Learning Robust Patient Representations from Multi-modal Electronic Health Records: A Supervised Deep Learning Approach

Xianli Zhang * Buyue Qian † Yang Li * Yang Liu * Xi Chen ‡ Chong Guan ‡ Chen Li †

Abstract

Predicting patients' future outcomes by analyzing Electronic health records (EHRs) is a hot topic in machine learning. The key challenge in this area is how to transform high dimensional, redundant, and heterogeneous EHRs into appropriate representations. In this paper, we argue for four desired properties of ideal patient representation learning, which are completeness, cross-modality invariance, anti-nuisance, and personality maintenance. To obtain such properties, We propose a Supervised Deep Patient Representation Learning Framework (SDPRL) to learn patients' representations that incorporates complete semantics of health conditions by using multi-modal EHR data. Furthermore, we propose to maximize the mutual information (MI) among each pair of different modal representations, as well as minimizing the task-specific loss function. This not only keeps the task-relevant semantic information into the learned representations, but also makes the resulting representations to be relatively invariant across the modalities, anti-nuisance, and maintain the personality. With experiments conducted on the publicly available MIMIC-III dataset on the mortality prediction and forecasting the length of stay (LOS) tasks, we empirically demonstrate that the proposed SDPRL achieves higher prediction performance than baseline frameworks. Moreover, we demonstrate that SDPRL can yield the desired properties we argued. It can well-handle the modal-missing issue in the test phase, as well as getting advance to the goal of personalized medicine.

1 Introduction

Learning better patient representation from Electronic Health Records (EHRs) is critical to improve the performance of the downstream clinical applications. Among various representation learning algorithms, Deep Representation Learning (DRL) has achieved successes in several domains with different data modalities, such as computer vision [1], speech and audio processing [2, 3], and natural language processing [4]. This attracts researchers in the healthcare domain to take advantage of DRL to learn patient representation from complicated EHR data. Based on the learned meaningful representations, various models can be readily constructed for different clinical tasks, such as measuring patient similarity [5], predicting in-hospital mortality [6], and forecasting length of stay (LOS) [7, 8], *etc.*

Although existing DRL-based models for clinical applications achieve successes to some extent, issues and challenges still exist. (i) Most existing approaches for learning patient representation are mostly based on single modal EHR data (*e.g.* only vital signs). However, from the perspective of the clinical diagnosis process, clinicians typically observe multi-aspect clinical observations (*e.g.* vital signs, notes, and interventions *etc.*) of the patient for having a comprehensive understanding of the patient's health conditions. In this sense, a model that lacks the capability of coding multi-modal input may lead to obtaining the representations without semantic integrity. (ii) Although several approaches [9, 10, 11, 12] were proposed for modeling multi-modal EHRs, they seem to lack the strength to handle the modal-missing issue in the test phase. This evidently undermines their applicability in real-world scenarios. The fundamental reason for this limitation is that these methods are feeble to learning modal-invariant representations. (iii) EHR data usually has high dimensions and contains many redundant features. What we might prefer is to keep the "good" factors into the representations and throw away the nuisance. Yet the most existing DRL-based models are not able to approach this goal well. Because either "good" or nuisance factors that have the correlations with outcomes will be involved in the learned representations in such models. (iv) One consensus is that personalized patient representations can drive healthcare decisions and strategies precisely [13, 14]. Nevertheless, patient representations derived from existing methods lack the person-

*National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China. {xlbryant, vigilee, ly97xjtu}@stu.xjtu.edu.cn

†School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China. {qianbuyue, cli}@xjtu.edu.cn

‡Tencent Jarvis Lab, Shenzhen, China. {jasonxchen, chongguan}@tencent.com

ality, which disobeying the goal of contemporary personalized medicine. The main reason is that they only optimize the Neural Network (NN) parameters by minimizing task-specific loss functions (*e.g.* cross entropy loss for classification), ignoring the instance-level (individual level) discrimination. Considering the above limitations, in this paper, we argue that the desired properties of a patient representation learning approach are as follows:

- Completeness: the representation should maintain the complete semantics of the patient's underlying healthcare condition.
- Cross-modality invariance: patient representations of different modalities within the same sample should as similar as possible.
- Anti-nuisance: redundant factors are supposed to be discarded whereas good ones should be kept in the representation.
- Personality: the representation should retain semantics of instance-level discrimination in extent, not only the task-specific labels differentiation.

To approach these goals, we propose a Supervised Deep Patient Representation Learning Framework (SDPRL) based on DRL. We adopt multi-modal EHR data and try to represent the patient's health status as comprehensive as possible. Each modal of EHRs is regarded as a separate, redundant, and incomplete observation perspective to the patient's health conditions. For each modal, we use a modal-specific neural network (NN) for encoding modal-specific information into the representation space. Then, we apply global Max-pooling to fuse all modal representations to a unified fixed-size vector as the final patient representation. Because such pooling opration is flexible to handle the varying number of the input modalities in the test phase.

Based on the above simple pipeline, SDPRL maximizes the mutual information (MI) between the different modal representations in a modal-pair at the instance-level while minimizing the task-specific loss function. This endows several advantages to the obtained representation. First, modal-specific encoders are guided by each other and represent the information of each other modal as far as possible, thus make the representation of each modal retain relatively complete information of the inputs. Beyond this, different modalities of the same patient can be mapped to nearby points in the representation space, which makes the overall framework robust towards the modal missing issue in the test phase. Second, "good" factors which shared task-relevant semantic and have dependencies between modalities can be kept in the representation, while the independent noise can be thrown away. Finally, SDPRL can yield personality properties by emphasizing the consistency of different modal representations at the individual level. With experiments conducted on the publicly avaliable MIMIC-III [15] dataset. we demonstrate that the proposed SDPRL can yield all aforementioned desired properties of patient representation. The main contributions of this paper are as follows:

- We propose SDPRL, a simple yet effective DRL-based patient representation learning framework for encoding patients' multi-modal EHRs. The representation derived from SDPRL has several advantages, including completeness, cross-modality invariance, anti-nuisance, and personality maintenance.
- We consider an extremely hard setting, which assumes that only the "weaker modal" [1] can be obtained in the test phase. We empirically show that maximizing the MI among each pair of different modal representations can maintain the performance even with the "weaker modal". This makes SDPRL cope well with the modal missing problem.
- By applying linear encoder to SDPRL, we visualize the activation of all input factors. We observe that "good" factors with all modalities that shared task-relevant semantic can be kept in the representation, whereas the nuisance factors, *i.e.* independent to both task-specific label and others modal factors, are discarded.
- Finally, we adopt t-SNE to visualize the obtained patient representation in the test set. We find that by emphasizing the consistency of different modal representations, the fused representation clustering into multiple clusters and each of which with a clear boundary between positive and negative patients.

2 Related Work

EHRs contain multi-modal information related to patient health conditions and behaviors in hospitals, *e.g.* demographics, clinical narratives, lab measurements, examination reports, and other patient encounter entries [16]. Appropriately represent multi-modal EHRs is a fundatial step before conducting downstream clinical tasks, such as measuring patient similarity [17, 18], mortality prediction [19, 20], intervention prediction [9, 21, 22], diagnosis prediction [23, 24, 25, 26], disease

[1] In this paper, we call the "weaker modal" as the modal with the lower prediction performance when training the model separately with the data of each modal.

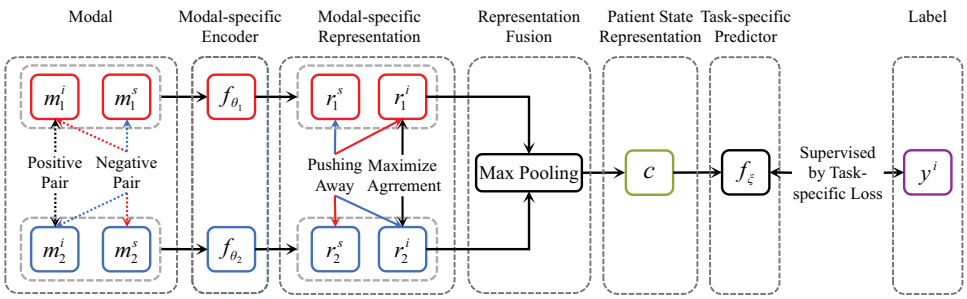


Figure 1: The overview pipeline of the proposed SDPRL in the two modalities case. We treat the two different modalities of the i-th patient as a positive pair (*i.e.* $\{m_1^i, m_2^i\}$, indicated by the black dashed line with arrows), whereas one modal of other patients with a different label from the i-th patient are chosen to form a negative pair with one modal of i-th patient (*i.e.* $\{m_1^i, m_2^s\}$ and $\{m_2^i, m_1^s\}$, indicated by the red and blue dashed line with arrows, respectively). The core idea of SDPRL is to maximize the consistency between the representations of the positive pair and separate that of negative pair away (as shown in the Modal-specific Representation block).

.

risk prediction [27, 28, 29], and forecasting LOS [6, 7, 8], *etc.* However, the inherent issues of EHRs, such as noisy, high-dimensionality, sparsity, and heterogenous, create obstacles to conventional machine learning approaches to learn robust and effective patient representations. In contrast, DRL shares the superior ability for encoding complex data into dense and distributed vectors in a continuous space. To learn meaningful patient representations by DRL, researchers usually leverage task-specific labels as the supervised information for optimization. Various NN architectures for gaining higher prediction performance have also been studied in literatures [16, 27, 24, 30].

However, most existing approaches only used single modal EHR data for learning representation of the patient [6, 7, 19, 20]. This would damage the completeness of the obtained semantic representations. To keep relative complete semantics of the patient in the representations, several pieces of literature have explored learning patient representations from multi-modal EHRs. For example, Suresh *et al.* [9] proposed to learn representations of multi-modal EHR data for predicting the onset and weaning of ICU interventions. Qian *et al.* [10] considered encoding information of both ICD codes and textual clinical notes into the representations, thus lead to an improvement of understanding the patient health conditions. Purushotham *et al.* [11] proposed a multimodal deep learning model namely MMDL to learn shared representations from multiple types of EHR data in ICU. Yin *et al.* [12] propose to incorporate multimodal clinical inputs (*e.g.* demographics, diagnosis, medications, and lab measurements) to impute the missing values in lab measurements. They showed the imputed data yield better performance than original data in a clinical clustering task. Although these approaches can learn representations with relative complete semantics towards patients' health conditions, they are powerless to the missing model issue. This problem is caused by both the strongly coupled feature fusion manner and incapable of learning cross-modal invariant representation. Besides, the absence of the extra supervision on capturing complex relationships and dependencies across disparate data types can not further throw away nuisance factors of inputs. Last, the personality of patient is also ignored by existing models, which should be an important feature in the area of precision and personalized medicine.

3 Method

In this section, we first introduce our notation, as well as the multi-modal sampling strategy for each patient. Then, we elaborate on our SDPRL framework mathematically in the scenarios of two modalities, under the classification setting. Finally, we present how to extend SDPRL to the case of multiple modalities.

3.1 Notation & Multi-modal Sampling Strategy for Each Patient. In this paper, an EHR dataset that consists of M modalities with N patients is denoted as $D = \{m_1^i, \cdots, m_M^i, y^i\}_{i=1}^N$, where m indicates the input with arbitrary modal and y represents the label.

We call any two modal data from the same patient as a positive pair, which is denoted by $p^i = \{m_j^i, m_k^i\}$, where $1 \leq j < k \leq M$. For each positive pair p^i, we denote its corresponding negative pairs with respect to j-th modal as $\{m_j^i, m_k^s\}$, where $i \neq s$. Similarly, a negative pair with respect to k-th modal is denoted by $\{m_j^s, m_k^i\}$, where $i \neq s$. Besides, we force the labels

for each modalities in a negative pair are different, *i.e.* $y^i \neq y^s$ for classification, and $|y^i - y^s| > \epsilon$ for regression, where ϵ is a hyperparameter. The positive pair and the negative pair will be used for training the consistency of different modalities representation within the same patient.

3.2 SDPRL on Two Modalities Fig. 1 illustrates an example of SDPRL applied to a dataset with two-modal (*e.g.* vital signs and interventions). The core idea of this paper is to maximize the representation consistency among each modal pair within the same patient while keeping the task-relevant semantics in the representations. To achieve this goal, SDPRL first encode each modal into the continuous space by the corresponding modal-specific encoder. Then, it maximizes the MI among each positive pair, as well as minimizing the task-specific loss function.

As shown in Fig. 1, given a patient's two-modal inputs $\{m_1^i, m_2^i\}$, SDPRL regard them as a positive pair p^i, and encode them into the representation space, respectively:

$$r_1^i = f_{\theta_1}(m_1^i); \quad r_2^i = f_{\theta_2}(m_2^i), \tag{3.1}$$

where $r_1^i, r_2^i \in \mathbb{R}^d$ are the representations of the corresponding modalities, $f_{\theta_1}, f_{\theta_2}$ are the modal-specific encoder, which can be implemented by proper NN architectures.

Subsequently, we treat each modal as an anchor to sample the corresponding negative pairs that will be used as the negative samples for training the consistency of r_1^i and r_2^i. Taking m_1^i as an example, we randomly sample N_{neg} negative pairs according to Section 3.1. Then, each negative pair $\{m_1^i, m_2^s\}$ is encoded by the corresponding modal-specific encoder according to Eq. (3.1), where we denote the resulting representations of a negative pair as $\{r_1^i, r_2^s\}$. Afterward, we attempt to emphasize the consistency between representations of the positive pair while reducing that of negative pairs. We propose to maximize the lower bound of MI among a positive pair. The objective function can be formalized as:

$$\mathcal{L}_{m_1^i, m_2^i} = -\log \frac{\exp(\text{sim}(\{r_1^i, r_2^i\})/\tau)}{\sum_{n=1}^{N_{neg}} \exp(\text{sim}(\{r_1^i, r_2^s\}_n)/\tau)}, \tag{3.2}$$

where τ is a temperature parameter, and $\text{sim}(\cdot,\cdot)$ indicates the cosine similarity.

Note that $\mathcal{L}_{m_1^i, m_2^i}$ in Eq. (3.2) treats the 1-th modal as an anchor, and sample negative pairs from the 2-th modal. Symmertrically, when anchoring at the 2-th modal, we can obtain $\mathcal{L}_{m_2^i, m_1^i}$ as:

$$\mathcal{L}_{m_2^i, m_1^i} = -\log \frac{\exp(\text{sim}(\{r_1^i, r_2^i\})/\tau)}{\sum_{n=1}^{N_{neg}} \exp(\text{sim}(\{r_1^s, r_2^i\}_n)/\tau)}. \tag{3.3}$$

In order to handle the modal-missing issue in the test phase, we adopt global Max-pooling opertation to fuse r_1^i and r_2^i to an unified vector:

$$c^i = \text{Max-pooling}(r_1^i, r_2^i), \tag{3.4}$$

where $c^i \in \mathbb{R}^d$ is the final patient representation.

Then, we minimize the corresponding task-specific loss function for further guiding the overall framework keep the task-relevant semantics in c^i, which can be formalized as follow:

$$\mathcal{L}_{task} = \begin{cases} \frac{1}{N_b} \sum_{i=1}^{N_b} \text{CE}(y_i, f_\xi(c^i)), & \text{for classification} \\ \frac{1}{N_b} \sum_{i=1}^{N_b} \text{SE}(y_i, f_\xi(c^i)), & \text{for regression} \end{cases} \tag{3.5}$$

where N_b indicates the batch size, $\text{CE}(\cdot,\cdot)$ indicates the cross-entropy loss, $\text{SE}(\cdot,\cdot)$ denotes the square error, and f_ξ is the classifier or regressor that can be implemented depending on downstream tasks.

In practice, we can optimize the overall framework in an end-to-end manner by summing all objective functions together:

$$\begin{aligned} \mathcal{L} &= \lambda_1 \sum_{i=1}^{N_b} \mathcal{L}_{m_1^i, m_2^i} + \lambda_2 \sum_{i=1}^{N_b} \mathcal{L}_{m_2^i, m_1^i} + \lambda_c \mathcal{L}_{task} \\ &= \sum_{i=1}^{N_b} (\lambda_1 \mathcal{L}_{m_1^i, m_2^i} + \lambda_2 \mathcal{L}_{m_2^i, m_1^i}) + \lambda_c \mathcal{L}_{task}, \end{aligned} \tag{3.6}$$

where $\lambda_1, \lambda_2, \lambda_c$ are three balancing hyperparameters.

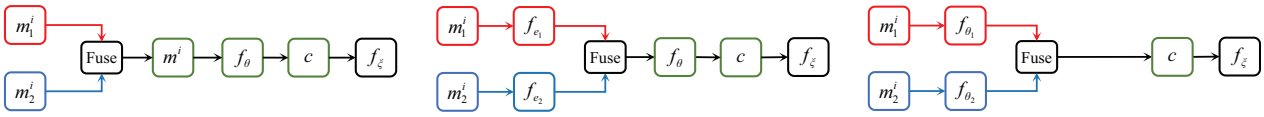
3.3 SDPRL with More than Two Modalities In this section, we extend SDPRL to the case of more than two modalities. Consider there are M modalities in the dataset (*i.e.* $D = \{m_1^i, \cdots, m_M^i, y^i\}_{i=1}^N$), the goal of SDPRL is to emphasize the consistency among representations of all modalities within individual patient. To this end, we can incorporate all modalities into Eq. (3.6), which can be mathematically described as:

$$\begin{aligned} \mathcal{L}_M = &\sum_{1 \leq j < k \leq M} \sum_{i=1}^{N_b} (\lambda_j \mathcal{L}_{m_j^i, m_k^i} + \lambda_k \mathcal{L}_{m_k^i, m_j^i}) \\ &+ \lambda_c \mathcal{L}_{task}, \end{aligned} \tag{3.7}$$

where $\lambda_j, \lambda_k, \lambda_c$ are balancing hyperparameters.

4 Experiments & Evaluations

4.1 Dataset & Experiment Settings MIMIC-III is a widely-used publicly data source that has catalyzed many research studies. In order to enhance the reproduction of our experiments, in this paper, we



(a) EFI (b) EFE (c) LF

Figure 2: Framework schemas of EFI, EFE, and LF in the two-modal case.

adopt an open source pipeline namely MIMIC-Extract [20] to construct dataset from the original MIMIC-III database. Using MIMIC-Extract, we extract the "level2" cohort that including demographics, vital signs, and interventions of each patient. The dataset statistics can be find in both the MIMIC-Extract paper [20] and the its code repositories [2]. For each patient, we inject her/his demographic into vital signs and interventions separately at each timestamp. We treat the vital signs and interventions (both injected with demographics) as two modalities. According to MIMIC-Extract, we build two classification tasks, which are in-ICU mortality prediction and LOS > 7 days prediction.

We randomly split the constructed dataset into training, validation, and testing sets in a 70:10:20 ratio. Besides, to explore whether the proposed SDPRL can handle well with the modal-missing issue, we consider an extremely hard setting, which assuming that only the "weaker modal" [1] can be obtained in the test phase.

Note that the positive/negative labels in the dataset for both tasks are usually imbalanced. Therefore, we use Area Under Receiver Operator Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC) to measure the performance of all approaches.

4.2 Baselines In order to fairly evaluate the effectiveness of the proposed SDPRL, we compare it with several popular NN-based frameworks for multi-modal EHR data, including early fusion on inputs (**EFI**), early fusion on embeddings (**EFE**) [12], and late fusion (**LF**) [9, 10, 11]. It is worth noting that these frameworks in the corresponding literatures are designed with specific NN architectures for different clinical tasks. In this paper, we treat them as abstract frameworks and extend them to two appropriate NNs, which are Recurrent NN, and Transformer [31]. The framework schemas of EFI, EFE, and LF are illustrated in Fig. 2, and the detailed descriptions of them are given as follows:

EFI: In this framework, we concatenate all modal data together to obtain a unified input vector. However, this is not usually a flexible approach, because in most cases, multi-modal data is difficult or impossible to be fused in the input-level (e.g. clinical notes and medical image).

EFE: This framework usually first transform each modal data to an embeddings vector at each timestamp, and then concatenate all the embeddings of each timestamp to a unified feature vector.

LF: This is the most popular framework for encoding multi-modal data to a unified vector. In this baseline, each modal data is encoded by the model-specific encoder into modal-specific representation space. Then, all modal representations are aggregated together by a concatenate operation followed with an MLP layer.

It worth noting that our SDPRL is similar to the schema of LF. The main difference between SDPRL and LF can be divided into two aspects. First, SDPRL adopts an global Max-pooling opertation to aggregate all representations of different modalities, whereas LF uses an MLP layer. Second, SDPRL emphasizes the consistency between each pair of modalities representations. Note that the use of MLP is a hard-coupled manner for feature fusion, which cannot cope with the modal-missing issue. Hence, to further evaluate the effectiveness of SDPRL, we replace the MLP layer in LF by the global Max-pooling operation, whereby the resulting framework denoted by $\mathbf{LF}_{pool}$ can be treated as an variants of both LF and SDPRL.

4.3 Implementation Details For each modal input, we use an MLP layer with ReLu activation as the embedding layer, the output dimension of which is set to 128. For all frameworks, we extend the encoder (i.e. f_θ) of them to two appropriate NN architectures, which are Bi-directional Long Short Memory Neural Network (Bi-LSTM) and the encoder part of Transformer with 2 heads and 2 layers. We set the output dimensions of all NNs at each timestamp to 128. The predictors (i.e. f_ξ) of all frameworks are implemented by 2 fully-connection layers with an inner dimension size of 256. For SDPRL, we set $\tau = 0.1, N_{neg} = 32, \lambda_1 = \lambda 2 = 0.05,$ and $\lambda_c = 1$. For training all models, we use Adam [32] with the batch size of 16 and the learning rate of 0.0001. The weight decay is set to 0.001. All frameworks are implemented with PyTorch 1.0 on two Nvidia Titan XP GPUs.

4.4 Results and Discussions

[2] https://github.com/MLforHealth/MIMIC_Extract

Table 1: Performance Results on In-ICU Mortality, and LOS > 7 Days.

Task	Model		AUROC (%)	AUPRC (%)
In-ICU Mortality	Bi-LSTM	EFI	**90.30**	52.54
		EFE	89.89	52.65
		LF	89.66	51.67
		LF_{pool}	89.49	51.82
		SDPRL	90.24	**52.77**
In-ICU Mortality	Transformer	EFI	89.53	52.10
		EFE	89.58	52.31
		LF	89.93	52.61
		LF_{pool}	89.81	52.68
		SDPRL	**90.60**	**53.66**
LOS > 7 Days	Bi-LSTM	EFI	76.60	19.23
		EFE	76.57	19.05
		LF	76.14	19.22
		LF_{pool}	76.33	**19.44**
		SDPRL	**76.64**	19.23
LOS > 7 Days	Transformer	EFI	76.44	19.28
		EFE	76.76	19.36
		LF	76.16	19.17
		LF_{pool}	76.04	19.01
		SDPRL	**77.91**	**19.96**

4.4.1 Performance Analysis: Table 1 reports the AUROC and AUPRC of all approaches on both the In-ICU mortality and LOS > 7 days tasks. Specifically, in Transformer-based implemented frameworks, the AUROC improves 0.75% and the AUPRC improves 1.86% when comparing the results of SDPRL with the best results of baselines frameworks on the In-ICU mortality task. On the LOS > 7 days task under the Transformer-based implementation, we can see that SDPRL achieves an improvement of 1.50% on AUROC and an improvement of 3.10% on AUPRC, that compared with the best results of all baseline frameworks. In Bi-LSTM-based implementation, although EFI has a little bit higher AUPRC (only 0.07%) than ours on the In-ICU mortality and LF_{pool} achieves a little bit higher AUPRC (1.09%) than ours, we can observe from Table 1 that, on average, the proposed SDPRL achieves almost the best performance compared with all baselines. The reason is that Eq. (3.2) and Eq. (3.3) provide additional supervision for mining the complex relationships among factors of different modalities. This can not only produce a complete representation of the patient but also can capture more complex dependencies among all modalities.

4.4.2 Handle With the Modal-missing Issue In order to show the capability of SDPRL in handling the modal-missing issue, we first report the performance of the Bi-LSTM and the 2-head 2-layer Transformer encoder that both training and testing on single modal data. See the top four raws in Table 2, we find that the interventions is the "weaker modal" in the dataset. Then, we train LF_{pool} and SDPRL with different implementation on the two-modal dataset, and test their performance on single-modal testing set. Note that we do not report the performances of EFI, EFE, and LF here, because their hard-coupled feature fusion manner makes them can not handle the modal-missing issue. As we can see from the last eight raws in table 2, SDPRL outperforms LF_{pool} with a significantly margin on both AUROC and AUPRC. Specifically, when only the "weaker modal" (*i.e.* interventions) is avaliable in the testing set, the AUROC and AUPRC improves 8.48% and 11.62%, respectively, when comparing SDPRL with LF_{pool} in the implementation of Bi-LSTM. In the Transformer-based implementation, SDPRL achieves improvements of 13.70% and 35.60% on AUROC and AUPRC, respectively. Besides, even testing them on the testing set that only contains vital signs, our framework also achieves higher AUROC and AUPRC compared with that of LF_{pool}.

The reasons for the above observations is that optimizing Eq. (3.2) and Eq. (3.3) yields relative invari-

Table 2: Performance results of SDPRL and LF_{pool} with different training sets and testing sets.

Methods	Training Set	Testing Set	AUROC(%)	AUPRC(%)
Bi-LSTM	Vital Signs	Vital Signs	89.38	50.44
Bi-LSTM	Interventions	Interventions	79.42	33.79
Transformer	Vital Signs	Vital Signs	89.52	51.10
Transformer	Interventions	Interventions	78.85	32.52
Bi-LSTM+LF_{pool}	Vital Signs & Interventions	Vital Signs	88.98	50.27
Bi-LSTM+SDPRL	Vital Signs & Interventions	Vital Signs	**89.67**	**50.58**
Bi-LSTM+LF_{pool}	Vital Signs & Interventions	Interventions	73.48	28.57
Bi-LSTM+SDPRL	Vital Signs & Interventions	Interventions	**79.71**	**31.89**
Transformer+LF_{pool}	Vital Signs & Interventions	Vital Signs	89.58	51.10
Transformer+SDPRL	Vital Signs & Interventions	Vital Signs	**90.33**	**51.47**
Transformer+LF_{pool}	Vital Signs & Interventions	Interventions	68.80	22.22
Transformer+SDPRL	Vital Signs & Interventions	Interventions	**78.23**	**30.13**

Figure 3: Heatmaps of feature activations in the vital signs, where the top raw is computed by parameters of SDPRL with linear NN implementation, and the bottom raw is computed by that of LF_{pool} with the same implementation. The AUROC and AUPRC of SDPRL are 89.57% and 51.35%, respectively, whereas AUROC and AUPRC of LF_{pool} are 89.01% and 50.09%, respectively. We can clearly observe that SDPRL relies on fewer factors but achieve higher performance than LF_{pool}.

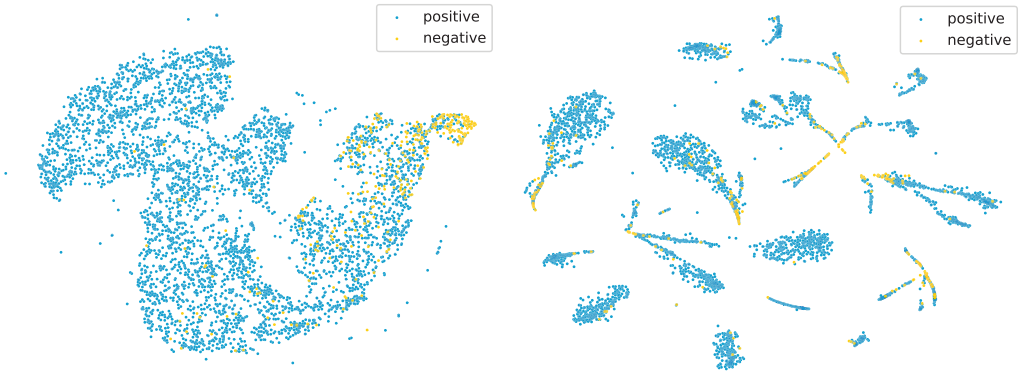
ant representations across modalities. In other words, model-specific encoders can be guided by each other, so modalities of the same patient can be mapped to relative nearby points in the representation space. Therefore, when one modal data miss in the test phase, the representation of another modal is still close to the semantic representation of the patient in the vector space.

4.4.3 Capability of Throwing Nuisance Away of Noisy and High Dimentional EHRs: In order to show the capability of SDPRL in discarding noisy factors in inputs, we implement f_θ of both SDPRL and LF_{pool} by two fully-connection layers followed with a global Average-pooling. Then, we training these two implemented frameworks on the two-modal dataset, and compute the activation of each trained model towards each factor of the vital signs, which contain over 300 factors. Fig. 3 shows the activations of all factors of both LF_{pool} and SDPRL. We can observe that SDPRL yields a more sparse factor activation heatmap than LF_{pool}. But the performance of SDPRL is higher than that of LF_{pool}. This demonstrates that SDPRL can keep less yet "good" factors into the representations while LF_{pool} may rely on some noisy factors that harm to performance. The reason for this is that by optimizing Eq. (3.6), only the factors that shared task-relevant semantics and have dependencies between modalities can be kept in the representation, while independent factors will be thrown away.

4.4.4 Personality Analysis: In order to show the personality property of the leaned representation of SDPRL, we visualize the patient representations of both LF_{pool} and SDPRL in Fig. 4, where the dimension is reduced to two by t-SNE. We can observe that LF_{pool} only learned a boundary between positive and negative patients, whereas SDPRL yields several clusters. As we can see from Fig. 4 (b), some clusters only contain negative patients, and some clusters contain both negative and positive patients. In each cluster that contains both negative and positive patients, there is a clear boundary between negative and positive patients. This demonstrates that SDPRL can not only learn the task-relevant semantic information but also can retain personalized information, which accords with the goal of personalized and precision medicine.

5 Conclusion

In this work, a novel patient representation learning framework was proposed, namely SDPRL. We experimentally proved that the proposed SDPRL can achieve better prediction performance than all baseline frameworks. Moreover, we showed that the learned representations of SDPRL contains several desired properties,



(a) LF_{pool} (b) SDPRL

Figure 4: Patient representations visualization learned by LF_{pool} and SDPRL for the test set. Dimension reduced via t-SNE. We can observe that LF_{pool} only learned a boundary between positive and negative patients, whereas SDPRL yields several clusters. This shows the personality property of the representations learned by SDPRL.

including cross-modality invariance, anti-nuisance, and personality maintenance. The learned cross-modality invariant representations can handle the modal-missing issues better than all baseline frameworks. This demonstrates that our SDPRL can be well-adapted to complex clinical prediction scenarios. When implementing SDPRL by a linear model and visualizing the activation of input factors, we proved that maximizing the MI between representations within each modality pair can get rid of more nuisance infromation. Last but not least, we visualized the learned representations by t-SNE to show the personality property of the representations. This may be a potential solution to getting more closer to the goal of personalized and precision medicine.

6 Acknowledgments

This work has been supported by the National Key Research and Development Program of China (2018YFC0910404); National Natural Science Foundation of China (61772409); The consulting research project of the Chinese Academy of Engineering (The Online and Offline Mixed Educational Service System for "The Belt and Road" Training in MOOC China); Project of China Knowledge Centre for Engineering Science and Technology; The innovation team from the Ministry of Education (IRT 17R86); and the Innovative Research Group of the National Natural Science Foundation of China (61721002).

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Unsupervised cross-modal alignment of speech and text embedding spaces. In *Advances in Neural Information Processing Systems*, pages 7354–7364, 2018.

[3] Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *Interspeech 2016*, pages 765–769, 2016.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[5] Zihao Zhu, changchang Yin, Buyue Qian, Yu Cheng, Jishang Wei, and Fei Wang. Measuring patient similarities via a deep architecture with medical concept embedding. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 749–758, 2016.

[6] Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: clinical time series analysis using attention models. In *Proc. AAAI Conf. Artificial Intelligence*, pages 4091–4098, 2018.

[7] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

[8] Emma Rocheteau, Pietro Liò, and Stephanie Hyland. Predicting length of stay in the intensive care unit with temporal pointwise convolutional networks. *arXiv preprint arXiv:2006.16109*, 2020.

[9] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for*

Healthcare Conference, pages 322–337, 2017.

[10] Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. MNN: multimodal sttentional neural networks for diagnosis prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5937–5943, 2019.

[11] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

[12] Changchang Yin, Ruoqi Liu, Dongdong Zhang, and Ping Zhang. Identifying sepsis subphenotypes via time-aware multi-modal auto-encoder. In *Proc. ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, KDD '20, pages 862–872, 2020.

[13] Jinghe Zhang, Kamran Kowsari, James H Harrison, Jennifer M Lobo, and Laura E Barnes. Patient2vec: a personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access*, 6:65333–65346, 2018.

[14] Qiuling Suo, Fenglong Ma, Ye Yuan, Mengdi Huai, Weida Zhong, Jing Gao, and Aidong Zhang. Deep patient similarity learning for personalized healthcare. *IEEE transactions on nanobioscience*, 17(3):219–227, 2018.

[15] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[16] Wei-Hung Weng and Peter Szolovits. Representation learning for electronic health records. *arXiv preprint arXiv:1909.09248*, 2019.

[17] Qiuling Suo, Weida Zhong, Fenglong Ma, Yuan Ye, Mengdi Huai, and Aidong Zhang. Multi-task sparse metric learning for monitoring patient similarity progression. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 477–486. IEEE, 2018.

[18] Jiazhi Ni, Jie Liu, Chenxin Zhang, Dan Ye, and Zhirou Ma. Fine-grained patient similarity measuring using deep metric learning. In *Proc. ACM Int'l Conf. Information and Knowledge Management*, pages 1189–1198, 2017.

[19] Aya Awad, Mohamed Bader-El-Den, James McNicholas, Jim Briggs, and Yasser El-Sonbaty. Predicting hospital mortality for intensive care unit patients: time-series analysis. *Health informatics journal*, 26(2):1043–1059, 2020.

[20] Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 222–235, 2020.

[21] Zaharah Allah Bukhsh, Irina Stipanovic, Aaqib Saeed, and Andre G Doree. Maintenance intervention predictions using entity-embedding neural networks. *Automation in Construction*, 116:103202, 2020.

[22] Marzyeh Ghassemi, Mike Wu, Michael C Hughes, Peter Szolovits, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. *AMIA Summits on Translational Science Proceedings*, 2017:82, 2017.

[23] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. GRAM: graph-based attention model for healthcare representation learning. In *Proc. ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 787–795, 2017.

[24] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proc. ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 1903–1911, 2017.

[25] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proc. ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 743–752, 2018.

[26] Yang Li, Buyue Qian, Xianli Zhang, and Hui Liu. Knowledge guided diagnosis prediction via graph spatial-temporal network. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 19–27. SIAM, 2020.

[27] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

[28] Fenglong Ma, Yaqing Wang, Jing Gao, Houping Xiao, and Jing Zhou. Rare disease prediction by generating quality-assured electronic health records. In *Proc. SIAM Int'l Conf. Data Mining*, pages 514–522. SIAM, 2020.

[29] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proc. ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 647–656, 2020.

[30] 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.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.