (He et al., 2019; Chen et al., 2020a) has garnered a significant amount of interest, mainly due to its encouraging promise of learning representations while requiring no human annotation that are on par with supervised learning. Originating from the unsupervised learning realm, contrastive learning has also been extended to the semi-supervised (Chen et al., 2020b) and fully-supervised (Khosla et al., 2020) settings, allowing for effectively leveraging label information when available. The underlying idea across these settings is to pull “similar points” (or points belonging to the same class) together, while simultaneously pushing apart “dissimilar points” (or points belonging to different classes), in the embedding space. Accordingly, each training sample is an *anchor* with its similar and dissimilar points in the same training serving as *positive* and *negative* samples, respectively. In another use case (Khosla et al., 2020), the supervised variant of contrastive loss was found to consistently perform better than cross-entropy on large-scale classification problems, while also yielding superior robustness to data noise and unseen corruptions during testing. Lately, it was found by (Yang and Xu, 2020; Kang et al., 2021) that when the data distribution is skewed, contrastive learning can learn a more balanced feature space than its vanilla supervised counterpart.

The above progress sheds light on a new opportunity to integrate contrastive learning into EHR classification tasks, with the hope for stronger discriminative power, robustness, and handling outcome imbalance. However, the incorporation of contrastive learning into EHR data faces one unique roadblock: the process of sampling patients with *positive* outcomes. For any "anchor" sample, a good positive sample has to be semantically aligned (e.g., the same class), yet also nontrivial and informative enough for learning meaningful features. It is well-known that the quality of positive sampling determines the effectiveness of contrastive learning to a large extent (Grill et al., 2020; Li et al., 2020). Previous literature either (in the unsupervised setting) performs data augmentation to create a similar anchor and positive examples (He et al., 2019; Chen et al., 2020a), or (in the supervised setting) randomly samples examples from the same class (Khosla et al., 2020; Kang et al., 2021). Unfortunately, EHR data is not as readily amendable to data augmentation due to the previously mentioned aspects (e.g., varying feature completeness). Furthermore, random class-wise sampling overlooks the ultra-high intra-class variance in certain clinical phenotypes and will easily collapse the learned features, which we find in our experiments. In summary, no off-the-shelf positive sampling is directly applicable for contrastive learning in EHR data.

In this work, we address the above challenges by presenting a holistic framework for classification tasks in EHR data using contrastive learning to explicitly take outcome imbalance and intra-class heterogeneity into consideration. We introduce contrastive loss into a focal loss-based classification pipeline and show that contrastive loss boosts both overall task accuracy in different scenarios of class imbalance in EHR data. As the key innovation, we design two unique positive sampling strategies specifically tailored for EHR data which is less amenable to data augmentation: a *feature-based* positive sampling that exploits the feature space neighborhood structure to reinforce the feature learning; and an *attribute-based* positive sampling that incorporates pre-generated patient similarity to define the sample proximity. Both sampling approaches are designed to capture the high intra-class heterogeneity in EHR data. Our contributions can be summarized in the following three-folds:

- **Framework:** We are the first to integrate a bespoke approach using contrastive learning into the challenging task of outcome classification for real-world, imbalanced, and heterogeneous EHR data. When combined with a strong baseline using focal loss, we demonstrate that integrating contrastive learning can further remarkably boost predictive accuracy and be robust to data imbalance.
- **Methodology:** We present two new positive sampling approaches that enable the usage of contrastive learning in EHR data. Both approaches take better care of the high intra-class variance in EHR data, and outperform existing vanilla options. These approaches may open a new set of possibilities for extending contrastive learning to many other domains where data augmentation is less feasible.
- **Experiments:** We test our model on predicting 24-hour mortality using real-world COVID-19 EHR data from a large, diverse health system. We assess our contrastive regularizer and two positive sampling strategies. We assess the robustness of this framework in chunks of data with different sample sizes and imbalance ratios. Our method, particularly the attribute-based positive sampling contrastive regularizer,

achieves a boost in performance over vanilla focal loss, reaching a high AUROC prediction score of 0.959, largely outperforming other alternatives.

2. Related Work

Phenotype Intra-class Heterogeneity and Subphenotypes. Clinical phenotypes, and therefore EHR data, are often heterogeneous by nature. Most complex diseases, for instance, have varying manifestations, presentations, sequelae, and outcomes. As such, these diseases are manifested by a variety of clinical data types, including lab test result ranges or disease diagnostic codes. Therefore, it is often the case that patients in the same outcome class (i.e., those that develop severe COVID-19) could have large intra-class variance in terms of etiology and presentation. Patterns in this phenomenon can be considered subphenotypes of a disease. Exploring different subphenotypes is valuable to precision medicine and can enhance the performance of the predictive tasks and lead to more personalized recommendations. There is a large body of work exploring computational methods for subphenotyping, such as Parkinson’s disease (Lewis et al., 2005), scleroderma (Schulam et al., 2015), and Glioblastoma (Verhaak et al., 2010). To better capture the pattern of subtypes, methods such as multi-task learning and hierarchical models (Suresh et al., 2018; Alaa et al., 2018) have been studied. Recently, Su et al. characterized the heterogeneity of COVID-19 into four distinct clinical subphenotypes Su et al. (2021).

Contrastive Learning for Data Imbalance. It is long known that real-world datasets, particularly EHR data, have issues with outcome imbalance which limits performance for many analyses (Santiso et al., 2019; Wu et al., 2010)(see section 3.2 for more details). Traditional methods like ensemble learning (Khalilia et al., 2011) or re-balancing classes (Buda et al., 2018) have been utilized for certain tasks in this realm but come with their own set of limitations. The incorporation of focal loss achieved higher accuracy on EHR-based classification tasks (Wang et al., 2018, 2020). These facets suggest that contrastive learning may serve to better address imbalance in addition to focal loss. Studies found that decoupling the data representation and classifier can lead to better classification for long-tailed datasets (Kang et al., 2019). Yang and Xu (2020) proposed either to use a simple pseudo-labeling strategy to alleviate label bias with extra data in a semi-supervised manner, or to abandon labels at the beginning and pre-train classifiers in a self-supervised manner, can both improve the performance class-imbalanced learning. Kang et al. (2021) compared self-supervised contrastive learning and supervised methods for long-tailed datasets and found that self-supervised contrastive learning constantly outperforms supervised methods on heavily imbalanced data. Their study shows that representation learned from the self-supervised contrastive loss performs well on both balanced and imbalanced datasets because it can generate a balanced feature space with similar separability for all classes.

Supervised Contrastive Learning. Self-supervised contrastive learning takes an augmented anchor as the single positive for each anchor without taking the advantage of pre-labeled class information. So, in one batch, images from the same class of the anchor have been treated as the negative samples which can reduce the performance (Khosla et al., 2020). Supervised contrastive learning can leverage label information to generate better positive samples that embeddings of objects from the same classes should be similar, while

embeddings of objects from different classes should be dissimilar. [Khosla et al. \(2020\)](#) proposed the multiple positives per anchor in addition to many negatives and provide a unified loss function which can be viewed as the generalization of both triplet ([Weinberger and Saul, 2009](#)) and N-pair ([Sohn, 2016](#)) losses. Their loss is less sensitive to hyperparameters, which can provide consistent boosts for accuracy for different datasets, and is robust to natural corruptions. But taking multiple positive samples from the same class in EHR data can be problematic because there are complex intra-class variances in the EHR data which can lead to over-collapsing feature spaces. [Kang et al. \(2021\)](#) proposed k-positive contrastive learning to take advantage of supervised contrastive learning and also solve the issue of imbalanced data. The proposed k-positive contrastive learning takes k instances of the same class of the anchor as the positives and demonstrates superior performance over the latest contrastive learning methods on both balanced data and long-tailed data. The proposed k-positive contrastive loss is different from ([Khosla et al., 2020](#))’s supervised contrastive learning which uses all the instances from the same class to be the positive pairs and cannot avoid the dominance of large classes in the representation space. While [Kang et al. \(2021\)](#)’s k-positive contrastive loss can carefully balance the equal number of positive pairs for all classes, especially for the long-tailed data which class instances vary dramatically. Therefore, k-positive contrastive loss can generate feature spaces with desirable balance and discriminative ability.

Applications of Contrastive Learning on Clinical Data. There have been encouraging studies that have demonstrated the potential utility of contrastive learning in health data, albeit only a few. Contrastive learning has been applied for more robust learning of various patient data modalities. [Kiyasseh et al. \(2020\)](#) created CLOCS, a family of contrastive learning methods on unlabeled cardiac physiological data for downstream tasks like better quantifying patient similarity for disease detection. [Kostas et al. \(2021\)](#) created BENDR, which leverages transformers and contrastive self-supervised learning to better learn representations of electroencephalogram data. Moving to the realm of EHR data, [Li et al. \(2019\)](#) built a framework that enhanced predictive performance for common diseases across multiple sites without the need to share data by leveraging Distributed Noise Contrastive Estimation. [Wanyan et al. \(2021\)](#) demonstrated that contrastive learning enhanced prediction of critical events in COVID-19 as well as led to better patient representations. [Chen et al. \(2021\)](#) used transformers and contrastive learning to learn embedding representations of EHR data and showed that these representations allow for better predictions in disease retrieval tasks. It is clear that the potential utility of this framework into the realm of healthcare, especially EHR, is just at the beginning.

3. Methodology

3.1. Addressing EHR Data Imbalance: Contrastive-Regularized Focal Loss

Focal loss ([Lin et al., 2017](#)) has been shown to work well in imbalanced EHR data and significantly improves performance in such tasks as predicting mortality from heart failure ([Wang et al., 2020](#)). However, focal loss may underperform in situations where there is a lot of patient intra-class heterogeneity. This may be the case because of the unimodal loss structure of focal loss which classifies based solely on label information, making it unable to leverage and learn from the rich features that constitute patient data within a group ([Oord](#)

et al., 2018). To leverage the above problem, we propose a learning framework by adding a contrastive regularizer to the base focal loss, for boosting the performance on EHR tasks on outcome imbalance as well as intra-class heterogeneity. We apply our contrastive learning framework similar to Chen et al. (2020a), then compile an end-to-end training strategy that incorporates our novel k-positive selection approaches (described in more detail in the next section):

$$\mathcal{L} = \hat{\mathcal{L}} + \alpha\mathcal{L}^* \quad (1)$$

Where $\hat{\mathcal{L}}$ is the focal loss, \mathcal{L}^* is the additional contrastive loss as a regularizer, α is the regularization coefficient that controls the loss magnitude. Different from (Chen et al., 2020a), we use the supervised contrastive learning framework (Khosla et al., 2020) to generate augmented positive samples from the same class group, for each batch, we sample N samples, and use our proposed sampling strategies to generate K positive samples from the same class for each data in batch, then use all $(N - 1) \times K$ samples as the negative samples:

$$\mathcal{L}^* = -\frac{1}{N} \sum_{p=1}^N \sum_{i=1}^K \log \frac{\exp(\mathbf{z}_p \cdot \mathbf{z}_i^+ / \tau)}{\exp(\mathbf{z}_p \cdot \mathbf{z}_i^+ / \tau) + \sum_{j=1}^{(N-1)K} \exp(\mathbf{z}_p \cdot \mathbf{z}_j^- / \tau)} \quad (2)$$

Where \mathbf{z}_p is the embedding vector of one patient data in batch, \mathbf{z}_i^+ are positive sample embeddings. \mathbf{z}_j^- are negative sample embeddings. K is the positive sample numbers, τ is the temperature hyper-parameter. Intuitively, we apply a vanilla k-random positive sampling strategy (Kang et al., 2021) as our first positive sampling strategy as well as a baseline for comparison with our next proposed two rational sampling strategies, the learning algorithm with this k-random positive sampling is shown in Alg 1. We highlight the part in red to mainly distinguish the K-random sampling algorithm from our next proposed two positive sampling strategies.

Algorithm 1 K-Random Positive Sampling Contrastive Learning

Input: longitudinal EHR features

Output: Embedding vector representation for patients

- 1: Initialize all weights W
 - 2: **for** each epoch **do**
 - 3: **while** not converged **do**
 - 4: sample a mini-batch training patients $P \in P_{all}$.
 - 5: **for** each $p \in P$, randomly sample k positive data $p_k^+ \in P_{all}$ that have the same label as p .
 - 6: Compute: $\mathcal{L} = \hat{\mathcal{L}} + \alpha\mathcal{L}^*$
 - 7: compute gradient of loss function $\nabla\mathcal{L}$ and update weight matrices W .
 - 8: **end while**
 - 9: **end for**
 - 10: **return** embedded representation $c_p, \forall p \in P_{all}$
-

3.2. Addressing EHR Data Heterogeneity: Two Proximity-based Positive Sampling Approaches

A key knob in contrastive learning is to find positive pairs for anchor examples and to maximize their learned features’ similarity to inject the desired invariance. As class labels are available for us, a vanilla option is to follow the k-random positive sampling strategy (Kang et al., 2021) for supervised contrastive learning, which just randomly picks samples belonging to the same class to form anchor-positive pairs. However, we demonstrate this naive baseline is unable to work well for EHR data since it neglects the high intra-class variance as seen in heterogeneous phenotypes like COVID-19, and such learned features will easily collapse and fail to generalize. To this end, we develop two unique positive sampling strategies specifically tailored for EHR data: a *feature-based* positive sampling that exploits the feature space neighborhood structure to reinforce the feature learning; and an *attribute-based* positive sampling that incorporates the raw features to define the sample proximity. Both sampling approaches are designed to capture the high intra-class variance in EHR data. The main difference between the two lies in how they compute sample similarity, which has been highlighted in Algorithm 2 (blue) and Algorithm 3 (orange), respectively.

Feature-based Sampling In our feature-based k nearest neighborhood (knn) positive sampling method, we construct a knn graph by ranking patients by their similarities among the embedding vectors within the same class and selecting positive samples for every training patient from its top k neighbors in the feature knn graph. We define the similarity score between one pair of features as their cosine similarity:

$$\mathbf{z} \cdot \bar{\mathbf{z}} / \|\mathbf{z}\| \|\bar{\mathbf{z}}\| \quad (3)$$

Where \mathbf{z} and $\bar{\mathbf{z}}$ are two embedding feature vectors. Since constructing a knn graph would take large computational resources, especially when data is huge, we update the knn graph every epoch instead of every iteration. Our feature-based knn sampling contrastive regularizer loss is then written as follow:

$$\mathcal{L}^*_{feature} = -\frac{1}{N} \sum_{p=1}^N \sum_{i=1}^K \log \frac{\exp(\mathbf{z}_p \cdot \mathbf{z}_{i(feature)}^+ / \tau)}{\exp(\mathbf{z}_p \cdot \mathbf{z}_{i(feature)}^+ / \tau) + \sum_{j=1}^{(N-1)K} \exp(\mathbf{z}_p \cdot \mathbf{z}_j^- / \tau)} \quad (4)$$

Where $\mathbf{z}_{i(feature)}^+$ are the top k sample embeddings from the feature knn graph. The knn feature based contrastive learning algorithm is shown in Alg 2.

Attribute-based Sampling In our attribute-based positive sampling model, we construct the knn graph to rank patients by their similarities by lab test features. Since the input values are in different scales, we scale the feature values into the range of [0,1] based on their mean and standard deviation, and define the attribute similarity score by their Euclidean distance. We included 34 lab test features that were present in 80% of the cohort (see Appendix A):

$$\sum_{\bar{k}=1}^m \|X_{i\bar{k}} - X_{j\bar{k}}\|_2 \quad (5)$$

Where X_i and X_j are two input vectors representing two individual patients, m is the input feature dimension. \bar{k} represents each highlighted feature. The contrastive regularizer loss

Algorithm 2 Feature-based K-nearest Neighborhood Sampling Contrastive Learning

Input: Longitudinal EHR features**Output:** Embedding vector representation for all patients

- 1: Initialize all weights W_p
 - 2: **for** each epoch **do**
 - 3: Compute similarities between all pairs of embedding feature representations based on equation 3, and build knn graph from it.
 - 4: **while** not converged **do**
 - 5: sample a mini-batch training patients $P \in P_{all}$.
 - 6: for each $p \in P$, sample k positive data $p_k^+ \in P_{all}$ that have the same label as p , and are connected to node p in the knn graph.
 - 7: Compute: $\mathcal{L} = \hat{\mathcal{L}} + \alpha \mathcal{L}^*_{feature}$
 - 8: compute gradient of loss function $\nabla \mathcal{L}$ and update weight matrices W_p .
 - 9: **end while**
 - 10: **end for**
 - 11: **return** embedded representation $c_p, \forall p \in P_{all}$
-

is written as follow:

$$\mathcal{L}^*_{attribute} = -\frac{1}{N} \sum_{p=1}^N \sum_{i=1}^K \log \frac{\exp(\mathbf{z}_p \cdot \mathbf{z}_{i(attribute)}^+ / \tau)}{\exp(\mathbf{z}_p \cdot \mathbf{z}_{i(attribute)}^+ / \tau) + \sum_{j=1}^{(N-1)K} \exp(\mathbf{z}_p \cdot \mathbf{z}_j^- / \tau)} \quad (6)$$

Where $\mathbf{z}_{i(attribute)}^+$ are the top k sample embeddings from the attribute knn graph. Alg 3 shows our attribute-based contrastive learning algorithm. While our feature-based sampling strategy updates the knn graph every epoch, the attribute-based knn graph is computed before training begins. In other words, we use pre-computed patient similarity metrics to select positive samples which are not learned by our algorithm.

4. Experiment

4.1. Dataset Description and Experimental Design

We assess our framework in a clinically-relevant EHR task, specifically predicting mortality from real-world COVID-19 data. Our dataset is comprised of patients from a large and diverse health system in an urban environment. We obtain such data for 5,712 patients who tested positive for COVID-19 and were hospitalized (23% mortality rate). The collected EHR data contains the following information: COVID-19 status, demographics, laboratory test results, vital signs, and comorbidities (see Appendix A for details).

Our primary task was to predict mortality in COVID-19 patients 24 hours before the event. We model the EHR data using a standard longitudinal RNN model framework as in Choi et al. (2016). Longitudinal data (i.e., features with multiple values), specifically lab tests and vital signs, were binned and averaged within 6-hour windows across their hospitalization. We concatenated non-longitudinal categorical features, specifically demographics and co-morbidities, into a separate shallow neural network layer. We then concatenated the output embedding vector from this layer with the embedding vector from the RNN model

Algorithm 3 Attribute Based K-nearest Neighborhood Sampling Contrastive Learning

Input: Longitudinal EHR features**Output:** Embedding vector representation for all patients

- 1: Initialize all weights W_p
 - 2: Pre-compute similarities between all pairs of embedding features based on equation 5, and build knn graph from it.
 - 3: **for** each epoch **do**
 - 4: **while** not converged **do**
 - 5: sample a mini-batch training patients $P \in P_{all}$.
 - 6: for each $p \in P$, sample k positive data $p_k^+ \in P_{all}$ that have the same label as p , and are connected to node p in the knn graph.
 - 7: Compute: $\mathcal{L} = \hat{\mathcal{L}} + \alpha \mathcal{L}^*_{attribute}$
 - 8: compute gradient of loss function $\nabla \mathcal{L}$ and update weight matrices W_p .
 - 9: **end while**
 - 10: **end for**
 - 11: **return** embedded representation $c_p, \forall p \in P_{all}$
-

(i.e., the last time frame) to form the final patient embedding representation z_p . All analyses were performed using TensorFlow 1.15.1 and utilized the Adam optimizer (Kingma and Ba, 2014). We set the batch size to be 32, with 6 training epochs, and set our embedding dimension to be 100. For all experiment below, we specifically pick our model parameters to be: $k = 5$, $\alpha = 0.2$, and $\tau = 1$. We conduct a thorough investigation of optimal model parameters in Appendix B.

4.2. Performance of Contrastive Regularizer Positive Sampling Strategies

For testing the overall performance of our contrastive regularizer on our task, we split our data into 70% for training and 30% for testing for performing the 7-fold cross-validation. The performance metrics are shown in Table 1. We first assess the effect of focal loss compared to the commonly used cross-entropy loss. In this comparison, we find focal loss outperforms cross-entropy loss in terms of both AUROC (6% improvement) and AUPRC (3% improvement).

We then assess any performance improvements over the base focal loss for our various positive sampling approaches, specifically k-random (random), feature-based (feature), and attribute-based (attribute). We find that all of our three positive sampling strategies confer performance improvements over the base focal loss, specifically 2%, 3%, and 3% for random, feature, and attribute respectively. We then compare the overall contribution of the two sampling approaches we developed (feature and attribute) compared to the random k-positive selection for focal loss.

We further assess the relative benefit to these sampling approaches in different EHR cohort scenarios in subsequent generalizability experiments described below.

Table 1: Overall Mortality Prediction Performance

Loss Model	AUROC	AUPRC
CE	0.873(0.003)	0.801(0.002)
FL	0.931(0.002)	0.830(0.002)
FL(random)	0.949(0.002)	0.879(0.001)
FL(feature)	0.956(0.002)	0.884(0.001)
FL(attribute)	0.959(0.002)	0.886(0.001)

Table 2: Prediction Performance on Various Sample Sizes in the same Test Set

Loss Model	Training data(size/positive label percentage)				
	399/23%	999/23%	1999/23%	2999/23%	3999/23%
CE	0.732(0.003)	0.845(0.002)	0.863(0.002)	0.876(0.002)	0.887(0.002)
FL	0.803(0.002)	0.916(0.002)	0.926(0.002)	0.934(0.002)	0.935(0.001)
FL(random)	0.886(0.001)	0.931(0.002)	0.941(0.001)	0.947(0.001)	0.949(0.001)
FL(feature)	0.897(0.001)	0.937(0.001)	0.944(0.001)	0.952(0.001)	0.956(0.001)
FL(attribute)	0.906(0.001)	0.938(0.001)	0.949(0.001)	0.953(0.001)	0.954(0.002)

4.3. Assessing the Robustness of the Contrastive Regularizer Framework with Different Training Sample Sizes and Imbalance Ratios

Next, we test the robustness of our model and baselines to predict the same outcome in subsets of our data at different sample sizes and imbalance ratios in the training data. Specifically, we train on varying dataset characteristics (i.e. size and imbalance ratio) and test in the fixed original size and imbalance ratio as the original experiment (N=1713). Like before, we assess various baselines and implementations of positive sampling strategies.

First, the AUROC results of varying training sample sizes are shown in Table 2. Specifically, we compared performance at N=399, 999, 1999, 2999, and 3999 all with the same imbalance ratio of 23% positive outcomes. Across all scenarios, we see a marked improvement of focal loss over cross-entropy loss with the larger improvements seen at smaller sample sizes (i.e., 7% improvement in N=1999). The addition of random positive sampling for focal loss increased performance across all experiments but was most prominent in the smallest sample size, specifically 8% improvement over base focal loss. The incorporation of feature- and attribute-based sampling strategies also demonstrated improvements over the random sampling strategy in focal loss: 1.1% for feature-based and 2.0% for attribute-based. For these experiments, we did not find a large difference between the feature-based and attribute-based strategies with slightly higher values in attribute-based.

We next performed a similar experiment varying degrees of data imbalance via restricting the number of positive outcome patients (same amount of negative) in the training data, specifically at 1%, 5%, 10%, 15%, and 20%. Table 3 shows the AUROC results for this experiment. Like before, focal loss has large improvements over cross-entropy loss, with the largest improvement at the most imbalanced ratio, specifically 15% improvement at

Table 3: Prediction Performance When Varying Training Outcome Imbalance Ratios

Loss Model	Training data(size/positive label percentage)				
	3099/1%	3245/5%	3445/10%	3645/15%	3845/20%
CE	0.693(0.002)	0.801(0.002)	0.828(0.002)	0.841(0.002)	0.847(0.002)
FL	0.849(0.002)	0.884(0.002)	0.901(0.002)	0.927(0.002)	0.932(0.002)
FL(random)	0.929(0.001)	0.934(0.001)	0.939(0.001)	0.944(0.001)	0.949(0.001)
FL(feature)	0.934(0.001)	0.942(0.001)	0.944(0.001)	0.943(0.001)	0.953(0.002)
FL(attribute)	0.940 (0.002)	0.942 (0.001)	0.948 (0.001)	0.949 (0.001)	0.959 (0.001)

1% imbalance. All contrastive regularizers with different positive sampling greatly boost the performance of focal loss, specifically 9% improvement. All three positive sampling strategies conferred improvements over the baseline focal loss with the trend of showing bigger improvements with higher training imbalance. The feature- and attribute-based approaches still show improvements over random sampling across different imbalance ratios, with an average 0.07% for feature-based and 0.11% for attribute-based. The attribute-based sampling has slight improvement over feature-based sampling, one average 0.04% at different ratios, note that all contrastive regularizer boosted focal losses have much stable performance at different ratios, with the largest degradation of 2% from 20% ratio to 1% ratio, comparing to the degradation of 15% from cross-entropy loss and 8.3% from focal loss, our contrastive regularizer boosted focal loss provides much stable performance.

4.4. The Impact of Different Positive Sampling Strategies on Representations Learned from EHR Data

In all previous experiments, we used the contrastive-regularized focal loss in a semi-supervised setting to directly classify patients from the trained model. Table 1 already demonstrated the benefit of our contrastive regularizer with an improved predictive performance in the semi-supervised setting. Here, we demonstrate the impact of different sampling strategies on the hidden representations learned via supervised contrastive learning (Khosla et al., 2020). In this setting, we learn 100-dimension hidden representation of patients using the contrasting loss functions shown in Eq. 2, 4, and 6. Fig. 1 shows the embeddings obtained from different sampling strategies.

Table 4: Pre-trained Embedding Evaluation for Sampling Strategies

Sampling Strategy	AUROC	ESS	SD (positive)	SD (negative)
Random	0.859	0.446	0.158	0.448
Feature	0.893	0.627	0.193	0.534
Attribute	0.904	0.803	0.255	0.702

Feature- and attribute-based positive samplings better separate classes in embedding space. To test the separation of classes in embedding space, we train a logistic linear classifier on patient embeddings obtained from contrastive learning with different

sampling strategies. To train the logistic regression model, we use the same split percentage for training and testing data as used in the above section. The first column of Table 4 shows that the prediction scores of the linear classifier trained on both attribute- and feature-based embedding outperform the random sampling strategy, specifically 3.4% for feature-based, and 4.5% for attribute-based. This result demonstrates that feature- and attribute-based samplings push positive and negative classes further apart in embedding space, which helps the linear classifier attain better accuracy.

To quantitatively measure the separation between the positive and negative patient groups, we define a simple inter-class distance metric (here we use Embedding Separation Score (ESS) for this metric) as follows:

$$\|z_p - z_n\| / (\|z_p\| + \|z_n\|), \quad (7)$$

where z_p and z_n are normalized embedding centers for positive patient group and negative patient group. Here, the value of inter-class distance is between $[0, 1]$, and the higher the distance the better the classes are separated in the embedding space. Table 4 column 2 shows that feature-based sampling and attribute-based sampling show higher inter-class distance, as expected. Fig. 1 visually confirms that the new sampling strategies indeed better separate positive and negative classes.

Feature- and attribute-based positive samplings capture more intra-class heterogeneity. Usually, it is desirable that points from the same class are pulled together in embedding space to create a compact cluster. However, COVID EHR data shows tremendous heterogeneity in terms of patients demographics, symptoms, and outcomes (Su et al., 2021). Hence, it is often beneficial to keep some intra-class variability (heterogeneity) to facilitate learning from local neighborhoods of patients across the spectrum within each class. We quantitatively measure the intra-class variance (in the embedding space) to represent intra-class heterogeneous by computing the standard deviation of distance metric between patient embedding with its cluster center embedding in terms of Eq. 7 for the positive class (mortality) patient group and negative patient group, respectively. Columns 4 and 5 in Table 4 show that feature-based and attribute-based sampling indeed preserve higher intra-class variances to reflect the heterogeneity within each class. This benefit is because random positive sampling contrastive loss approximates one uniform distribution within each group, resulting in collapsing the embedding space, while our feature- or attribute-based positive sampling contrastive loss strategies approximate different subgroup distributions within the same class and avoid collapsing of the embedding space. Such collapsing could result in a much closer distribution center between the positive and negative group as shown in Table 4, thus leading to worse prediction performance.

5. Conclusion

In this work, we introduce a general framework that adds a contrastive regularizer on top of focal loss for boosting predictive performance. We further propose two novel positive sampling strategies, feature-based and attribute-based, that outperform k random sampling for contrastive learning especially in datasets with high intra-class heterogeneity and data imbalance. Through experiments predicting mortality in real-world COVID-19 EHR data, we demonstrate the contrastive regularized framework greatly boosts the performance over

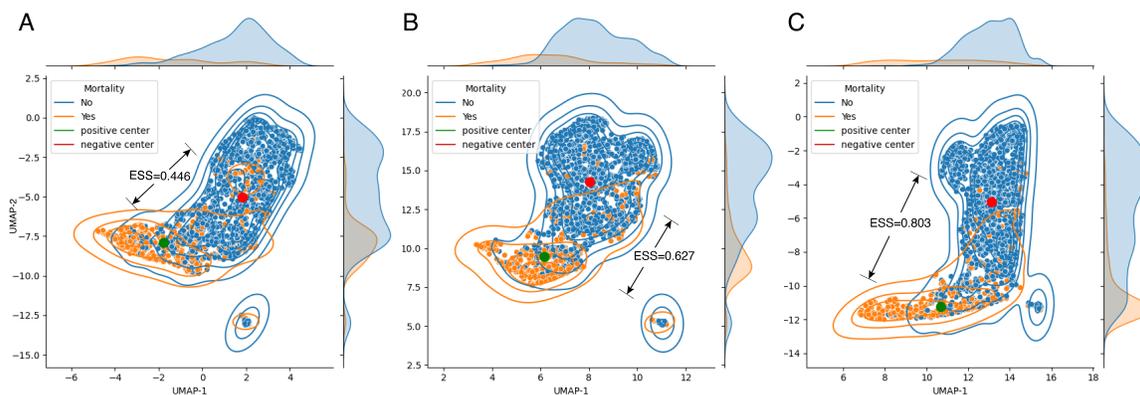


Figure 1: UMAP visualizations of pre-trained embedding with different sampling strategies: (A) random positive sampling; (B) feature-based positive sampling; (C) attribute-based positive sampling. Blue dots are non-death patient embeddings, orange dots are death patient embeddings, the orange and blue lines represent distribution contour of the positive and negative groups.

focal loss at various sample sizes and imbalance ratios. Our results show that the two sampling strategies both outperform k random sampling in this task with the attribute-based approach having a slight edge over the feature-based approach. Our experiments further confirm that the two proposed sampling strategies in our contrastive regularized framework can achieve better inter-class separation and leverage intra-class heterogeneity.

References

- A. M. Alaa, J. Yoon, S. Hu, and M. van der Schaar. Personalized risk scoring for critical care prognosis using mixtures of gaussian processes. *IEEE Transactions on Biomedical Engineering*, 65(1):207–218, 2018. doi: 10.1109/TBME.2017.2698602.
- Peter C Austin, Jack V Tu, Jennifer E Ho, Daniel Levy, and Douglas S Lee. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology*, 66(4):398–407, 2013.
- Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.
- Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

- Yen-Pin Chen, Yuan-Hsun Lo, Feipei Lai, and Chien-Hua Huang. Disease concept-embedding based on the self-supervised method for medical information extraction from electronic health records and disease retrieval: Algorithm development and validation study. *Journal of Medical Internet Research*, 23(1):e25113, 2021.
- Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.
- Shahram Ebadollahi, Jimeng Sun, David Gotz, Jianying Hu, Daby Sow, and Chalapathy Neti. Predicting patient’s trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics. In *AMIA annual symposium proceedings*, volume 2010, page 192. American Medical Informatics Association, 2010.
- Jean-Bastien Grill, Florian Strub, Florent Althché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.
- Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606, 2011.
- Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.
- Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021.
- Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11(1):1–13, 2011.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals. *arXiv preprint arXiv:2005.13249*, 2020.

- Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *arXiv preprint arXiv:2101.12037*, 2021.
- Gilad J Kuperman, Anne Bobb, Thomas H Payne, Anthony J Avery, Tejal K Gandhi, Gerard Burns, David C Classen, and David W Bates. Medication-related clinical decision support in computerized provider order entry systems: a review. *Journal of the American Medical Informatics Association*, 14(1):29–40, 2007.
- S J G Lewis, T Foltynie, A D Blackwell, T W Robbins, A M Owen, and R A Barker. Heterogeneity of parkinson’s disease in the early clinical stages using a data driven approach. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(3):343–348, 2005.
- Chunyuan Li, Xiujun Li, Lei Zhang, Baolin Peng, Mingyuan Zhou, and Jianfeng Gao. Self-supervised pre-training with hard examples improves visual representations. *arXiv preprint arXiv:2012.13493*, 2020.
- Ziyi Li, Kirk Roberts, Xiaoqian Jiang, and Qi Long. Distributed learning from multiple ehr databases: contextual embedding models for medical events. *Journal of biomedical informatics*, 92:103138, 2019.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Sara Santiso, Arantza Casillas, and Alicia Pérez. The class imbalance problem detecting adverse drug reactions in electronic health records. *Health informatics journal*, 25(4):1768–1778, 2019.
- Peter Schulam, Fredrick Wigley, and Suchi Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.
- Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1857–1865, 2016.
- Chang Su, Yongkang Zhang, James H Flory, Mark G Weiner, Rainu Kaushal, Edward J Schenck, and Fei Wang. Novel clinical subphenotypes in covid-19: derivation, validation, prediction, temporal patterns, and interaction with social determinants of health. *medRxiv*, 2021.

- Harini Suresh, Jen J Gong, and John V Guttag. Learning tasks for multitask learning: Heterogenous patient populations in the icu. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 802–810, 2018.
- Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.
- Liansheng Wang, Qiuhaio Xu, and Shuo Li. Utility balanced classification for automatic electronic medical record analysis. In *2018 5th International Conference on Systems and Informatics (ICSAI)*, pages 1093–1098. IEEE, 2018.
- Zhe Wang, Yiwen Zhu, Dongdong Li, Yichao Yin, and Jing Zhang. Feature rearrangement based deep learning system for predicting heart failure mortality. *Computer methods and programs in biomedicine*, 191:105383, 2020.
- Tingyi Wanyan, Hossein Honarvar, Suraj K Jaladanki, Chengxi Zang, Nidhi Naik, Sulaiman Somani, Jessica K De Freitas, Ishan Paranjpe, Akhil Vaid, Riccardo Miotto, et al. Contrastive learning improves critical event prediction in covid-19 patients. *arXiv preprint arXiv:2101.04013*, 2021.
- Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.
- Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, pages S106–S113, 2010.
- Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Appendix A. Input Features For EHR data

For our experiment, the collected EHR data contains the following specific features: COVID-19 status and demographics (age, gender, and race). We also collect vital signs, specifically: heart rate, respiration rate, pulse oximetry, blood pressure (diastolic and systolic), temperature, height, and weight. We compile 12 comorbidities: alcoholism, asthma, atrial fibrillation, coronary artery disease, cancer, chronic kidney disease, chronic obstructive pulmonary disease, diabetes mellitus, heart failure, hypertension, stroke, and liver disease. Lastly, we collect 34 relevant laboratory test results, these lab tests are specifically listed as follows:

Albumin, Alkaline Phosphatase, Alanine Aminotransferase, Amylase, Anion Gap, Partial Thromboplastin Time, Aspartate transaminase, Atypical Lymphocytes, Band Cells, Basophil %, Basophil Count, Direct Bilirubin, Total Bilirubin, Myeloblasts, Brain Natriuretic Peptide, Blood Urea Nitrogen, C-reactive protein, Ionized Calcium, Calcium, Chloride, Creatine phosphokinase, Creatine kinase-MB, Bicarbonate, Creatinine, D-dimer, Eosinophil %, Eosinophil, Ferritin, Fibrogen, Glucose, Hematocrit, Hemoglobin, International Normalized Ratio, IL6, Iron, Ketone, Lactate, Lactate Dehydrogenase, Lymphocyte %, Lymphocytes, Mean Corpuscular Hemoglobin Concentration, Mean Corpuscular Volume, Mean Platelet Volume, Monocyte %, Monocytes, Neutrophil %, Neutrophils, Oxygen Saturation, pH, Platelets, Partial Pressure of Oxygen, Potassium, PT, Protein, Red Blood Cell Count, Red Cell Distribution Width, Sodium, Total Iron-Binding Capacity, Transferrin saturation, Troponin I, Uric Acid, White Blood Cell Count

Appendix B. Studying the Effect of Model Parameters for Contrastive Regularizer

We assess the effect of different parameters for our three positive sampling contrastive regularizer models. There are three model parameters that could be customized and optimized in our contrastive learning framework, specifically: k, α, τ , where k is the positive sample numbers, α is the weight for contrastive regularizer, τ is the temperature parameter for contrastive loss. First, we test on the performance using different k values, with $\alpha = 0.2$ and $\tau = 1$ fixed. the result is shown in Table 5.

Table 5: Prediction Performance on Different k

Loss Model	k=1	k=3	k=5	k=9	k=11	k=15
FL(random)	0.949	0.951	0.950	0.953	0.952	0.953
FL(feature)	0.950	0.957	0.956	0.957	0.953	0.955
FL(attribute)	0.956	0.957	0.954	0.961	0.956	0.961

Second, we test the performance on the impact of different regularizer weight coefficient α with $k = 5$ and $\tau = 1$ fixed. The results of this experiment are shown in Table 6

Finally, we test the parameter τ with $k = 5$ and $\alpha = 0.2$ fixed. The results are shown in Table 7.

Table 6: Predictive Performance using Different Weight Coefficients α

Loss Model	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
FL(random)	0.946	0.949	0.953	0.955	0.956
FL(feature)	0.952	0.953	0.956	0.955	0.955
FL(attribute)	0.951	0.954	0.958	0.958	0.957

Table 7: Predictive Performance using Different τ

Loss Model	$\tau = 0.1$	$\tau = 0.5$	$\tau = 1$	$\tau = 1.5$
FL(random)	0.961	0.962	0.953	0.954
FL(feature)	0.963	0.965	0.956	0.955
FL(attribute)	0.967	0.966	0.958	0.957