2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669798

# Domain Adaptation for Trauma Mortality Prediction in EHRs with Feature Disparity

Xinlu Zhang[†], Shiyang Li[†], Zhuowei Cheng[†], Rachael Callcut[¶], Linda Petzold [†] ,
University of California, Santa Barbara[†]
University of California, Davis[¶]
{xinluzhang, shiyangli,zwcheng,petzold}@cs.ucsb.edu racallcut@ucdavis.edu

***Abstract*—Trauma mortality prediction from electronic health records (EHRs) with machine learning models has received growing attention in medical fields, but EHRs in different hospitals and sub-medical domain populations are often scarce due to expensive collection processes or privacy issues. Domain Adaptation (DA) has emerged as a promising approach in computer vision and natural language processing to improve model performance in small data regimes by leveraging domain-invariant knowledge learned from a different yet related large source dataset. However, its applicability in trauma mortality prediction is challenging since EHRs collected from different hospital systems encounter feature disparity, i.e. distinct features between the source and target domain data. This paper demonstrates the effectiveness of three DA techniques in trauma mortality prediction, with a private encoding strategy that maps EHRs in both source and target domains with different raw features into the same latent space to alleviate feature disparity issues. Our experimental results on two real-world EHR datasets with various training data scenarios show that DA can improve mortality prediction consistently and significantly with private encoding. Finally, an ablation study manifests the importance of modeling feature disparity in DA, and 2-d t-SNE analysis explains its effectiveness.**

***Index Terms*—Electronic Health Record, Mortality Prediction, Domain Adaptation, Adversarial Learning, Contrastive Learning**

## I. Introduction

Trauma is the leading cause of death from age 15 to 49 worldwide, resulting in the death of more than 5 million people each year [1]. After admission to intensive care units (ICUs), most of these deaths occur in the first several hours or days. Treatment decisions and actions in the first several minutes or hours after injury are critical [2], and medical decision errors are more likely to be made during this period than at later times [3]. Thus, tools that can provide efficient and real-time prediction are critical for clinicians to maximize the impact of treatment and improve survival rates.

Machine learning approaches are increasingly being used to detect adverse events in clinical settings. Different from risk scores, e.g. Modified Early Warning Score (MEWS) [4], which are developed on general populations with specific clinical measurements, machine learning techniques can be customized to different patient subpopulations or professional care facilities by training models on different datasets [5]. Recently, with the rapid development of deep learning (DL), a variety of DL techniques and frameworks have been applied to clinical predictions, such as in-hospital mortality prediction, length-of-stay prediction and phenotype classification [6] [7], achieving state-of-the-art performance. DL models often require a large corpus of labelled training data to work well [8]. Therefore, researchers often use DL approaches and draw conclusions on the basis of large public electronic health record (EHR) datasets, such as Medical Information Mart for Intensive Care (MIMIC III) [9], assuming that sufficient training data is available for different clinical tasks and the test set follows the same distribution of training data. Nevertheless, models trained on large public datasets often achieve suboptimal performance when directly deployed to smaller private EHRs due to distribution shift [10], e.g. differences in lab procedures and instrumentation, injury types and population groups based on location etc. On the other hand, only limited private EHRs are available to serve as the training set at a single medical system [5], especially when further narrowed into a medical sub-domain, e.g. trauma, limiting possible application of DL methods.

Domain adaptation (DA) is a subcategory of transfer learning that leverages knowledge from a different but related source domain as additional information to improve model performance for a target domain with limited training data [11]. DA has made remarkable progress in computer vision [12] and natural language processing [13]. Some prior works [10][14][15] have also successfully applied DA to clinical predictions across multiple hospital systems, assuming that the distribution shift between the source and target domain of EHRs is caused by heterogeneous patient populations (*covariate shift*) and variations in data collection procedures (*systematic bias*). However, these methods are used only on datasets with overlapping features in both domains, and ignore information provided by distinct features in the target and source (*feature disparity*), which describe the characteristics of different datasets and can be essential for prediction performance. For example, clinicians tend to order particular blood tests for trauma patients, compared to other ICU patients, to identify disease states associated with coagulopathy [16], a known contributor to trauma mortality [17]. Two challenges remain to directly apply DA to multi-hospital system EHRs with feature disparity in a target data scarcity setting. First, not all clinical features included in the source and target datasets are precisely the same, and even some overlapping features collected from different hospital systems are represented in different ways. Second, it is impracticable to align the target and source distributions when the target training set is extremely inadequate.

In this paper we aim to overcome the aforementioned challenges to fully utilize DA techniques to improve DL performance on a small target dataset, by leveraging domain-invariant knowledge from another different yet related large source dataset with feature disparity. Specifically, we introduce the private encoding technique to map target and source datasets from different feature spaces into the same hidden representation space, and utilize pairing sampling techniques [18],[19] to pair each target data point to abundant source data points, to effectively align the target and source distributions for applying different DA approaches. We demonstrate the effectiveness of DA in mortality prediction for trauma patients by taking feature disparity into account on two real-world datasets. To summarize, our contributions include:

- We extend DA techniques with a proposed private encoding to enable early-stage mortality prediction for trauma patients in multi-hospital system EHRs with feature disparity. To the best of our knowledge, this is the first paper to consider feature disparity between the source and target domain in leveraging DA methods.
- Experimental results on two small target datasets show that DA techniques with our proposed methodology improve mortality prediction performance consistently and significantly in various training scenarios, with F1-scores up to $100\%$.
- We provide an ablation study and a 2d t-SNE [20] analysis on two datasets, which further underscores the importance of modeling feature disparity and validates the effectiveness of DA techniques, respectively.

## II. Related Work

Machine learning models have been trained for many clinical tasks, with several public datasets available, e.g. MIMIC III and eICU Collaborative Research Database (eICU-CRD)[21]. [22] presents benchmarking results for clinical prediction tasks such as mortality prediction, length of stay prediction, and ICD-9 code group prediction using the MIMIC III dataset. [23] performs a length of stay prediction, utilizing a Bayesian neural network, and conducts experiments on the eICU-CRD datasets. Although DL approaches attain state-of-the-art performance in medical domains, most of the works are only applied to large EHRs but difficult to yield similar performance when retrospective data are limited, which is common in real-world hospital settings.

DA leverages knowledge from a source domain to improve a target domain performance with limited training data, and has been applied in computer vision [12] and natural language processing [13]. One of the simplest DA methods is fine-tuning (FT) [24], which first pretrains model parameters on a source dataset and then updates them with a target dataset. Although widely used [25], FT tends to be sub-optimal, especially when target training data is insufficient [18], due to catastrophic forgetting [26]. More advanced techniques have been proposed to deal with the challenge in few-shot scenarios. In particular, [19] proposes a Siamese architecture [27] to address visually supervised DA by learning an embedding subspace, in which mapped raw feature domains are semantically aligned but maximally separated with few labeled target data samples available. Under the same few-shot learning scenario, [18] provides a framework to exploit adversarial learning to identify the embedding subspace for feature alignment. However, it is not immediately apparent how to apply these methods to multi-hospital system EHRs with feature disparity because these EHRs are within a different input space.

For DA in multi-hospital system EHRs, [10] shows the effectiveness of FT on mortality prediction by pretraining on MIMIC III or eICU-CRD and fine-tuning on the other one with overlapping features. [14] seeks domain-invariant representations between two systems by adversarial learning for clinical task predictions. [15] proposes two causes of discrepancies between multi-hospital system EHRs: 1) covariate shift, caused by different patient distributions in different hospitals, and 2) systematic bias, caused by different administrative policies and workflows of different medical systems. However, they utilize only target and source datasets with overlapping features, and ignore the feature disparity among EHRs. This can significantly degrade performance in the target domain since the information provided by overlapping features between source and target is limited, and distinctive properties provided by domain-specific features are not fully utilized.

## III. Background

**Mortality Prediction** Let $\mathcal{X}^{|d|}$ be a $|d|$-dimensional space, where $d$ is a selected feature set based on particular hospital systems and/or sub-medical domains, and $\mathbf{x}$ is a data vector which represents clinical measurements, taking values in $\mathcal{X}^{|d|}$ following distribution $p(\mathbf{x})$. Supposing that $y \in \{0,1\}$ is the binary outcome indicator for each sample, where $y=0$ and $y=1$ denote discharge and death respectively, we can represent an EHR dataset as a collection of $N$ i.i.d. samples $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$.

In early-stage mortality prediction, given selected clinical measurements $\mathbf{x}$ in the first several hours after admission, we would like to predict whether a patient will die after certain periods:

$$y = \Phi(\mathbf{x}), \tag{1}$$

where $\Phi$ is a probabilistic model. In this paper, we focus on the training data scarcity scenario for the problem of interest.

**DA in EHRs** To learn a representative $\Phi$ in Eq. 1 in a data-scarce EHR dataset, given that another relevant and large EHR dataset is accessible, we cast the problem in terms of a DA problem. In the DA setting, we have a source dataset $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_s}$, where $\mathbf{x}_i^s \in \mathcal{X}^{|d_s|}$, and a target dataset $\mathcal{D}_t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{N_t}\}$, where $\mathbf{x}_i^t \in \mathcal{X}^{|d_t|}$, with $N_s \gg N_t$. Our goal is to produce an accurate survival outcome on the target domain $t$ with training data scarcity, by leveraging knowledge from the source domain $s$ with a sufficient amount of available data information, but with feature disparity, i.e. $|d_s \cup d_t| < |d_s| + |d_t|$, $|d_s| > |d_s \cap d_t|$ and $|d_t| > |d_s \cap d_t|$.

## IV. Methodology

In this section we introduce the private encoding technique to deal with feature disparity between multi-hospital system EHRs, and explain how it is equipped with three DA approaches: FT, ALPCA [18] and CLPSA [19] to fully utilize source knowledge and improve prediction performance in the target domain during inference.

### A. Private Encoding

There is no universal representation for raw features in EHRs. We have $\mathcal{X}^{|d_s|} \neq \mathcal{X}^{|d_t|}$, given that EHRs from different hospital systems or sub-domain populations meet discrepancies. To leverage DA, which requires target and source representations in the same space, we define a new space $\tilde{\mathcal{X}}$ as the hidden feature space. Instead of obtaining the space utilizing only overlapping features between domains, private encoders, $K_s$ and $K_t$, map all selected features, $\mathbf{x}^s$ and $\mathbf{x}^t$ from $\mathcal{X}^{|d_s|}$ and $\mathcal{X}^{|d_t|}$, respectively, to $\tilde{\mathcal{X}}$ to obtain the hidden feature vectors $\tilde{\mathbf{x}}^s$ and $\tilde{\mathbf{x}}^t$, by

$$\tilde{\mathbf{x}}^s = K_s(\mathbf{x}^s) \tag{2}$$

$$\tilde{\mathbf{x}}^t = K_t(\mathbf{x}^t), \tag{3}$$

following distributions $p(\tilde{\mathbf{x}}^s)$ and $p(\tilde{\mathbf{x}}^t)$. We assume that there is a covariate shift [28] between $\tilde{\mathbf{x}}^s$ and $\tilde{\mathbf{x}}^t$, such that the distribution $p(\tilde{\mathbf{x}}^s) \neq p(\tilde{\mathbf{x}}^t)$ in $\tilde{\mathcal{X}}$. Encoding all selected features in both domains encourages $\tilde{\mathbf{x}}^s$ and $\tilde{\mathbf{x}}^t$ to have more comprehensive representations for corresponding domains, and can further improve the mortality prediction performance.

### B. DA techniques with private encoding

DA approaches attempt to align the distributions of the source and target domains to obtain more domain-invariant information from the source, improving model performance in the target domain when training data is scarce. To utilize DA techniques in EHRs with feature disparity, we first map both source and target domain data into the same hidden space by private encoders, and then utilize two more components for both domains: shared encoders $H_s, H_t : \tilde{\mathcal{X}} \to \mathcal{Z}$, which take outputs from the source and target private encoders, respectively, to obtain shared domain-invariant representations; and classifiers, $C_s, C_t : \mathcal{Z} \to \mathcal{Y}$, which take outputs from $H_s$ and $H_t$ to predict mortality for the source and target domains, respectively. To improve mortality prediction in the target domain, we optimize binary cross entropy losses in both domains,

$$\hat{y} = C(H(\tilde{\mathbf{x}})), \tag{4}$$

$$\mathcal{L}_C = -\big[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\big], \tag{5}$$

with different DA techniques, where $\hat{y}$, $H$ and $C$ are symbolic representations of the predicted mortality, shared encoders and classifiers, respectively, for both the source and target domains. To encourage representations of the source and target towards becoming domain-invariant in the embedding space $\mathcal{Z}$, we share parameters between $H_s$ and $H_t$, i.e. $H_s = H_t = H$. To ensure that representations carry the domain-specific characteristics of source and target in $\mathcal{Y}$, we set $C_s \neq C_t$.

**Fine-tuning (FT)** FT is one of the most direct ways to apply DA. Specifically, FT is a method that adopts a model that has already been trained for a given task, and tunes or tweaks the model to perform on a second different but related task [29]. Here, we first pretrain networks with the source data for mortality prediction,

$$\hat{y}^s = C_s(H_s(\tilde{\mathbf{x}}^s)), \tag{6}$$

and then fine-tune with the available target data,

$$\hat{y}^t = C_t(H_t(\tilde{\mathbf{x}}^t)), \tag{7}$$

where $H_t$ is initialized by $H_s$ from Eq. 6.

**Adversarial Learning with Pairing Classes Alignment (ALPCA)** Traditionally, adversarial learning [30] introduces a min-max game training strategy to obtain domain-invariant knowledge by seeking a discriminator, $D$, that can identify samples from source and target distributions. After $D$ is learned, $H$, in the role of a generator, is updated to render $D$ unable to distinguish samples from the source and target domains. However, due to data scarcity, $H$ cannot estimate the target population accurately. Even with a perfect $H$, i.e. $D$ cannot distinguish whether a sample is from the source or target domain, $H$ still cannot guarantee that samples from different domains but with the same class label will map nearby in the embedding space, since no class information is provided to $D$ in standard adversarial training[18].

We follow [18] to alleviate the target training sample shortage issue in adversarial domain adaptation and encourage networks to learn the properties of death and discharge patients in two domains. Specifically, a pairing strategy [18] is used to overcome the scarcity of training data by creating four groups based on domain and class information, and a multi-class discriminator $D$ [18] is introduced by distinguishing four pairing groups to encourage $H$ to generate domain-invariant representations with class information.

As shown in Fig. 1 (encouraged by [18]), two groups ($\mathcal{G}_1$ and $\mathcal{G}_2$) consist of positive pairs and two ($\mathcal{G}_3$ and $\mathcal{G}_4$) consist of negative pairs. Each positive pair is composed of two samples with the same class, either `death-death` or `discharge-discharge`; while each negative pair is composed of two samples with different classes, i.e. `death-discharge`. Pairs in $\mathcal{G}_1$ and $\mathcal{G}_3$ consist of items both from the source domain, generated by randomly pairing samples drawn from the source distribution based on class information; while pairs in $\mathcal{G}_2$ and $\mathcal{G}_4$ consist of one item from the source and another from the target distribution, created by pairing each target sample with a number of randomly drawn source samples based on classes. In $\mathcal{G}_1$ and $\mathcal{G}_2$, we generate `death-death` and `discharge-discharge` with approximate ratio $1:1$ to encourage the networks to learn more similarity information between death patients, which is hard to achieve with very imbalanced medical data. We construct each group of the same size for training $D$ by matching the other three groups' size with the smallest one.

As demonstrated in Fig. 2, $H$ tries to fool $D$ by taking hidden representations from corresponding private encoders

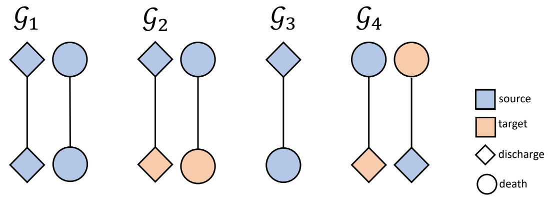


Fig. 1: An example illustration of pairs in four groups for ALPCA.

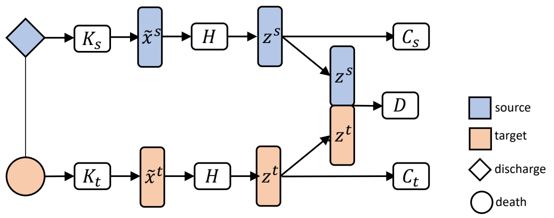


Fig. 2: ALPCA with a `discharge-death` example from $\mathcal{G}_4$.

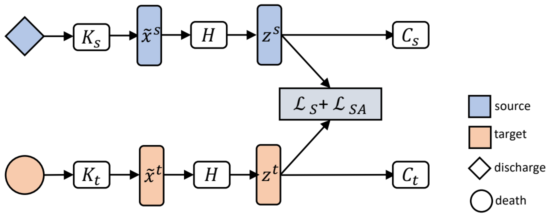


Fig. 3: CLPSA with a negative example.

($K_s$ and $K_t$) and outputting domain-invariant feature representations. $D$ takes the concatenation of domain-invariant representations to distinguish which group a sample pair comes from, trained via a standard cross entropy loss,

$$\mathcal{L}_D = -E\big[\sum_{i=1}^{4} y_{\mathcal{G}_i} \log(D(H(K(\mathcal{G}_i))))\big], \qquad (8)$$

where $y_{\mathcal{G}_i}$ denotes the label of group $i = 1, 2, 3, 4$, and $K$ denotes a symbolic representation of private encoders based on different pair groups, i.e. $K$ is two $K_s$ for $\mathcal{G}_1$ and $\mathcal{G}_3$, while one $K_t$ and one $K_s$ for $\mathcal{G}_2$ and $\mathcal{G}_4$. To output domain-invariant representations carrying class information, $H$ fools $D$ to identify sample pairs from $\mathcal{G}_2$ as $\mathcal{G}_1$, and pairs from $\mathcal{G}_4$ as $\mathcal{G}_3$, so that target samples can have indistinguishable representations with more diverse source samples. Mathematically, $H$ along with $K_t$ and $K_s$ are updated, with

$$\mathcal{L}_G = -E\big[y_{\mathcal{G}_1} \log(D(H(K(\mathcal{G}_2)))) + y_{\mathcal{G}_3} \log(D(H(K(\mathcal{G}_4))))\big]. \qquad (9)$$

Minimizing (9) together with the source and target classification losses,

$$\mathcal{L}_{G\&C} = \alpha\mathcal{L}_G + \beta\mathcal{L}_{C_s} + \mathcal{L}_{C_t}, \qquad (10)$$

where $\alpha$ and $\beta$ are hyper-parameters, encourages networks to obtain domain-invariant representations with class information and achieve good performance on classification tasks by considering distribution differences in $\mathcal{Y}$, simultaneously.

**Contrastive Learning with Pairing Semantic Alignment (CLPSA)** Instead of training an additional $D$ for aligning the feature distributions in $\mathcal{Z}$ to achieve the goal of DA, CLPSA [19] learns $\mathcal{Z}$ by pulling samples from different domains but the same class as close together as possible yet pushing samples from different domains and classes as far away as possible. Following [19], as shown in Fig.3, $H$ takes $\tilde{\mathbf{x}}^s$ and $\tilde{\mathbf{x}}^t$ to obtain domain-invariant representations by training with contrastive semantic alignment losses, consisting of a semantic alignment loss $\mathcal{L}_{SA}$ and a separation loss $\mathcal{L}_S$.

Specifically, to align the distributions of samples in the embedding space, a semantic alignment loss is introduced,

$$\mathcal{L}_{SA} = \sum_{a=0}^{1} d(p(H(\tilde{\mathbf{x}}_a^s)), p(H(\tilde{\mathbf{x}}_a^t))), \qquad (11)$$

where $\tilde{\mathbf{x}}_a^s$ and $\tilde{\mathbf{x}}_a^t$ are vectors from the source and target domains in $\tilde{\mathcal{X}}$ belonging to the same class $a$. $d$ is a metric to measure the distance between the distributions of $\tilde{\mathbf{x}}_a^s$ and $\tilde{\mathbf{x}}_a^t$ in $\mathcal{Z}$. $\mathcal{L}_{SA}$ prompts samples with the same class from two different domains to map nearby in the embedding space.

Although pulling the same class samples from different domains close together in the embedding space encourages target groups to obtain information from similar points in the source, minimizing $\mathcal{L}_{SA}$ does not guarantee that points in different classes from two domains are separated enough, which would significantly degrade performance in the target domain. Therefore, we leverage a separation loss,

$$\mathcal{L}_S = \sum_{a=0,b=0|a\neq b}^{1} k(p(H(\tilde{\mathbf{x}}_a^s)), p(H(\tilde{\mathbf{x}}_b^t))) \qquad (12)$$

where $k$ is a metric to measure the similarity between the distributions $\tilde{\mathbf{x}}_a^s$ and $\tilde{\mathbf{x}}_b^t$ in $\mathcal{Z}$. $\mathcal{L}_S$ encourages class separation by pushing the representations of different classes in two domains farther away, i.e. adding a penalty if the distance between distributions $p(H(\tilde{\mathbf{x}}_a^s))$ and $p(H(\tilde{\mathbf{x}}_b^t))$ is small.

Finally, CLPSA is jointly trained with classification losses from both domains and contrastive semantic alignment losses,

$$\mathcal{L}_{CCSA} = \mathcal{L}_{C_t} + \gamma\mathcal{L}_{C_s} + \delta(\mathcal{L}_{SA} + \mathcal{L}_S), \qquad (13)$$

where $\gamma$ and $\delta$ are hyper-parameters.

Similar to ALPCA, it is difficult to minimize Eq.13 when the target training data is scarce because $\mathcal{L}_{SA}$ and $\mathcal{L}_S$ depend on calculating distances and similarities between distributions, and those need to learn with sufficient samples. Therefore we pair each target sample to a large number of randomly selected source samples and compute average pairwise distances between positive pairs

$$d(p(H(\tilde{\mathbf{x}}_a^s)), p(H(\tilde{\mathbf{x}}_a^t))) = \sum_{i,j} d(H(\tilde{x}_i^s), H(\tilde{x}_j^t)), \qquad (14)$$

where $y_i^s = y_j^t = a$; or similarities between negative pairs

$$k(p(H(\tilde{\mathbf{x}}_a^s)), p(H(\tilde{\mathbf{x}}_b^t))) = \sum_{i,j} k(H(\tilde{x}_i^s), H(\tilde{x}_j^t)), \qquad (15)$$

where $y_i^s = a \neq y_j^t = b$, between points in the embedding space to achieve semantic alignment. Here, the ratio between positive and negative sample pairs is 1:1, and the ratio between the `death-death` and `discharge-discharge` pairs is also 1:1 in the positive group, to encourage the network to learn more information from the records of patients that died, in an imbalanced dataset.

We implement $\mathcal{L}_{SA}$ and $\mathcal{L}_{S}$ with contrastive loss following [31]

$$d(H(\tilde{x}_i^s), H(\tilde{x}_j^t)) = \frac{1}{2}\|H(\tilde{x}_i^s) - H(\tilde{x}_j^t)\|^2, \quad (16)$$

$$k(H(\tilde{x}_i^s), H(\tilde{x}_j^t)) = \frac{1}{2}max(0, m - \|H(\tilde{x}_i^s) - H(\tilde{x}_j^t)\|)^2, \quad (17)$$

where $\|\cdot\|$ denotes the Frobenius norm and $m$ is a margin to define the separability in the embedding space.

## V. Experiments

### A. Datasets

We conducted experiments on a source dataset extracted from MIMIC III and two target datasets: the UCSF dataset and the EICU dataset. For the missing data issue [32], each dataset was preprocessed in the same way. First, we excluded patients with more than $50\%$ of features missing, and then we applied MICE [33] data imputation for the remaining missing values. The statistics of these processed datasets are summarized in Table I, and details are described below.

TABLE I: Dataset Statistics

| | Source | Target | |
|---|---|---|---|
| | MIMIC III | UCSF | EICU |
| # patients | 29914 | 2069 | 2565 |
| # death | 3063 | 342 | 204 |
| # discharge | 26851 | 1727 | 2361 |
| # feature | 21 | 31 | 35 |
| #overlapping | 21 | 18 | 12 |

**MIMIC III** is a public data source of de-identified EHRs, which contains 53,423 patients admitted to ICUs at a Boston-area hospital from 2001–2012 [9]. We extracted the data following [34] and selected the 17 most common clinical features (e.g. heart rate, blood pressure, temperature and respiratory rate, etc.), as well as 4 demographic features (ethnicity, gender, age and weight) for the mortality prediction. As we focus on prediction with early-stage measurements, we took the first appearance of each clinical feature measurement in the first two hours after admission if it is available; otherwise, it was regarded as a missing value. After data preprocessing, the source dataset consisted of 29914 patients with 21 features in total. We randomly selected $80\%$ of the data points as the training set, and the rest as the validation set.

**UCSF Dataset**, collected from the UCSF/San Francisco General Hospital and Trauma Center, contains 2,190 patients admitted to a Level I trauma center. We selected demographic information, injury score, physical vital signs and laboratory results[1] measured at the time of admission or during the first two hours after admission, as features for mortality prediction. After data preprocessing, we have 2,069 patients with 31 features, including 18 features overlapping with MIMIC III. We randomly selected 784 and 785 patients as the validation and testing sets, respectively, and further randomly drew different training sizes from the remaining patients to simulated data scarcity at various levels.

**EICU Dataset** is extracted from the eICU-CRD, a multi-center ICU database with high granularity data for over 200,000 admissions to ICUs monitored by eICU programs across the United States. We selected the first-time ICU admission of each adult patient (age $>$ 18) with a diagnosis related to trauma. Then we queried the minimum and maximum of clinical measurements (e.g. blood urea nitrogen, white blood cell count and hemoglobin, etc.) and demographics information taken in the first 24 hours of a patient's ICU stay, following the code shared by the eICU research community[2], to predict the mortality after the first admission day. After data prepossessing, we have 2565 trauma patients with a death and discharge ratio of approximately 1:12, and 35 features containing 12 overlaps with MIMIC III. 1032 and 1033 patients were randomly selected as validation and test sets, and the remaining were randomly drawn for various training size scenarios, which is the same as UCSF dataset.

### B. Experiment settings

*1) Evaluation Metric:* We measured the performance of three DA methods and baselines by the F1-score, with 0.5 as the prediction threshold following the previous work [35].

*2) Model configurations:*

- **MLP-target (baseline)**: This is a multi-layer perceptron (MLP) composed of $K_t$, $H$ and $C_t$, trained only on the target dataset.
- **FT**: An MLP model composed of $K_s$, $H$ and $C_s$ was first pretrained on source data. The best performing model, evaluated by the source validation set, was saved. Then another MLP model composed of $K_t$, $H$ and $C_t$ were trained on target data, with $K_t$ and $C_t$ randomly initialized and $H$ inherited from the pretraining step.
- **ALPCA**: As shown in Fig. 2, all six networks ($K_s$, $K_t$, $H$, $D$, $C_s$ and $C_t$) in the framework are MLPs with random initialization, and the discriminator is trained with Eq. 8. The other five networks are trained with Eq. 10 following standard adversarial training schema [30].
- **CLPSA**: As shown in Fig. 3, five networks ($K_s$, $K_t$, $H$, $C_s$ and $C_t$) in the framework are MLPs with random initialization. They are trained with Eq. 13.

For the neural networks above, we used batch normalization to normalize the input layer by re-centering and re-scaling. For fair comparison, we assigned the same structures for $K_t$, $H$ and $C_t$ in the MLP-target and FT, respectively. In ALPCA and CLPSA, $K_s$, $K_t$, $H$, $C_s$ and $C_t$ have the same structure as FT. $D$ in ALPCA is a two-layer MLP. The size of each hidden layer in all networks was selected by grid search among $\{8, 16, 32\}$. We implemented all models in PyTorch [36] and all of the neural networks were trained with Adam [37], whose learning rates were selected by grid search among $\{0.0001, 0.0002, 0.0005\}$[3].

[1] Lab tests focusing on trauma patients: Protein C, D-Dimer, ATIII, Factor II, V, VII, VIII, IX and X

[2] https://github.com/mit-lcp/eicu-code

[3] All other hyper-parameters, e.g. $\alpha$, $\beta$ and $m$ etc. were selected by grid search in the same ranges for fair comparison. We omitted these due to space limitations.

TABLE II: F1 comparison (%) of MLP-target, FT, ALPCA and CLPSA with different training data size on the UCSF and EICU datasets. Mean values along with their standard deviations in the subscript were calculated with 5 data splits.

| | **Training size** | 50 | 100 | 200 | 300 | 400 | 500 |
|---|---|---|---|---|---|---|---|
| **Dataset** | **Model** | | | | | | |
| UCSF | MLP-target | $36.2_{4.7}$ | $36.4_{0.4}$ | $39.5_{6.1}$ | $46.8_{4.8}$ | $46.7_{2.0}$ | $47.4_{0.3}$ |
| | FT | $50.3_{7.1}$ | $51.4_{7.7}$ | $58.1_{4.6}$ | $59.9_{4.4}$ | $60.1_{2.8}$ | $58.0_{2.6}$ |
| | ALPCA | $48.6_{7.3}$ | $\mathbf{57.5}_{2.1}$ | $\mathbf{59.4}_{1.4}$ | $60.4_{1.2}$ | $59.9_{0.9}$ | $60.7_{1.7}$ |
| | CLPSA | $\mathbf{52.3}_{3.4}$ | $54.2_{4.3}$ | $59.3_{2.6}$ | $\mathbf{62.3}_{2.3}$ | $\mathbf{64.5}_{1.5}$ | $\mathbf{63.7}_{2.2}$ |
| EICU | MLP-target | $17.7_{7.7}$ | $20.6_{2.7}$ | $21.4_{1.3}$ | $21.9_{1.9}$ | $20.4_{1.7}$ | $22.3_{1.4}$ |
| | FT | $23.5_{7.0}$ | $21.8_{5.1}$ | $26.3_{3.7}$ | $26.7_{5.6}$ | $25.7_{4.5}$ | $27.9_{1.7}$ |
| | ALPCA | $26.1_{5.4}$ | $25.3_{6.5}$ | $30.7_{3.2}$ | $29.8_{3.2}$ | $35.6_{4.5}$ | $36.3_{1.9}$ |
| | CLPSA | $\mathbf{28.1}_{6.4}$ | $\mathbf{30.4}_{5.1}$ | $\mathbf{34.0}_{4.5}$ | $\mathbf{37.5}_{2.9}$ | $\mathbf{40.8}_{3.5}$ | $\mathbf{41.0}_{2.1}$ |

TABLE III: F1 comparison (%) of ablation study on three reasons for discrepancies of DA methods on the UCSF and EICU datasets.

| **Model** | MLP-target | FT | | | ALPCA | | | CLPSA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | `C-only` | `S+C` | `F+S+C` | `C-only` | `S+C` | `F+S+C` | `C-only` | `S+C` | `F+S+C` |
| UCSF | $47.4_{0.3}$ | $29.0_{2.2}$ | $38.9_{3.5}$ | $\mathbf{58.0}_{2.6}$ | $54.6_{4.1}$ | $55.6_{1.7}$ | $\mathbf{60.7}_{1.7}$ | $54.8_{2.1}$ | $53.7_{2.6}$ | $\mathbf{63.7}_{2.2}$ |
| EICU | $22.3_{1.4}$ | $18.3_{0.7}$ | $18.3_{0.7}$ | $\mathbf{27.9}_{1.7}$ | $17.34_{0.8}$ | $15.5_{3.11}$ | $\mathbf{36.3}_{1.9}$ | $16.5_{2.4}$ | $15.8_{0.7}$ | $\mathbf{41.0}_{2.1}$ |

### *C. Performance comparison*

To evaluate three DA methods, we randomly selected $\{50, 100, 200, 300, 400, 500\}$ data points as the training set, to simulate insufficient training data situations for both the UCSF and EICU datasets. We randomly selected 5 data splits for each training size scenario, and performed 10 different runs for each of them. For each data split, we report the corresponding testing set results based on the top 1 run on validation sets. The mean value along with the standard deviation of these 5 data splits were calculated for each training data size.

Our results are summarized in Table II. All three DA methods yielded better performance than the MLP-target in both datasets across various training data scenarios, demonstrating the effectiveness of DA in small training data regimes. Both ALPCA and CLPSA achieved better performance than FT. CLPSA outperformed ALPCA or achieved comparable results in the UCSF dataset, and consistently performed better than ALPCA in the EICU dataset, across the entire range of training data sizes. The result that ALPCA underperforms CLPSA may be primarily due to introducing an additional network (discriminator), making the whole framework more challenging to optimize, with more parameters to update [38]. Surprisingly, the three DA methods with 50 training data points consistently outperformed the MLP-target with 500 training data points (10 times larger), which further strengthens the powerful capability of DA in small EHR data regimes.

### *D. Ablation study on modeling feature disparity.*

We have demonstrated the effectiveness of DA methods in Table II, where feature disparity (`F`) is modeled by private encoding strategy, as well as systematic bias (`S`) and covariate shift (`C`). We denote this setting as `F+S+C`. To verify the effectiveness of modeling feature disparity in DA, we considered two variants for both EHRs in the 500 patients training data scenario. First, we considered both systematic bias and covariate shift but not feature disparity, denoting as `S+C`. The networks were trained on both the source and target data, only including overlapping features and not sharing parameters on private encoders. Second, we trained the networks on two domains with overlapping features and shared parameters on private encoders, counting only the discrepancies caused by covariate shift, which is denoted as `C-only`.

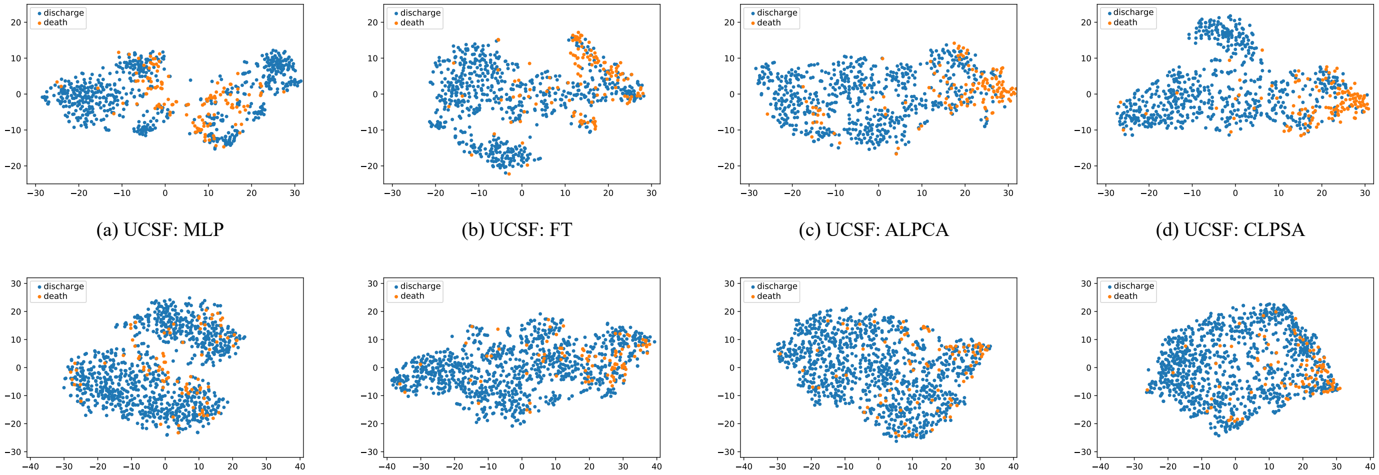
Table III presents the results of our ablation study on three different causes of discrepancies, as well as the MLP-target. The empirical analysis shows that `S+C` is not always better than `C-only`, or vice versa and they may even underperform MLP-target, which utilizes all features in target domain including overlapping ones. However, our method, `F+S+C`, consistently yields significantly better performance than `S+C` and `C-only` , demonstrating the importance of modelling feature disparity in DA of multi-hospital systems EHRs, especially when the overlapping features between systems are limited.

### *E. Analysis*

To understand why DA improves prediction performance on target datasets and why CLPSA yields the best performance compared to other DA strategies, we further analyzed 2-dimensional (2-d) embeddings of testing sets on the UCSF and EICU datasets, respectively. Specifically, we obtain 2-d embeddings by reducing high dimensional representations before feeding into the $C_t$ using t-SNE [39] in Scikit-learn [40]. Then we calculate the average difference between inter-cluster and intra-cluster distance in the death group, which is more important than the discharge group in the medical domain, yielding $d_{\texttt{diff}} = d_{\texttt{inter}} - d_{\texttt{intra}} = \frac{\sum_i^n \|e_i - c_{\texttt{discharge}}\|^2 - \sum_i^n \|e_i - c_{\texttt{death}}\|^2}{n}$, where $n$ represents the number of patients who died, $e_i$ is $i$-th patient's 2-d embedding, and $c_{\texttt{discharge}}$ and $c_{\texttt{death}}$ are centers of the discharge and death clusters in the 2-d embedding space, respectively. Inter-cluster distance is an average distance between members of a cluster and another cluster's center; meanwhile, intra-cluster distance is an average distance between members and their own center. We want the inter-cluster distance to be large to push the cluster far away from the other, but the intra-cluster distance to be small to pull members in a cluster as close as possible. Thus

TABLE IV: 2d t-SNE embedding distance evaluation of testing set in the UCSF and EICU datasets.

| | **Training Size** | 50 | 100 | 200 | 300 | 400 | 500 |
|---|---|---|---|---|---|---|---|
| **Dataset** | **Model** | | | | | | |
| UCSF | MLP-target | 5.7 | 5.9 | 5.1 | 8.2 | 9.3 | 8.3 |
| | FT | 6.9 | 10.8 | 11.8 | 13.0 | 13.9 | 12.9 |
| | ALPCA | **11.3** | 11.7 | **13.4** | 11.1 | 13.8 | 11.9 |
| | CLPSA | 8.2 | **12.1** | 12.9 | **14.5** | **15.8** | **14.0** |
| EICU | MLP -target | −0.2 | 0.8 | 2.3 | 1.7 | 2.7 | 2.9 |
| | FT | 3.7 | 3.6 | 4.7 | 3.1 | 3.6 | 6.1 |
| | ALPCA | **4.2** | 4.6 | **4.9** | 3.1 | 7.1 | 3.3 |
| | CLPSA | 3.6 | **5.5** | 4.1 | **5.0** | **7.2** | **6.5** |



(a) UCSF: MLP

(b) UCSF: FT

(c) UCSF: ALPCA

(d) UCSF: CLPSA

(a′) EICU: MLP

(b′) EICU: FT

(c′) EICU: ALPCA

(d′) EICU: CLPSA

Fig. 4: Testing set 2d t-SNE embedding for the UCSF and EICU dataset with the 500 training data scenario. The blue and orange dots represent the discharged and dead patients, respectively.

the $d_{\texttt{diff}}$ should be large to make the cluster easy to identify by $C_t$.

Average $d_{\texttt{diff}}$ results corresponding to various training data sizes in Table II are summarized in Table IV. Consistent with the F1 scores in Table II, CLPSA achieved a greater or comparable $d_{\texttt{diff}}$, compared to other methods for both the UCSF and EICU datasets, which indicates that hidden representations generated by CLPSA are prone to be identified by $C_t$ compared to other models.

We further visualized the 2-d t-SNE of all models from the same data split on both datasets with 500 training points in Fig. 4. The 2d t-SNEs of death and discharge patients on both the UCSF and the EICU dataset for MLP-target in Fig.(a) and Fig.(a′) are almost overlapping, indicating that it is difficult to find a straight line to distinguish the two groups, making $C_t$ prediction of mortality challenge. For the other DA models, the cluster of the death group often aggregates at the right in each plot, making it more straightforward to separate with the discharge patients compared to the MLP-target, and illustrating the reason for improvement by utilizing DA. Comparing the 2-d t-SNEs of different DA methods, we find that the cluster of death patients with CLPSA in Fig.(d) and Fig.(d′) is more concentrated than that with FT in Fig.(b) and Fig.(b′) and with ALPCA in Fig.(c) and Fig.(c′), which explains the better performance of CLPSA compared to other DA strategies.

## VI. Conclusion

In this paper we showed how DA methodologies, in particular FT, ALPCA and CLPSA, can be used to improve the performance of mortality prediction for trauma patients in regimes with limited training data. In contrast to existing DA methodologies in multi-hospital system EHR predictive tasks, which consider only the discrepancies caused by covariate shift and systematic bias, we bridge the gap of feature disparity by introducing a private encoding strategy which maps clinical measurements from different raw feature spaces to a hidden feature space and follows with various DA techniques. Extensive experimental results on two datasets demonstrate the usefulness of DA, and ablation studies and 2-d t-SNE analysis further explain the effectiveness of private encoding and DA methods, respectively.

## Acknowledgment

This work was funded by the National Institutes for Health (NIH) grant NIH 7R01HL149670.

## References

[1] W. H. Organization *et al.*, "Injuries and violence: the facts 2014," 2014.

[2] O. Blow, L. Magliore, J. A. Claridge, K. Butler, and J. S. Young, "The golden hour and the silver day: detection and correction of occult hypoperfusion within 24 hours improves outcome from major trauma," *Journal of Trauma and Acute Care Surgery*, vol. 47, no. 5, p. 964, 1999.

[3] M. J. Otero-López, P. Alonso-Hernández, J. A. Maderuelo-Fernández, B. Garrido-Corro, A. Domínguez-Gil, and A. Sánchez-Rodríguez, “Preventable adverse drug events in hospitalized patients,” *Medicina clinica*, vol. 126, no. 3, pp. 81–87, 2006.

[4] C. Subbe, A. Slater, D. Menon, and L. Gemmell, “Validation of physiological scoring systems in the accident and emergency department,” *Emergency Medicine Journal*, vol. 23, no. 11, pp. 841–845, 2006.

[5] T. Desautels, J. Calvert, J. Hoffman, Q. Mao, M. Jay, G. Fletcher, C. Barton, U. Chettipally, Y. Kerem, and R. Das, “Using transfer learning for improved mortality prediction in a data-scarce hospital setting,” *Biomedical informatics insights*, vol. 9, p. 1178222617712994, 2017.

[6] B. Shickel, P. J. Tighe, A. Bihorac, and P. 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*, vol. 22, no. 5, pp. 1589–1604, 2017.

[7] C. Xiao, E. Choi, and J. Sun, “Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review,” *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 2018.

[8] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, “Self-training with noisy student improves imagenet classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

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

[10] Z. Sun, S. Peng, Y. Yang, X. Wang, and F. Li, “A general fine-tuned transfer learning model for predicting clinical task acrossing diverse ehrs datasets,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 490–495.

[11] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, “A theory of learning from different domains,” *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010.

[12] G. Csurka, “Domain adaptation for visual applications: A comprehensive survey,” *arXiv preprint arXiv:1702.05374*, 2017.

[13] Y. Ziser and R. Reichart, “Neural structural correspondence learning for domain adaptation,” *arXiv preprint arXiv:1610.01588*, 2016.

[14] S. Purushotham, W. Carvalho, T. Nilanon, and Y. Liu, “Variational recurrent adversarial deep domain adaptation,” 2016.

[15] F. Khoshnevisan and M. Chi, “An adversarial domain separation framework for septic shock early prediction across ehr systems,” in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 64–73.

[16] Y. Zhang, T. B. Wu, B. J. Daigle, M. Cohen, and L. Petzold, “Identification of disease states associated with coagulopathy in trauma,” *BMC medical informatics and decision making*, vol. 16, no. 1, pp. 1–9, 2016.

[17] J. B. MacLeod, M. Lynn, M. G. McKenney, S. M. Cohn, and M. Murtha, “Early coagulopathy predicts mortality in trauma,” *Journal of Trauma and Acute Care Surgery*, vol. 55, no. 1, pp. 39–44, 2003.

[18] S. Motiian, Q. Jones, S. M. Iranmanesh, and G. Doretto, “Few-shot adversarial domain adaptation,” *arXiv preprint arXiv:1711.02536*, 2017.

[19] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, “Unified deep supervised domain adaptation and generalization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5715–5725.

[20] L. van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[21] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, “The eicu collaborative research database, a freely available multi-center database for critical care research,” *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.

[22] S. Purushotham, C. Meng, Z. Che, and Y. Liu, “Benchmarking deep learning models on large healthcare datasets,” *Journal of biomedical informatics*, vol. 83, pp. 112–134, 2018.

[23] J. Fang, J. Zhu, and X. Zhang, “Prediction of length of stay on the intensive care unit based on bayesian neural network,” in *Journal of Physics: Conference Series*, vol. 1631, no. 1. IOP Publishing, 2020, p. 012089.

[24] S. Kornblith, J. Shlens, and Q. V. Le, “Do better imagenet models transfer better?” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, 2019.

[25] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, “Convolutional neural networks for medical image analysis: Full training or fine tuning?” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[26] Z. Li and D. Hoiem, “Learning without forgetting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 2935–2947, 2018.

[27] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, “Siamese neural networks for one-shot image recognition,” in *ICML deep learning workshop*, vol. 2. Lille, 2015.

[28] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.

[29] L. Torrey and J. Shavlik, “Transfer learning,” in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.

[31] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[32] C. Zhang, V. Maroufy, B. Chen, and H. Wu, “Missing data issues in ehr,” in *Statistics and Machine Learning Methods for EHR Data*. Chapman and Hall/CRC, 2020, pp. 149–173.

[33] S. v. Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in r,” *Journal of statistical software*, pp. 1–68, 2010.

[34] S. Wang, M. B. McDermott, G. Chauhan, M. Ghassemi, M. C. Hughes, and T. Naumann, “Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii,” in *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 222–235.

[35] K. Lin, Y. Hu, and G. Kong, “Predicting in-hospital mortality of patients with acute kidney injury in the icu using random forest model,” *International journal of medical informatics*, vol. 125, pp. 55–61, 2019.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

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

[38] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

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

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.