Multi-modal Contrastive Learning for Healthcare Data Analytics

Rui Li
Computer Science and Engineering
University at Buffalo
Buffalo, NewYork, USA
rli35@buffalo.edu

Jing Gao
School of Electrical and Computer Engineering
Purdue University
West lafayetta, IN, USA
jinggao@purdue.edu

2022 IEEE 10th International Conference on Healthcare Informatics (ICHI) | 978-1-6654-6845-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICHI54592.2022.00029

***Abstract*—Electronic Health Record (EHR) is a digital version of patient's medical charts. EHR consists of longitudinal multi-modal data including demographics, diagnosis, clinical notes and clinical features. Plenty of data analytics works have been performed on EHR data. Among them, predictive modeling has been widely explored and most researches use single modality to perform the prediction task. Comparing with previous researches using single modal data, utilizing the multi-modal data can boost the prediction performance for a variety of downstream tasks. In this paper, in order to maintain the hierarchy of diagnosis codes, we embed diagnosis codes in hyperbolic space, and we utilize a hyperbolic transformer to model the sequential diagnosis information in multiple admissions. Meanwhile, we use multi-modal contrastive loss to capture the relation between diagnosis and clinical features. And we propose supervised contrastive loss in the multi-label setting. We perform two downstream tasks including diagnosis prediction and mortality prediction on two public datasets. Experiments on real-world datasets demonstrate the effectiveness of multi-modal contrastive loss in healthcare.**

***Index Terms*—multi-modal contrastive learning, diagnosis prediction, mortality prediction, hyperbolic neural networks**

I. Introduction

Recently, AI techniques have been widely used in healthcare, and powerful AI techniques can find clinically relevant information hidden in the massive amount of data, and in turn are able to support clinical decision making. The prevalence of AI in healthcare benefits from not only the development of AI techniques, but also the increasing availability of healthcare data. Electronic Health Record (EHR) data is one of the most representative healthcare data, and it contains longitudinal multi-modal data including demographics, diagnosis, clinical notes and clinical features. Usually EHR data is heterogeneous and noisy. Various types of features (variables) are included in EHR, including categorical variables (e.g., medical codes), numerical variables (e.g., clinical measurements), and textual variables (e.g., clinical notes). Meanwhile, data in each modality has its own characteristics. For example, diagnosis codes are extremely sparse and have the hierarchy tree-like structure, while clinical features contain a lot of missing data due to the irregular and incomplete recording.

Plenty of predictive models have been applied on EHR data. The key idea is to use sequence models to keep track of sequential information, and performing the prediction tasks based on the representation learned via the sequence model. Most predictive models use single modal data as input. However, EHR contains multi-modal data, and data of different modalities is closely related. Thus, incorporating multi-modal data is able to improve the prediction performance for a variety of downstream tasks.

Recently, multi-modal contrastive learning has been explored in the field of video representations [1], [2]. Inspired by these works, we use multi-modal contrastive learning framework to improve the performance of predictive models in healthcare. Our multi-modal contrastive learning framework uses diagnosis modality and clinical feature modality as input, and we perform two downstream tasks including diagnosis prediction and mortality prediction.

Our multi-modal contrastive learning framework uses different modules to learn the representations of different modalities. Because hyperbolic space is naturally equipped to model hierarchical tree-like structure, we learn the Poincaré embeddings in hyperbolic space and use hyperbolic transformer to learn the diagnosis representation. We choose GRU-D proposed in [3] as backbones to learn the clinical feature representation based on the irregular and incomplete characteristic of clinical features. Note that we can use other backbones as well. For diagnosis prediction task, we propose multi-label contrastive loss which also utilizing the hierarchy of the labels. And we further extend it to multi-modal version. We perform two downstream tasks, diagnosis prediction and mortality prediction, on two public avaliable EHR datasets. The performance shows the effectiveness of our multi-modal contrastive learning framework.

Our main contributions are summarized as follows:

- We apply the multi-modal contrastive learning framework to improve the performance of predictive models in healthcare by using the data of diagnosis modality and the data of clinical feature modality. We perform two downstream tasks including diagnosis prediction and mortality prediction.
- Based on the tree-like structure of diagnosis codes, we learn the Poincaré embeddings of diagnosis codes, and use hyperbolic transformer to model the sequential diagnosis codes among multiple admissions.
- For diagnosis prediction task, we propose multi-label contrastive loss. And we further extend it to multi-modal version.

2575-2634/22/$31.00 ©2022 IEEE
DOI 10.1109/ICHI54592.2022.00029

- We empirically show that the multi-modal contrastive learning outperforms existing methods on two real-world EHR datasets. The ablation study shows the effectiveness of every module in the framework.

II. Related Work

A. Diagnoses Prediction

Diagnosis prediction, which aims to predict the patient's health condition of the next admission based on their historical EHR, is an important task in health informatics, and thus has been widely studied. Most of the previous studies apply sequence models such as Long Short-Term Memory networks (LSTM) to model the sequential EHR data [4], [5], and some focus on learning the embedding of diagnoses codes based on the hierarchy of disease concepts [6]–[9]. Recently, researchers also explore the use of multi-modal data to perform the diagnosis prediction task. Some studies explore to incorporate patients' demographic information and clinical features to boost the performance [10]–[12].

B. Mortality Prediction

Mortality prediction refers to predict in-hospital mortality based on the clinical features collected on the first 48 hours of the patient admission. Clinical features include vital signs and lab test results, and they may be recorded irregularly, and different clinical features may be missing at different time stamps. In this way, most previous researches regard mortality prediction as a binary prediction task based on multivariate irregularly-sampled time series with missing data [3], [13]–[17]. Some studies try to impute the missing data first, and feed it into a time-aware GRU to take into account the different time intervals [3], [13]–[15]. The imputation methods include simple methods such as mean imputation [3] and complex methods such as Generative Adversarial Networks(GAN) [14], [15]. Meanwhile, other studies interpolate the input time series against a set of evenly distributed reference time stamps, and then use the original GRU as the prediction network [17].

C. Multi-modal Contrastive Learning

Contrastive learning algorithm creates different augmented views of the same data example, and maximize the agreement of the original data and the augmented versions via a contrastive loss in the latent space, and it is widely used in self-supervised and semi-supervised learning [18]–[20]. Recently, some researches extend the concept of contrastive loss to multi-modal domain. Besides computing the intra-modal contrastive loss between the original data and the augmented version, they also compute the inter-modal contrastive loss between different modalities, such as video, audio and text [1], [2]. In this paper, we regard diagnoses code and clinical features as two related modalities and compute the multi-modal contrastive loss to boost the performance.

D. Hyperbolic Neural Networks and Embeddings

Hyperbolic space is a space with constant negative curvature and can be thought of as a continuous version of trees, which is naturally equipped to model hierarchical tree-like structure. Plenty of previous works explore learning the embedding and constructing neural networks in hyperbolic space [21]–[23]. The disease taxonomy contains the hierarchy of disease concepts in the form of a tree-like relationship, thus hyperbolic embedding can be applied to boost the prediction performance. Although there are several researches use the hyperbolic embedding of diagnoses codes [9], [24], they only use the embedding in Hyperbolic space and the whole model is constructed on Euclidean space, which may cause performance degradation. In this paper, we use the hyperbolic transformer proposed in [23] to model the sequential EHR data.

III. Problem Statement

EHR Data. For each patient, the clinical record can be viewed as a sequence of admissions $V_1, \ldots, V_T$, where each admission record V_t contains diagnosis information $\mathbf{x}_t$ and clinical features $\mathbf{c_t}$. The diagnosis information $\mathbf{x_t} \in \{\mathbf{0}, \mathbf{1}\}^{|\mathcal{D}|}$ is a multi-hot binary vector, where $|\mathcal{D}|$ is the number of unique diagnosis codes, and $\mathcal{D} = \{d_1, d_2, \ldots, d_{|\mathcal{D}|}\}$. $x_{t,i} = 1$ indicates that the patient was diagnosed with disease d_i in the t-th admission; otherwise 0. For clinical features $\mathbf{c_t} \in \mathbb{R}^{N \times T_t}$, N is the number of clinical features that we are interested in, and T_t is the length of the ICU stay in the t-th admission. The clinical features include vital signs and some lab test results, and $\mathbf{c_t}$ may contain some missing data at different time stamps for different clinical features.

Disease Taxonomy. Let $\mathcal{G}$ denote the disease taxonomy, which contains the hierarchy of disease concepts in the form of a tree-like relationship, and the diagnosis codes in $\mathcal{D}$ are the leaf nodes. We define $\mathcal{D}' = \{d_{|\mathcal{D}|+1}, d_{|\mathcal{D}|+2}, \ldots, d_{|\mathcal{D}|+|\mathcal{D}'|}\}$ as the set containing the ancestor codes, and all nodes in $\mathcal{G}$ form the set $\mathcal{C} = \mathcal{D} + \mathcal{D}'$. We construct $\mathcal{G}$ using the multi-level diagnoses CCS categories[1].

Diagnosis Prediction Task. Given the patient's diagnosis information $\mathbf{x_t}$, clinical features $\mathbf{c_t}$, and the disease ontology $\mathcal{G}$, the goal of this task is to predict diagnosis codes of the next admission denoted as $\hat{\mathbf{y}}_{t+1}$. This is a multi-label classification task.

Mortality Prediction Task. Given the patient's diagnosis information $\mathbf{x_t}$, clinical features $\mathbf{c_t}$, and the disease ontology $\mathcal{G}$, the goal of this task is to predict the in-hospital mortality based on the first 48 hours of the current admission denoted as $\hat{y}_t$. This is a binary classification task.

IV. Methodology

Figure 1 shows the overview of the proposed multi-modal contrastive learning framework, which mainly contains three parts: (1) Diagnosis code encoder that learns the patient's diagnosis representation in hyperbolic space; (2) Clinical feature encoder that learns the patient's clinical feature representation

[1]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

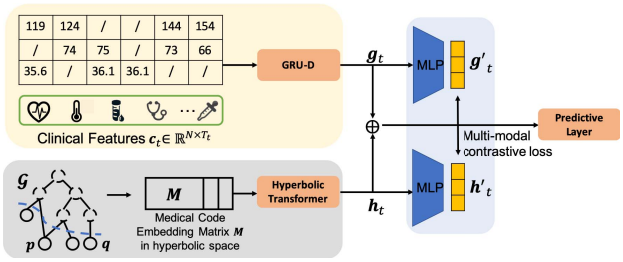


Fig. 1. The overall framework

in euclidean space; (3) Multi-modal contrastive loss that captures the relation between diagnosis and clinical feature modalities. First, we introduce the Poincaré ball model in hyperbolic space, and then introduce the above three parts and show how they are adapted for different prediction tasks. Finally, we show the prediction layer and objective functions.

A. Poincaré Ball Model in Hyperbolic Space

A hyperbolic space is a Riemanniam manifold with a constant negative curvature, one of the most representative hyperbolic space is Poincaré ball model [25]. The n-dimensional Poincare ball $\mathbb{H}_c^n = \{\boldsymbol{x} \in \mathbb{R}^n | c\|\boldsymbol{x}\| < 1\}$ is an open ball of radius $c^{-\frac{1}{2}}$. Given two points $\boldsymbol{u} \in \mathbb{H}_c^n$ and $\boldsymbol{v} \in \mathbb{H}_c^n$, the hyperbolic distance between them is defined as

$$d_{\mathbb{H}_c^n}(\boldsymbol{u}, \boldsymbol{v}) = (2/\sqrt{c}) tanh^{-1}(\sqrt{c}\| - \boldsymbol{u} \oplus_c \boldsymbol{v}\|) \tag{1}$$

Here, $\oplus_c$ is the Möbius addition defined in Equation 4. The exponential map $exp_{\mathbf{0}}^c$ gives a way to project a vector $\boldsymbol{v}$ from euclidean space $\mathbb{E}^n$ to hyperbolic space $\mathbb{H}_c^n$.

$$exp_{\mathbf{0}}^c(\boldsymbol{v}) = tanh(\sqrt{c}\|\boldsymbol{v}\|)\frac{\boldsymbol{v}}{\sqrt{c}\|\boldsymbol{v}\|} \tag{2}$$

The logarithmic map $log_{\mathbf{0}}^c$ gives a way to project a vector $\boldsymbol{y}$ from hyperbolic space $\mathbb{H}_c^n$ to euclidean space $\mathbb{E}^n$.

$$log_{\mathbf{0}}^c(\boldsymbol{y}) = tanh^{-1}(\sqrt{c}\|\boldsymbol{y}\|)\frac{\boldsymbol{y}}{\sqrt{c}\|\boldsymbol{y}\|} \tag{3}$$

If $\boldsymbol{x} \in \mathbb{H}_c^n$ and $\boldsymbol{y} \in \mathbb{H}_c^n$, the Möbius addition is defined as

$$\boldsymbol{x} \oplus_c \boldsymbol{y} := \frac{(1 + 2c\langle \boldsymbol{x}, \boldsymbol{y}\rangle + c\|\boldsymbol{y}\|^2)\boldsymbol{x} + (1 - c\|\boldsymbol{x}\|^2)\boldsymbol{y}}{1 + 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle + c^2\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2} \tag{4}$$

Given the disease taxonomy with a hierarchical tree-like structure, we compute the Poincaré embeddings $\boldsymbol{m}_i \in \mathbb{H}_c^n$ $(1 \leq i \leq |\mathcal{D}|)$ of diagnosis codes via the method proposed in [21]. We use the following loss function to obtain the pretrained embeddings of diagnosis codes.

$$\mathcal{L}_{pre} = -\sum_{(\boldsymbol{p_i}, \boldsymbol{p_j}) \in \mathcal{A}} log \frac{e^{-d(\boldsymbol{p_i}, \boldsymbol{p_j})}}{\sum_{\boldsymbol{p'_j} \in \mathcal{N}(\boldsymbol{p_i})} e^{-d(\boldsymbol{p_i}, \boldsymbol{p'_j})}} \tag{5}$$

Here, $\mathcal{A} = \{(\boldsymbol{p_i}, \boldsymbol{p_j})\}$ represents the set of hypernymy relations between diagnosis codes, and $\mathcal{N}(\boldsymbol{p_i}) = \{\boldsymbol{p'_j} | (\boldsymbol{p_i}, \boldsymbol{p_j}) \notin \mathcal{D}\} \cup \{\boldsymbol{p_j}\}$ represents the set of negative samples for diagnosis code $\boldsymbol{p_i}$. We first use Equation 5 to obtain the pretrained embeddings of diagnosis codes $\boldsymbol{m}_i$, and then feed $\boldsymbol{m}_i$ into the following modules.

B. Diagnosis Code Encoder in Hyperbolic Space

For the t-th admission, we need to compute the vector $\boldsymbol{v}_t$ that contains all current diagnosis information. Previous methods in euclidean space compute $\boldsymbol{v}_t$ by simply adding the corresponding diagnosis embeddings. However, due to the non-commutative and non-associative properties of the Möbius addition, we can not use the same method in hyperbolic space. In this way, we first project the Poincaré embedding $\boldsymbol{m}_i$ to euclidean space via Equation 3 and get $\boldsymbol{m}'_i$, then compute $\boldsymbol{v}'_t \in \mathbb{E}^n$ via $\boldsymbol{v}'_t = tanh(\boldsymbol{M}'\boldsymbol{x}_t)$, $\boldsymbol{M}'$ is the diagnosis code embedding matrix obtained via concatenating the diagnosis embedding vectors $\boldsymbol{m}'_1, \boldsymbol{m}'_2, \cdots, \boldsymbol{m}'_{|\mathcal{D}|}$. Then we map $\boldsymbol{v}'_t$ back to hyperbolic space via Equation 2 and obtain $\boldsymbol{v}_t$.

For diagnosis prediction task, we use the hyperbolic transformer proposed in [23] to capture the dependencies among multiple admissions and we use the hidden state $\boldsymbol{h}_t \in \mathbb{H}_c^n$ as the diagnosis representation of the t-th admission.

For mortality prediction task, because it only makes prediction based on the current admission, without considering historical admission information, so we just use $\boldsymbol{v}_t \in \mathbb{H}_c^n$ as the diagnosis representation.

C. Clinical Feature Encoder in Euclidean Space

Clinical features are recorded irregularly, and most previous researches regard them as multivariate irregularly-sampled time series with missing data. We use GRU-D [3] to obtain the clinical feature representation $\boldsymbol{g}$. Note that we can also use other backbones to learn the clinical feature representation.

For mortality prediction task, because MIMIC dataset is highly skewed, we perform data augmentation for samples with true label in the training process. We perturb the clinical features by adding noises following Gaussian distribution with mean 0 and variant 0.1, and the augmented samples have the same diagnosis codes comparing with the original sample. This can be viewed as oversampling examples in the minority class.

D. Multi-modal Contrastive Loss

Before we compute the multi-modal contrastive loss, we first map the diagnosis representation $\boldsymbol{h}$ and clinical feature representation $\boldsymbol{g}$ into the same domain via a multiple layer perceptron (MLP) module, and we get $\boldsymbol{h}'$ and $\boldsymbol{g}'$. Then we define multi-modal contrastive loss for two prediction tasks respectively.

Diagnosis prediction is a multi-label classification task, and previous supervised contrastive loss [26] defining for single label classification setting can not be used directly. Considering the hierarchy of diagnosis codes, we define the multi-label contrastive loss below.

Firstly, we define the code-level similarity which measures the similarity between a pair of ICD codes as $f_c(p, q) = 1/2^n$, n is the number of steps to the first common parent of p and q. For example, in Figure 1, $f_c(p, q) = 1/2^2$. Then we define the set-level similarity. For two sets containing ICD codes, $S_1 = \{p_1, p_2, \cdots, p_n\}$ and $S_2 = \{q_1, q_2, \cdots, q_m\}$, the set-level similarity is defined as $f_s(S_1, S_2) = (\sum_{i=1}^{n}\sum_{j=1}^{m} f_c(p_i, q_j))/(f_s(S_1, S_1) +$

$f_s(S_2, S_2)$). Here, the denominator is for normalization, if two sets S_1 and S_2 have the same ICD codes, $f_s(S_1, S_2) = 1$. $\boldsymbol{x}_i^t$ is the set of diagnosis codes for sample i at the t-th admission, because the multi-label contrastive loss for diagnosis prediction task is computed among the same admission, we omit the superscript t in the following definition. The multi-label contrastive loss is defined as

$$\mathcal{L}_{ml} = -\sum_{i \in I} \frac{1}{\sum_{j \in A(i)} f_s(\boldsymbol{x}_i, \boldsymbol{x}_j)} \sum_{j \in A(i)} f_s(\boldsymbol{x}_i, \boldsymbol{x}_j) log \frac{exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_j / \tau)}{\sum_{j \in A(i)} f_s(\boldsymbol{x}_i, \boldsymbol{x}_j) exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_j / \tau)} \quad (6)$$

I is the set of samples in the current batch, $A(i) = I \backslash \{i\}$, $\boldsymbol{z}_i$ is the diagnosis representation of sample i, and τ is the temperature that controls the strength of penalties on dissimilar samples. We can further extend the above supervised multi-label contrastive loss to the multi-modal version through the alignment of diagnosis representation and clinical feature representation. The contrastive loss from diagnosis modality to clinical feature modality is defined as

$$\mathcal{L}_{d \to c} = -\sum_{i \in I} \frac{1}{\sum_{j \in I} f_s(\boldsymbol{x}_i, \boldsymbol{x}_j)} \sum_{j \in I} f_s(\boldsymbol{x}_i, \boldsymbol{x}_j) log \frac{exp(\boldsymbol{h}_i' \cdot \boldsymbol{g}_j' / \tau)}{\sum_{j \in I} f_s(\boldsymbol{x}_i, \boldsymbol{x}_j) exp(\boldsymbol{h}_i' \cdot \boldsymbol{g}_j' / \tau)} \quad (7)$$

$\boldsymbol{h}_i'$ is the transformed diagnosis representation, and $\boldsymbol{g}_i'$ is the transformed clinical feature representation for sample i. The contrastive loss from clinical feature modality to diagnosis modality is defined vice versa.

Mortality prediction task is a binary classification task, and we can extend the supervised contrastive loss proposed in [26] to multi-modal version directly. Inspired by [1], we align diagnosis representation and clinical feature representation via multi-modal contrastive loss. For each anchor, positives are defined as the samples having the same mortality label comparing with the anchor, and the rest samples in the same batch are defined as negatives. The contrastive loss from diagnosis modality to clinical feature modality is defined as

$$\mathcal{L}_{d \to c} = -\sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(\boldsymbol{h}_i' \cdot \boldsymbol{g}_p' / \tau)}{\sum_{j \in A(i)} exp(\boldsymbol{h}_i' \cdot \boldsymbol{g}_j' / \tau)} \quad (8)$$

Here, I is the set of samples in the current batch, $A(i) = I \backslash \{i\}$, and $P(i) \coloneqq \{p \in I | y_j = y_i\}$, which is the set of positives given sample i as anchor. $\boldsymbol{h}_i$ is the diagnosis embedding, and $\boldsymbol{g}_i$ is the clinical feature embedding. The contrastive loss from clinical feature modality to diagnosis modality is defined vice versa.

For both prediction tasks, the overall multi-modal contrastive loss is defined as

$$\mathcal{L}_{ct} = \mathcal{L}_{d \to c} + \mathcal{L}_{c \to d} \quad (9)$$

Through contrastive loss, diagnosis representation can be adjusted according to clinical feature representations via back propagation, and vice versa.

E. Prediction Layer and Objective Function

For diagnosis prediction task, for the t-th admission, we first obtain the diagnosis representation $\boldsymbol{h}_t \in \mathbb{H}_c^n$ via the Diagnosis Code Encoder, then we obtain the clinical feature representation $\boldsymbol{g}_t \in \mathbb{E}^{n'}$ via backbone GRU-D [3]. Note that $\boldsymbol{h}_t$ and $\boldsymbol{g}_t$ are in different space and have different dimension, we first extend $\boldsymbol{g}_t$ to the same dimension of $\boldsymbol{h}_t$ via a linear layer, and then we map $\boldsymbol{g}_t$ to $\mathbb{H}_c^n$ via Equation 2 to obtain $\boldsymbol{g}_t' \in \mathbb{H}_c^n$. And we get $\boldsymbol{f}_t = \boldsymbol{h}_t \oplus_c \boldsymbol{c}_t$ via Equation 4, we consider that $\boldsymbol{f}_t$ contains the diagnosis representation and clinical feature representation at the same time. And we use the unidirectional Poincaré MLR defined in [23] as the predictive layer. The multi-label classification loss for sample i is defined as

$$\mathcal{L}_c = -\frac{1}{T-1} \sum_{t=1}^{T=1} (\mathbf{y}_t^T log(\hat{\mathbf{y}}_t) + (1 - \mathbf{y}_t)^T log(1 - \hat{\mathbf{y}}_t)). \quad (10)$$

Here, $\boldsymbol{y}_t$ is the ground truth, which is a multi-hot binary vector indicating the patient's diagnosis information, and $\hat{\boldsymbol{y}}_t$ is the predicted result.

For mortality prediction task, we obtain the diagnosis representation $\boldsymbol{h} \in \mathbb{E}^n$ and the clinical feature representation $\boldsymbol{g} \in \mathbb{E}^n$, and we concatenate $\boldsymbol{h}$ and $\boldsymbol{g}$ to obtain $\boldsymbol{f}$. $\boldsymbol{f}$ is fed into a linear layer with output dimension 1 and followed by a sigmoid layer for classification. The binary classification loss for sample i is defined as

$$\mathcal{L}_c = -(y \cdot log(\hat{y}) + (1 - y) \cdot log(1 - \hat{y})). \quad (11)$$

y is the mortality label, either 0 or 1, and $\hat{y}$ is the predicted result.

For both prediction tasks, we define the objective function as

$$\mathcal{L} = \mathcal{L}_c + \alpha \cdot \mathcal{L}_{ct} \quad (12)$$

$\mathcal{L}_c$ is the classification loss, $\mathcal{L}_{ct}$ is the contrastive loss, and α is the hyper-parameter.

V. Experimental Settings

Tasks and Evaluation Measures. We perform two prediction tasks, diagnoses prediction and mortality prediction. Diagnoses prediction aims to predict the patient's diagnosis codes in the next admission, and mortality prediction aims to predict the patient's in-hospital mortality after the first 48 hours of the current admission.

For diagnoses prediction, we use Recall@K and MAP@K as the evaluation criteria. For every admission V_t, we get a 1 if the target diagnoses code appears in the top k predictions and 0 otherwise. Recall@K is defined as the number of diagnoses codes that are predicted correctly in the top k of $\hat{\mathbf{y}}_t$ divided by the total number of diagnosis code in V_t. MAP@K refers to mean average precision, and it considers not only the precision and accuracy but also the order of diagnoses codes which are predicted correctly. We vary K from 10 to 30. Instead of predicting diagnosis categories like most of the previous research [5]–[7], we aim to predict the real diagnosis codes.

TABLE I
DATASET STATISTIC

Statistics	MIMIC-III		MIMIC-IV	
	D	M	D	M
# of patients	6,461	-	4233	-
# of admissions	17,189	50,260	10,928	26,062
# of unique diagnoses code	4,537	6,668	4,366	6,218
Avg. # of diagnoses code	13.31	11.18	18.364	16.285
# of unique CCS code	657	664	643	699
# of levels of CCS code	4	4	4	4
# of clinical features	12	12	12	12
Avg. # of hours per admission	134.091	-	152.162	-
Avg. miss rate of clinic feature	0.619	0.646	0.662	0.661
Percentage of hospital death	-	0.082	-	0.092

This means that our task is more difficult since the target space is much larger.

For mortality prediction, we use AUC-ROC and AUC-PRC as the evaluation criteria. MIMIC dataset is highly imbalanced with less than 10% patients died before hospital discharge, and died patients are regarded as positive samples. AUC-ROC is the area under the receiver operator characteristic curve, and AUC-PRC is the area under the precision-recall curve. For the above two measures, the higher the better.

Dataset. MIMIC-III and MIMIC-IV [27] are publicly available EHR datasets, which consist the admission records of ICU patients. EHR data include diagnosis information such as diagnosis codes and drug codes, clinical features such as vital signs and lab tests results, and clinical notes charted by care providers. For diagnoses codes, we use ICD-9 codes, because MIMIC-IV contains ICD-9 and ICD-10 simultaneously, we filter patients containing ICD-10 codes. For clinical features, we select 12 clinical features following IPN [17] including heart rate, pH, blood pressure, etc. Meanwhile, we filter admissions with less than 10 records of clinical features, and we also filter admissions with no diagnoses codes.

For diagnoses prediction task, we select the patients who made at least two visits and we only use the clinical features collected in the first 500 records. We aim to predict the real diagnosis codes, instead of predicting diagnosis categories like most of the previous research [4]–[7]. For mortality prediction task, we use clinical features recorded in the first 48 hours. Meanwhile, we note that both datasets are highly skewed with less than 10% positives. And the average missing rate of clinical features shows that over 60% of clinical features are missing. Table I shows the details about the datasets, 'D' is the abbreviation of diagnosis prediction and 'M' is the abbreviation of mortality prediction. Because mortality prediction task makes prediction for every admission, the $\#ofpatients$ is meaningless, so we omit this term. Meanwhile, we use the clinical features collected in the first 48 hours of the admission, we also omit $Avg.\#ofhoursperadmission$ for mortality prediction task.

Baselines. For diagnosis prediction, we select seven baselines. RNN and Dipole [5] only use sequence models to model sequential EHR data, and they do not use the disease taxonomy information. GRAM [6], KAME [7], and HAP [8] use the disease taxonomy information, and they only use patients' historical diagnosis codes as input. CAMP [10] and MDP [12] use the disease taxonomy information and use the multi-modal data as input.

For mortality prediction, we select six baselines. GRU-D [3] and Brits [13] first impute the missing data , and feed the imputed data into a time-aware GRU whose hidden state fades according to time intervals. IPN [17] is an interpolation method, and the interpolation network interpolates the input data against a set of evenly distributed time stamps. P-VAE and P-BiGAN [16] incorperate a continuous convolutional layer into the encoder-decoder framework, and utilize the framework on partial Variational Autoencoder and partial Bidirectional GAN. Concare [28] uses the patient's demographic data and clinical features as input, and it learns the inter-dependencies among clinical features, and it improves the multi-head self-attention via the cross-head decorrelation.

Implementation Details. For diagnosis prediction task and mortality prediction task, we randomly split the datasets into the training, validation and testing sets based on the number of patients and the number of admissions in a 0.75:0.1:0.15 ratio respectively. For fair comparison, all models are implemented with Pytorch, and we use the Adam optimizer with learning rate 0.001 and weight decay 0.001. For all models of diagnosis prediction task, the dimension of the diagnosis code embedding and the corresponding hidden state size are set to 128, batch size is 100. For all models of mortality prediction task, the hidden state size is set to 32, and the batch size is set to 256.

VI. EXPERIMENTAL RESULT

In this section, we demonstrate the performance of diagnosis prediction and mortality prediction on two public EHR dataset MIMIC-III and MIMIC-IV respectively.

A. Diagnoses Prediction

Prediction Results. Table II shows the performance of **M**ulti-modal **C**ontrastive learning framework for **D**iagnosis **P**rediction task (MCDP) comparing with seven baselines. The proposed model MCDP outperforms all baselines and achieves 2% higher Recall@10 and MAP@10 over the best baseline. This demonstrates the effectiveness of our model. We can note that MCDP only uses two modalities, while the best baseline MDP uses three modalities. Among these baselines, RNN and Dipole [5] do not use the disease taxonomy information, and they directly learn the diagnosis code embedding from the input data.

GRAM [6], KAME [7] and HAP [8] all use the disease taxonomy information, and the inputs are the patient's diagnosis codes. The performance increases from GRAM to HAP. GRAM only updates the embedding of the leave nodes, and uses the diagnosis code representation to make predictions. KAME is built upon GRAM but learns a knowledge vector that contains the coarse-grained information of

TABLE II
DIAGNOSES PREDICTION RESULT

Methods		MIMIC_III						MIMIC_IV					
		Recall@K			MAP@K			Recall@K			MAP@K		
		K=10	k=20	k=30	k=10	k=20	k=30	k=10	k=20	k=30	k=10	k=20	k=30
Baselines	RNN	0.248	0.347	0.409	0.176	0.21	0.226	0.218	0.32	0.383	0.159	0.200	0.220
	DIPOLE	0.255	0.358	0.426	0.178	0.214	0.231	0.201	0.305	0.376	0.140	0.180	0.200
	GRAM	0.254	0.359	0.425	0.177	0.214	0.231	0.223	0.331	0.399	0.163	0.207	0.229
	KAME	0.254	0.357	0.427	0.177	0.213	0.231	0.220	0.329	0.403	0.159	0.203	0.227
	HAP	0.257	0.361	0.429	0.18	0.217	0.235	0.233	0.337	0.408	0.172	0.216	0.238
	CAMP	0.260	0.366	0.435	0.183	0.221	0.239	0.228	0.333	0.406	0.168	0.212	0.235
	MDP	0.268	0.375	0.446	0.190	0.229	0.248	0.232	0.341	0.410	0.173	0.219	0.241
Our Approaches	MCDP^{-cf}	0.278	0.389	0.458	0.204	0.245	0.263	0.253	0.358	0.431	0.195	0.245	0.270
	MCDP^{-c}	0.278	0.393	0.462	0.203	0.246	0.264	0.253	0.359	0.429	0.196	0.245	0.269
	MCDP	**0.283**	**0.396**	**0.464**	**0.208**	**0.247**	**0.266**	**0.258**	**0.361**	**0.433**	**0.198**	**0.248**	**0.271**

ancestor codes. KAME outperforms GRAM, which suggests that general knowledge of the ancestors helps to represent the patient's health condition and further boosts the performance of the diagnosis prediction. However, both GRAM and KAME ignore the order among the ancestors. HAP fills the gap by designing a two-round attention propagation mechanism, and the embedding of diagnosis code is updated layer by layer. The performance of HAP exceeds GRAM and KAME, which indicates that using the full ontology hierarchy improves the models' expressibility.

CAMP [10] and MDP [12] uses the multi-modal data as input, and the performance of CAMP and MDP exceeds HAP. CAMP incorporates the patient's demographic data, and MDP uses the clinical features and demographic data at the same time. MDP can adjust the weight of clinical features based on the patient's current health condition. MCDP uses the hyperbolic transformer to model the sequential EHR data, and uses multi-modal contrastive loss to find the relation between diagnosis modality and clinical feature modality.

Ablation Study. To further study the effectiveness of each module in MCDP, we compare three reduced models with some modules removed.

- MCDP^{-cf}: MCDP removing the embedding of clinical features and removing contrastive loss.
- MCDP^{-c}: MCDP removing contrastive loss.

Table II shows the prediction result. We can find that in both datasets, the performance of MCDP^{-cf} surpasses the best baseline MDP 2% for Recall@10. We should notice that MDP uses three modalities while MCDP^{-cf} only uses single modality. This means that hyperbolic transformer is superior at modeling sequential data comparing with LSTM or GRU, and embedding in hyperbolic space has greater expressivity than euclidean space for tree-like structures. Comparing with MCDP^{-cf}, MCDP^{-c} performs slightly better on MIMIC-III and the performance is almost the same on MIMIC-IV. By contrast, MDP incorporates clinical features based on CAMP, and achieves significant improvement. This shows that simply adding the clinical feature representation can not boost the performance without mechanisms to adjust the weight of clinical features based on patient's health condition, which is proposed in MDP. And we can also find that MCDP performs better than MCDP^{-c}, this demonstrates the effectiveness of the multi-modal contrastive loss.

B. Mortality Prediction

Prediction Results. Table III shows the performance of **M**ult-modal **C**ontrastive learning framework for **M**ortality **P**rediction task (MCMP) comparing with six baselines. The proposed model MCMP outperforms all baselines and achieves 5.6% higher AUC-ROC and 12% higher AUC-PRC over the best baseline for MIMIC-III. And we can observe similar result for MIMIC-IV. This demonstrates the effectiveness of the diagnoses code in the mortality prediction task.

Among these baselines, GRU-D [3] impute the missing data with mean value which decays with different time intervals, and Brits [13] impute the missing data with the linear transformation of hidden state of a bi-directional time-aware GRU, and it also computes the estimation loss. The result of Brits outperforms GRU-D, it shows that bi-directional GRU is superior to GRU. IPN [17] is an interpolation method, which applies an interpolation network to obtain a collection of interpolants of inpute data defined at the T evenly distributed reference time points. IPN achieves comparable performance to GRU-D, it shows that interpolation is also effective comparing with imputation. P-VAE and P-BiGAN [16] introduce the continuous convolutional layer into the encoder-decoder framework, and apply this framework to partial VAE and partial BiGAN. The performance of P-BiGAN is better than P-VAE. All the above baselines use clinical features as input, the last baseline Concare [28] incorporates demographic data, and it learns the inter-dependencies among clinical features. Meanwhile it improves the multi-head self-attention via the cross-head decorrelation.

The proposed model MCMP uses the clinical features and diagnoses code as inputs. MCMP concatenates the diagnoses code representation and the clinical feature representation learned via GRU-D to make predictions, and the multi-modal supervised contrastive loss are computed to further improve the performance. When computing the supervised contrastive

TABLE III
MORTALITY PREDICTION RESULT

Methods		MIMIC-III		MIMIC-IV	
		AUC-ROC	AUC-PRC	AUC-ROC	AUC-PRC
Baseline	Gru-d	0.847	0.397	0.847	0.410
	Brits	0.858	0.429	0.850	0.459
	IPN	0.850	0.409	0.847	0.413
	P-VAE	0.839	0.382	0.838	0.360
	P-BiGAN	0.851	0.412	0.845	0.422
	Concare	0.864	0.425	0.860	0.465
Our Approaches	MCMP^{-ac}	0.902	0.494	0.915	0.564
	MCMP^{-c}	0.916	0.520	0.912	0.571
	MCMP^{-a}	0.912	0.511	0.914	0.565
	MCMP	**0.920**	**0.543**	**0.915**	**0.573**

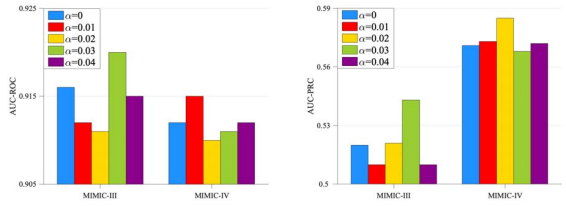
loss, the positives are defined as the samples having the same label with the anchor, and the other samples in the same batch are called the negatives. The result shows that incorporating diagnoses code can greatly boost the performance of mortality prediction, and it is grounded in the common sense. Usually, patients having diseases with high fatality rate may be more likely to die in an admission. The result of MCMP in Table III is computed with hyper-parameter α equals to 0.03 for MIMIC-III and 0.01 for MIMIC-IV.

Ablation Study. To further study the effectiveness of each module in MCMP, we compare three reduced models with some modules removed.

- MCMP^{-ac}: MCMP removing data augmentation for samples with true label and removing contrastive loss.
- MCMP^{-c}: MCMP removing contrastive loss.
- MCMP^{-a}: MCMP removing data augmentation for samples with true label.

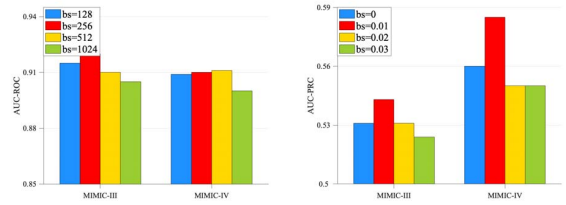
Table III shows the result on two datasets. Firstly, we can observe that MCMP achieves the best performance among all reduced models, and MCMP^{-ac} has the worst performance. Comparing with the backbone GRU-D, MCMP^{-ac} performs better because it concatenates the clinical feature representation and the diagnoses code representation, and uses the concatenated embedding to make prediction, this shows that diagnoses code representation is very useful for mortality prediction. MCMP^{-c} performs data augmentation for samples with true label in the training process. It perturbs the clinical features to augment the samples, and the augmented version has the same diagnoses codes and mortality label comparing with the original sample. Because MIMIC dataset is highly skewed, this method can be viewed as oversampling examples from the minority class. MCMP^{-c} performs better than MCMP^{-ac}, it shows that data augmentation can improve the performance. MCMP^{-a} computes the multi-modal contrastive loss, and it also performs better than MCMP^{-ac}. This demonstrates the effectiveness of multi-modal contrastive loss.

Detailed Analysis of Hyper-parameter. In this section, we conducted detailed analysis to show how prediction result changes with different hyper-parameters. Figure 2 shows the mortality prediction performance with respect to hyper-parameter α. α is used to adjust the weight of the multi-modal contrastive loss. We can observe that our method achieves the best performance when α equals to 0.03 for MIMIC-III and α equals to 0.01 for MIMIC-IV. Figure 3 shows the mortality prediction performance with respect to hyper-parameter batch size. We can observe that for both dataset, the performance achieves the best when batch size is 256.



(a) AUC-ROC (b) AUC-PRC

Fig. 2. Mortality Prediction Result V.S. Hyper-parameter α



(a) AUC-ROC (b) AUC-PRC

Fig. 3. Mortality Prediction Result V.S. Hyper-parameter batch size

VII. CONCLUSIONS

In this paper, we propose a multi-modal contrastive learning framework which uses the multi-modal contrastive loss to boost the prediction performance in healthcare. Our framework uses the diagnosis codes and clinical features as two related modalities. To maintain the hierarchy of diagnosis codes, we learn the Poincaré embeddings of diagnosis codes and we use hyperbolic transformer as the diagnosis code encoder. We compute the multi-modal contrastive loss between the

diagnosis modality and the clinical feature modality. And we perform two downstream tasks to show the effectiveness of the multi-modal contrastive learning framework. Meanwhile, we also define the multi-label contrastive loss and then extend it to the multi-modal version.

REFERENCES

[1] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6995–7004.

[2] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *Advances in Neural Information Processing Systems*, 2021.

[3] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[4] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[5] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1903–1911.

[6] 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.

[7] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "Kame: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 743–752.

[8] M. Zhang, C. R. King, M. Avidan, and Y. Chen, "Hierarchical attention propagation for healthcare representation learning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 249–256.

[9] C. Lu, C. K. Reddy, and Y. Ning, "Self-supervised graph learning with hyperbolic embedding for temporal health event prediction," *IEEE Transactions on Cybernetics*, 2021.

[10] J. Gao, X. Wang, Y. Wang, Z. Yang, J. Gao, J. Wang, W. Tang, and X. Xie, "Camp: Co-attention memory networks for diagnosis prediction in healthcare," in *2019 IEEE International Conference on Data Mining (ICDM)*, 2019, pp. 1036–1041.

[11] Z. Qiao, Z. Zhang, X. Wu, S. Ge, and W. Fan, "Mhm: Multi-modal clinical data based hierarchical multi-label diagnosis prediction," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1841–1844.

[12] R. Li, F. Ma, and J. Gao, "Integrating multimodal electronic health records for diagnosis prediction," in *AMIA Annual Symposium Proceedings*, 2021.

[13] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li, "Brits: Bidirectional recurrent imputation for time series," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[14] J. Yoon, J. Jordon, and M. Schaar, "Gain: Missing data imputation using generative adversarial nets," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5689–5698.

[15] Y. Luo, X. Cai, Y. ZHANG, J. Xu, and Y. xiaojie, "Multivariate time series imputation with generative adversarial networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[16] S. C.-X. Li and B. M. Marlin, "Learning from irregularly-sampled time series: A missing data perspective," in *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[17] S. N. Shukla and B. Marlin, "Interpolation-prediction networks for irregularly sampled time series," in *International Conference on Learning Representations*, 2019.

[18] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[20] 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.

[21] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[22] I. Chami, Z. Ying, C. Ré, and J. Leskovec, "Hyperbolic graph convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 32, pp. 4868–4879, 2019.

[23] R. Shimizu, Y. Mukuta, and T. Harada, "Hyperbolic neural networks++," in *International Conference on Learning Representations*, 2021.

[24] Q. Lu, N. de Silva, S. Kafle, J. Cao, D. Dou, T. H. Nguyen, P. Sen, B. Hailpern, B. Reinwald, and Y. Li, "Learning electronic health records through hyperbolic embedding of medical ontologies," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019, pp. 338–346.

[25] O. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[26] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 18 661–18 673.

[27] A. E. W. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard, "The mimic code repository: enabling reproducibility in critical care research," *Journal of the American Medical Informatics Association*, vol. 25, no. 1, pp. 32–39, 2018.

[28] 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, 2020, pp. 833–840.