Unifying Domain Adaptation and Domain Generalization for Robust Prediction across Minority Racial Groups

Farzaneh Khoshnevisan[✉][1] and Min Chi[2]

[1] Intuit Inc., San Diego, USA
farzaneh_khoshnevisan@intuit.com
[2] North Carolina State University, Department of Computer Science, Raleigh, USA
mchi@ncsu.edu

Abstract. In clinical deployment, the performance of a model trained from one or more medical systems often deteriorates on another system and such deterioration is especially evident among minority patients who often have limited data. In this work, we present a multi-source adversarial domain separation (MS-ADS) framework which unifies domain adaptation and domain generalization. MS-ADS is designed to address two types of discrepancies: *covariate shift* stemming from differences in patient populations, and *systematic bias* on account of differences in data collection procedures across medical systems. We evaluate MS-ADS for early prediction of *septic shock* on three tasks. On *a task of domain adaptation* across three medical systems, we show that by leveraging data from multiple systems while accounting for both types of discrepancies, MS-ADS improves the prediction performance across all three systems; on *a task of domain generalization to an unseen medical system*, we show that MS-ADS can perform better than or close to the gold standard supervised models built for the system; last but not least, on a task that involves both domain adaptation and domain generalization: *generalization to unseen racial groups across medical systems*, MS-ADS shows robust out-performance by addressing covariate shift across different racial groups and systematic bias across medical systems simultaneously.

Keywords: Domain Adaptation · Domain Generalization · Cross-racial Transfer · Septic Shock.

1 Introduction

Machine learning is used increasingly in clinical care to improve diagnosis, treatment policy, and healthcare efficiency. Because machine learning models learn from historically collected data, electronic health records (EHRs), populations that are under-represented in the training data are often vulnerable to harm by incorrect predictions. For example, between the two medical systems involved in this work, the percentages of White vs. African American are 71% vs. 22.5% in Christiana Care whereas 91% vs. 3% in Mayo clinic. For certain

diseases like sepsis, different racial groups often exhibit distinct progression patterns [35]. Therefore, a model that can leverage EHRs across multiple medical systems to improve prediction among minority racial groups is needed. However, EHRs across medical systems can vary dramatically because different systems serve different demographic populations and often employ different infrastructure, workflows and administrative policies [1]. For this work, we refer to the discrepancies caused by the heterogeneous patient populations as *covariate shift* and those caused by incompatible data collection procedures as *systematic bias*.

We propose a multi-source adversarial domain separation (MS-ADS) framework which unifies domain adaptation and domain generalization. More specifically, MS-ADS separates the local representation of each domain from the global latent representation across all domains to address *systematic bias* and leverages multi-domain discriminator in conjunction with gradient reversal layer to address the *covariate shift* across each pair of domains. More specifically, our MS-ADS is built atop variational recurrent neural networks (VRNN) [5] due to VRNN's ability to handle variabilities in EHRs, such as missing data, and its ability to capture complex conditional and temporal dependencies [26,39]; it is shown that VRNN significantly outperforms commonly-used variations of RNN such as long short-term memory (LSTM) on EHRs [16, 39]. The effectiveness of MS-ADS is compared against another strong VRNN-based domain adaptation framework called VRADA [28] for early prediction of a challenging condition in hospitals, septic shock. Sepsis is a life-threatening condition caused by a dysregulated body response to infection [32]. Septic shock is the most severe complication of sepsis, associated with high mortality rate and prolonged length of hospitalization [32]. Timing is critical for this condition as every hour delay in antibiotic treatment leads to 8% increase in the chance of mortality. Early prediction of septic shock is challenging due to vague symptoms and subtle body responses [19]. Also, sepsis, like cancer, involves various disease etiologies that span a wide range of syndromes, and different patient groups may show vastly different symptoms [35].

To investigate the early prediction of septic shock, we leverage EHRs collected from three large medical systems located in different parts of the US. The effectiveness of MS-ADS is evaluated on three tasks involving domain adaptation (DA), domain generalization (DG), or both. First, on a task of *DA across three medical systems*, we compare MS-ADS against VRADA and a VRNN model trained on all three domains and show that MS-ADS improves the prediction performance across the three domains and outperforms all baselines. Further, through visualization we show that MS-ADS indeed capture both covariate shift and systematic bias. Second, on a task of *DG to an unseen system*, we evaluate the performance of MS-ADSs trained with two medical systems on a third target system. The results suggest that MS-ADS can perform as well as or better than the gold standard: supervised model trained on the target domain. Finally, probably the most important, we evaluate MS-ADS on the task of *generalization to an unseen racial group across medical systems*. We demonstrate that by treating each medical system and each racial group as a separate domain, our MS-ADS is capable of addressing both covariate shifts across different racial groups and

systematic bias across medical systems. Our results suggest that MS-ADS significantly improves generalization performance to African American population in Mayo as compared to the other baselines. Our contributions are:

- By tackling two different types of discrepancies, MS-ADS can effectively leverage EHRs from multiple *medical systems* to improve prediction performance on each system individually and also combined.
- Domain-invariant representations generated by MS-ADS are generalizable to new domains such that they perform close to or better than the gold standard supervised models trained on those systems.
- By unifying DA and DG, as far as we know, MS-ADS is the first framework that shows great potential on generalization to unseen racial groups across medical systems.

2 Methodology

Problem Description We have K domains: $\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K\}$ and a domain contains n patient visits represented as $\boldsymbol{X} = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$. Each visit $\mathbf{x}^i$ is a multivariate time-series that is composed of T^i medical events and can be denoted as $\mathbf{x}^i = (\boldsymbol{x}_t^i)_{t=1}^{T^i}$ where $\boldsymbol{x}_t^i \in \mathbb{R}^D$. Additionally, each visit has a visit-level outcome label represented as $\boldsymbol{Y} = \{\mathrm{y}^1, ..., \mathrm{y}^n\}$ where $\mathrm{y}^\mathrm{i} \in \{1, 0\}$ indicates the outcome of visit i: septic shock or non-septic shock. By combining $\boldsymbol{X}$ and $\boldsymbol{Y}$ for each domain, we have: $\mathcal{D}_k = \{\mathbf{x}_{\mathcal{D}_k}^i, \mathrm{y}_{\mathcal{D}_k}^i\}_{i=1}^{n_k}$, where n_k is the number of visits in $\mathcal{D}_k$; Here we assume each $\{\mathbf{x}_{\mathcal{D}_k}^i, \mathrm{y}_{\mathcal{D}_k}^i\}_{i=1}^{n_k}$ is drawn from distribution $p_k(\mathbf{x}, \mathrm{y})$ that is different from $\{p_j(\mathbf{x}, \mathrm{y}) : j \neq k\}$. Our objective is to minimize the discrepancies between these K domains in a common latent space by aligning their latent representations: $\mathbf{z}_{\mathcal{D}_1}, ..., \mathbf{z}_{\mathcal{D}_K}$, so that to create a unified, generalizable classifier $C : \mathbf{z} \mapsto \mathrm{y}$ that predicts the outcome optimally in *all* K domains. To do so, we adversarially learn $\binom{K}{2}$ discriminators to minimize the distance between global latent representations of each pair of domain $\mathbf{z}_{\mathcal{D}_i}$ and $\mathbf{z}_{\mathcal{D}_j}$. We describe this framework in detail in the following.

Multi-Source Adversarial Domain Separation (MS-ADS) Fig. 1 illustrates MS-ADS architecture: it separates one globally-shared latent representation for all domains from domain-specific (local) information. This architecture would allow global information to be purified so that the discrepancies caused by systematic bias are addressed. MS-ADS employs VRNN as the base model to process sequential input EHRs. VRNN has an encode-decoder structure where its four internal operations interact with each other to capture dependencies between latent random variables across time steps (please see [5] for more details). MS-ADS ensures that the global latent representations are different from the local ones by maximizing a dissimilarity measure. Additionally, multiple domain discriminators and a label predictor are employed to ensure domain-invariant and class-discriminative projection. In the following, we briefly describe the two steps for training the MS-ADS framework.

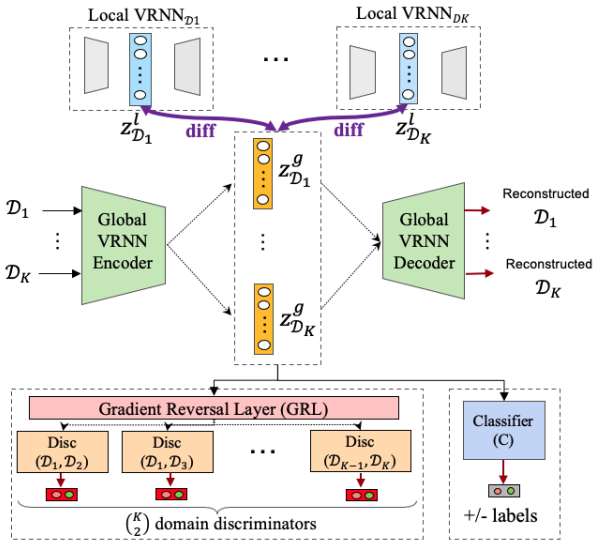


Fig. 1: Multi-Source Adversarial Domain Separation (MS-ADS) Framework

Step 1: Pre-train Source and Target VRNNs. Optimal local latent representations $\mathbf{z}^l_{\mathcal{D}_1}, ..., \mathbf{z}^l_{\mathcal{D}_K}$ are obtained by pre-training a local VRNN per each domain separately. The VRNN's loss objective ($\mathcal{L}^l_{\text{vrnn}}$) optimizes the inference (encoder) and the generative (decoder) processes to minimize the reconstruction loss [5].

Step 2: Discriminative Adversarial Separation. As shown in Fig. 1, MS-ADS is composed of the K pre-trained local VRNNs from step 1, and one global VRNN that takes the concatenation of all K domains as input. The global Encoder will generate global latent representations, and the global decoder reconstructs the input for each domain. The set of discriminators align the global latent representations between every two domains from $\mathcal{D}_i$ and $\mathcal{D}_j$. Finally, the unified classifier learns to predict the outcome labels using all latent representations regardless of their source domain. Following formalizes each component's loss objective.

1. **Global and Local VRNNs**: The parameters of local VRNN$_{\mathcal{D}_1}$,..., VRNN$_{\mathcal{D}_K}$ are initialized based on the K pre-trained VRNNs to generate local representations: $\mathbf{z}^l_{\mathcal{D}_1}, ..., \mathbf{z}^l_{\mathcal{D}_K}$. The global VRNN also takes concatenation of all domain's data as input and the global encoder generates $\mathbf{z}^g_{\mathcal{D}_1}, ..., \mathbf{z}^g_{\mathcal{D}_K}$. Further, for optimizing reconstruction loss in each of the local and global VRNNs we follow the original VRNN loss as follows:

$$\mathcal{L}^l_{\text{vrnn}}(\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}; \Theta^l) = \sum_{i=1}^{K} \mathcal{L}_{\text{vrnn}}(\mathbf{x}_{\mathcal{D}_i}; \theta_{e_i}, \theta_{d_i}) \tag{1}$$

$$\mathcal{L}^g_{\text{vrnn}}([\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}]; \Theta^g) = \mathcal{L}_{\text{vrnn}}([\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}]; \Theta^g) \tag{2}$$

where $\Theta^l = \bigcup_{i=1}^{K} (\theta_{e_i}, \theta_{d_i})$ and $\Theta^g = (\theta^g_e, \theta^g_d)$ indicate the local and global VRNN parameters, respectively.

The main novelty of MS-ADS is to *separate* local and global features by maximizing the distance between them so that they extract system specific features such as systematic bias. To do so, we add a dissimilarity measurement between $(\mathbf{z}^g_{\mathcal{D}_i}, \mathbf{z}^l_{\mathcal{D}_i})$ for all $\mathcal{D}_i, i \in \{1, ..., K\}$ for each sample, defined by a *Frobenius norm* which measures the orthogonality between global and local representation from each domain. Let us denote matrix $\boldsymbol{Z}^g_{\mathcal{D}_i}$ as global matrix of $\mathcal{D}_i$ where each row j of it is composed of $\mathbf{z}^g_{\mathcal{D}_i}$ for sample j in this domain. Similarly, $\boldsymbol{Z}^l_{\mathcal{D}_i}$ indicates local matrix of $\mathcal{D}_i$. Therefore, the difference loss is defined as:

$$\mathcal{L}_{\text{diff}}(\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}; \Theta^l, \Theta^g) = \sum_{i=1}^{K} \left\| {\mathbf{Z}^g_{\mathcal{D}_i}}^\top \mathbf{Z}^l_{\mathcal{D}_i} \right\|^2_F \tag{3}$$

where $\|\cdot\|^2_F$ refers to the squared Frobenius norm where zero indicates orthogonal vectors. Finally, the overall separation loss is:

$$\begin{aligned}\mathcal{L}_{\text{sep}}(\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}; \Theta) =& \mathcal{L}^l_{\text{vrnn}}(\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}; \Theta^l) + \mathcal{L}^g_{\text{vrnn}}([\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}]; \Theta^g) \\ &+ \alpha \mathcal{L}_{\text{diff}}(\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}; \Theta^l, \Theta^g).\end{aligned} \tag{4}$$

2. **Classifier:** A simple fully connected neural network is used as a classifier that consumes the global latent representations from the last time step T. This network is optimized based on the binary cross-entropy loss ($\mathcal{L}_B$) for all domains as:

$$\mathcal{L}_{\text{clf}}(\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}; \theta_c, \theta^g_e) = \sum_{i=1}^{K} \mathcal{L}_B(C_{\theta_c}(E_g(\mathbf{x}_{\mathcal{D}_i}; \theta^g_e)_T), \mathrm{y}_{\mathcal{D}_i}) \tag{5}$$

where θ_c indicates the classifier parameters.

3. **Discriminator**: To minimize the difference between source domains, we propose to build a domain discriminator for each pair of domains. Therefore, each discriminator $D_{i,j}$ is a fully connected neural network that takes the last time step from global representations $\mathbf{z}^g_{\mathcal{D}_i}$ and $\mathbf{z}^g_{\mathcal{D}_j}$ as input to infer a domain label. This will result in $\binom{K}{2}$ binary classifiers and the total discriminator loss would become:

$$\mathcal{L}_{\text{disc}}(\mathbf{z}^g_{\mathcal{D}_1}, ..., \mathbf{z}^g_{\mathcal{D}_K}; \theta_{\text{disc}}) = \binom{K}{2}^{-1} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \mathcal{L}_B(D_{i,j}(\mathbf{z}^g_{\mathcal{D}_i}, \mathbf{z}^g_{\mathcal{D}_j}); \theta^{i,j}_{\text{disc}}) \tag{6}$$

where $\theta_{\text{disc}}^{i,j}$ indicates the parameters of discriminator $D_{i,j}$. The discriminator's objective is to minimize this loss while the global VRNN tries to maximize this loss. Therefore, the adversarial learning process captures the notion of invariant latent representations between different domains. We have explored multiple other discriminative adversarial learning designs for multi-source problems such as a single discriminator with one vs. rest discrimination or with accumulated gradients [33, 38], but the results show that the pairwise architecture performs the best.

Inspired by Ganin et al. [10] we use the gradient reversal layer (GRL) to effectively combine and optimize all three loss components using backpropagation. GRL can be represented as $\mathcal{R}(x)$ with different forward and backward propagation behavior, where I is the identity matrix and λ is a constant (a specified schedule during training can be used):

$$\mathcal{R}(x) = x; \frac{\partial \mathcal{R}}{\partial x} = -\lambda I \tag{7}$$

The GRL would handle the gradients from the discriminators that should be optimized in the reverse order and the overall optimization becomes:

$$\begin{aligned} \arg \min_{\Theta, \theta_c, \theta_{\text{disc}}} & \mathcal{L}_{\text{sep}}(\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}; \Theta) + \mathcal{L}_{\text{clf}}(\mathbf{x}_{\mathcal{D}_1}, ..., \mathbf{x}_{\mathcal{D}_K}; \theta_c, \theta_e^g) + \\ & \mathcal{L}_{\text{disc}}(\mathcal{R}(\mathbf{z}_{\mathcal{D}_1}^g), ..., \mathcal{R}(\mathbf{z}_{\mathcal{D}_K}^g); \theta_{\text{disc}}) \end{aligned} \tag{8}$$

Equation 8 yields a multi-source domain adaptation framework that can separate domain-specific features from the globally-shared latent representations and adversarially learn an invariant representation between each pair of the source domains. We hypothesize that MS-ADS will address both systematic bias and covariate shift effectively in a multi-source learning environment and builds a unified classifier that is robust across all source domains. We assess this hypothesis through experimentation in the following sections.

3 Experimental Setup

Three EHR Datasets: 210,289 visits of adult patients (i.e. age > 18) admitted to *Christiana Care Health System (CCHS)* in Newark, Delaware (07/2013-12/2015); 106,844 adult patient visits from *Mayo* Clinic in Rochester, Minnesota (07/2013-12/2015); and 53,423 ICU visits of patients admitted to Beth Israel Deaconess Medical Center in Boston, Massachusetts (2001-2012), *MIMIC-III* [15]. *Note that the nature of MIMIC-III data is different from CCHS and Mayo.* To be consistent among all datasets, we define our target population as *suspected of infection*, identified by administration of any anti-infectives, or a positive PCR test result. This definition and the following data pre-processing steps are determined by three leading clinicians with extensive experience on this subject.
Labeling: We adopt the agreement between International Classification of Diseases, Ninth Revision (ICD-9) codes recorded in EHRs, and our expert-defined

rules based on the Third International Consensus Definitions for Sepsis and Septic Shock [32] to achieve the most reliable population across all datasets. Our clinicians identify septic shock at event-level as having received vasopressor(s) or persistent hypotension for more than 1 hour (systolic blood pressure (SBP)<90; or mean arterial pressure<65; or drop in SBP>40 in an 8-hour window).

Sampling: Using the agreement criteria results in 2,963 positive cases in CCHS, 3,499 in Mayo, and 2,459 cases in MIMIC-III. To balance the number of positive and negative cases, we perform a stratified random sampling by 1) maintaining the same underlying age, gender, ethnicity, and length of stay distribution, and 2) having the same level of severity as positive samples. The severity of septic shock visits is identified as the presence of different stages of sepsis in their visits: infection, inflammation, and organ failure as defined by experts.

Aggregation: To align the sampling frequency across all datasets, we use a 30-minutes aggregation window to summarized all records into a single event and missing if none. Our feature set includes 7 vital signs (e.g.: SBP, Temperature), 2 oxygen information (FIO2 and OxygenFlow), and 10 lab results (e.g.: WBC, BUN). To handle the remaining missing values, we first use expert rules to carry forward vital signs (for 8 hours) and lab results (for 24 hours), then we apply the mean imputation along with the missing indicator. Our experiments show that this strategy will help VRNN address such variabilities in data more efficiently.

Prediction Task: Fig. 2 shows our prediction task setup: using EHRs in an observation window to predict whether a patient is going to develop septic shock n hours later; n varies from 24 to 72 hours denoted as *prediction window* and *observation window* is set to be capped at 48 hours as suggested by the leading physicians. All the sequences are aligned by their end time, which is the shock onset for positive visits and a truncated time point for non-shock visits. To prevent the potential bias in models, negative visits are truncated such that they have the same distribution of length as positives. As the prediction window expands, the number of visits remaining in the observation window will drop. For a fair comparison, we sample the same number of positive/negative visits in all domains. This results in 1,315 total visits from each domain for 24 hours early prediction and 620 visits for 72 hours. Therefore, as the number of samples decreases, it is more crucial to integrate different domains to build more robust classifiers.

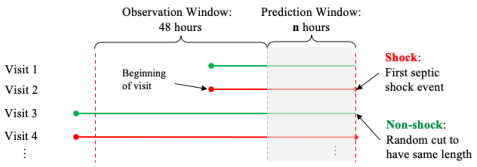


Fig. 2: Septic shock early prediction task

Parameters and Training: As illustrated in Equation 8, there are three sets of parameters: discriminator (θ_{disc}), classifier (θ_c), and VRNN (Θ) parameters optimized through a GRL for adversarial training using NAdam optimizer [34], with learning rate $\alpha_{\text{total}} = 8e^{-4}$. Then the classifier and VRNN models are optimized in an additional step to compete against the gradients from the discriminator,

with learning rates: $\alpha_c = 10e^{-4}$, $\alpha_{\text{VRNN}} = 10e^{-4}$. In every epoch, the order of optimization between the three optimizers is altered from the previous epoch to prevent over-training of a specific network. All the models are implemented in Tensorflow using mini-batch with batch size 32. The same experimental setup is used for all the models with 160 epochs and early stopping. The VRNN's hidden size is set to 30 and the latent size is defined as 50. All the sequences are zero-padded to have the same length and the zero-paddings are masked for reconstruction loss calculation.

Evaluation Metrics: Our evaluation metrics include accuracy, recall, precision, F_1 score, and area under ROC curve (AUC) obtained from 2-fold cross-validation in three independent runs. We mainly use F_1 and AUC as the main metrics as they offer a trade-off between precision, recall, and specificity.

4 Multi-Source DA across the Three Medical Systems

By leveraging data from multiple medical systems while accounting for both *covariate shift* and *systematic bias* across them, we expect MS-ADS would improve the prediction performance across all three systems. Therefore, in this task, the test set is composed of an equal number of visits from CCHS, Mayo, and MIMIC. MS-ADS is compared against six baselines:

1. *VRNN(CCHS)*: a VRNN trained on CCHS only.
2. *VRNN(Mayo)*: a VRNN trained on Mayo only.
3. *VRNN(MIMIC)*: a VRNN trained on MIMIC-III only.
4. *VRNN(Separate)*: will use the individual VRNN trained above to predict the corresponding test data.
5. *VRNN(All)*: a VRNN trained on a combined data from CCHS, Mayo, MIMIC.
6. *Multi VRADA* [28]: a modified version of VRADA to address multi-source DA by changing the domain classifier loss to categorical cross-entropy loss.

24 Hours Early Prediction: Table 1 presents the DA results for 24 hours early prediction on the combined test data (ALL) first (top) and then on test data in each system separately. The top row shows that 1) among the five non-adaptive baselines (1-5), VRNN(All) outperforms all single-domain VRNNs and VRNN(Separate). This result suggests that a more effective classifier can be achieved by leveraging more training samples; 2) By comparing the two multi-source DA models against VRNN(All), we show that VRADA is not able to outperform VRNN(All) while MS-ADS performs robustly and achieves the best performance on all measures except on recall. The highest recall is achieved by VRNN(MIMIC) at a cost of low precision.

The bottom three rows in Table 1 show whether the performance of these models differ across different medical systems (domains). Due to the space limitation, for each domain, we only listed the performance of the corresponding VRNN trained on the same domain compared with the best of the remaining six models. Table 1 shows MS-ADS consistently to be the best model on CCHS and MIMIC but for Mayo, VRNN (Mayo) has a higher F1 score and very close

Table 1: Multi-source DA performance (± std) evaluated on integration of ALL domains and each domain separately for 24 hours early prediction task.

Test Domain	Model	Accuracy	Precision	Recall	F_1 Score	AUC
ALL	1. VRNN(CCHS)	0.735(±0.012)	**0.823**(±0.019)	0.6(±0.048)	0.692(±0.026)	**0.815**(±0.014)
	2. VRNN(Mayo)	0.741(±0.017)	0.753(±0.021)	0.718(±0.053)	0.734(±0.026)	0.81(±0.015)
	3. VRNN(MIMIC)	0.732(±0.016)	0.677(±0.023)	**0.894****(±0.037)	**0.769**(±0.009)	0.814(±0.015)
	4. VRNN(Separate)	**0.803**‡(±0.012)	**0.817**‡(±0.017)	0.781(±0.037)	**0.797**‡(±0.017)	0.864(±0.01)
	5. VRNN(All)	0.795(±0.004)	0.791(±0.014)	**0.801**(±0.022)	0.796(±0.006)	**0.882**‡(±0.003)
	6. Multi VRADA	0.78(±0.029)	0.778(±0.046)	0.766(±0.034)	0.771(±0.021)	0.855(±0.031)
	7. MS-ADS	**0.81****(±0.011)	**0.828****(±0.018)	**0.782**‡(±0.027)	**0.804****(±0.014)	**0.893****(±0.009)
CCHS	1. VRNN(CCHS)	**0.778**(±0.008)	**0.833**(±0.022)	0.698(±0.034)	0.759(±0.014)	0.837(±0.012)
	7. MS-ADS	0.777(±0.012)	0.791(±0.016)	**0.75**(±0.028)	**0.77**(±0.014)	**0.862**(±0.013)
Mayo	2. VRNN(Mayo)	**0.731**(±0.011)	0.732(±0.016)	**0.729**(±0.04)	**0.73**(±0.017)	0.795(±0.004)
	7. MS-ADS	0.73(±0.022)	**0.752**(±0.034)	0.688(±0.046)	0.718(±0.028)	**0.796**(±0.02)
MIMIC	3. VRNN(MIMIC)	0.9(±0.018)	0.888(±0.014)	**0.917**(±0.038)	0.902(±0.02)	0.961(±0.014)
	7. MS-ADS	**0.921**(±0.015)	**0.935**(±0.018)	0.907(±0.018)	**0.921**(±0.016)	**0.974**(±0.005)

· The *best* and the *second best* models are labeled with ** and ‡, respectively.

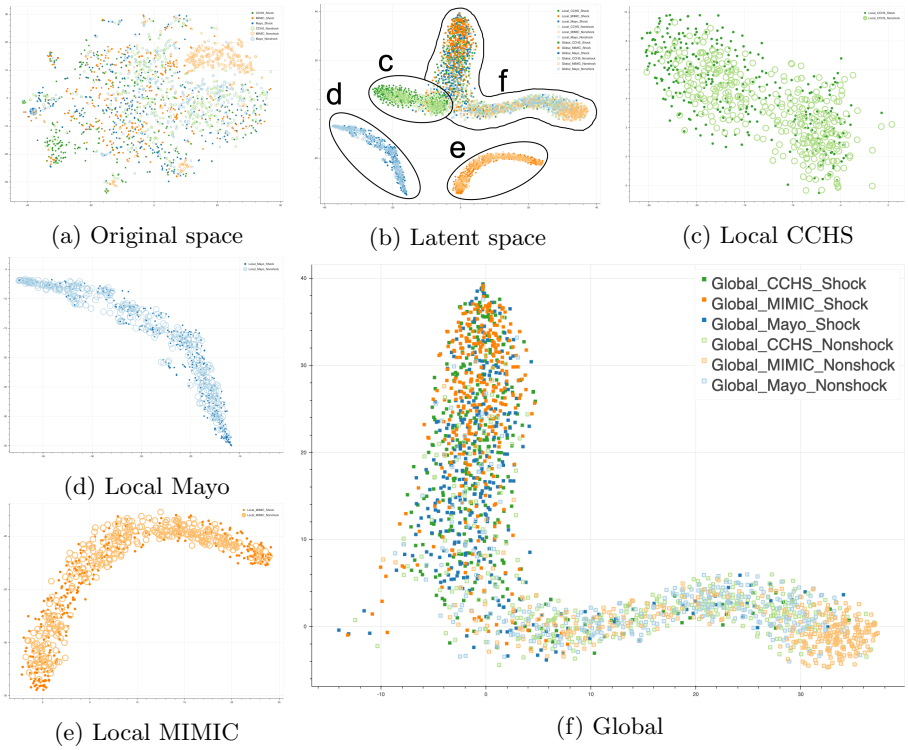
Table 2: Multi-source DA performance evaluated for 24-72 hours early prediction.

Model	Accuracy	Precision	Recall	F_1 Score	AUC
1. VRNN(CCHS)	0.674(±0.04)	0.734(±0.055)	0.559(±0.052)	0.63(±0.041)	0.728(±0.046)
2. VRNN(Mayo)	0.66(±0.031)	0.678(±0.037)	0.624(±0.075)	0.645(±0.044)	0.712(±0.032)
3. VRNN(MIMIC)	0.689(±0.014)	0.633(±0.015)	**0.909****(±0.029)	0.745(±0.009)	0.759(±0.016)
4. VRNN(Separate)	**0.755**‡(±0.025)	**0.775**‡(±0.031)	0.719(±0.053)	0.742(±0.031)	0.804(±0.023)
5. VRNN(All)	0.749(±0.012)	0.757(±0.021)	0.743(±0.051)	**0.747**‡(±0.019)	**0.835**‡(±0.01)
6. 3-d VRADA	0.746(±0.021)	0.751(±0.034)	0.739(±0.043)	0.743(±0.022)	0.829(±0.024)
7. MS-ADS	**0.771****(±0.016)	**0.782****(±0.016)	**0.75**‡(±0.033)	**0.765****(±0.02)	**0.85****(±0.012)

· The *best* and the *second best* models are labeled with ** and ‡, respectively.

AUC score to MS-ADS. Additionally, MIMIC data has extremely good results while the performance on Mayo is the worst. Such results suggest that early sepsis shock prediction is relatively trivial for MIMIC dataset probably because MIMIC only includes ICU visits. Therefore, in the following, we mainly focus on generalization to CCHS and Mayo only.

Varying 24-72 Hours Early Prediction: Table 2 shows the average performance by varying the prediction window from 24 to 72 hours, with every 12 hours interval. For each prediction window, our test set has an equal number of visits from each domain. Table 2 shows MS-ADS significantly outperforms all other baselines including VRADA and VRNN(All) for *all* metrics except recall. VRNN(MIMIC) performs with the highest recall across all domains, but at the cost of very low precision. Comparing VRNN(Separate) with MS-ADS shows a ∼3% improvement for recall and ∼4.5% improvement for AUC across all three domains. This result demonstrates that by integrating EHRs across medical systems, MS-ADS can address insufficient labeled data problems and by adopting an effective domain adaptation architecture, MS-ADS can address both systematic bias and covariate shift across medical systems.



(a) Original space (b) Latent space (c) Local CCHS

(d) Local Mayo

(e) Local MIMIC (f) Global

Fig. 3: Visit-level t-SNE visualization of (a) Original vs. (b) Latent space of MS-ADS. (c)-(e) show domain-specific representations while (f) illustrates the globally-shared representation. Solid dots represent septic shock visits.

Visit-level Visual Investigation. Fig. 3 illustrates t-SNE visualization of the original and latent representation of all visits for 24 hours early prediction. In all graphs, different colors represent different medical systems and solid and hollow points represent shock and non-shock visits, respectively. Fig. 3a illustrates the original space and 3b shows that the latent space generated by MS-ADS can separate the three local representations (c), (d), (e) (enlarged in Fig. 3c-3e) from the global ones (f) (enlarged in Fig. 3f). Fig. 3c-3e suggests that MS-ADS can address systematic bias effectively while Fig. 3f shows that in the global space, samples from different domains are close together and mostly aligned and mixed. This shows that MS-ADS can address covariate shift effectively as well.
Event-level Visual Investigation. Further, we look at the original and global latent space at the event level to validate if covariate shift is addressed by MS-ADS along the temporal axis. We select two similar septic shock visits across CCHS and Mayo such that both develop inflammation and multiple organ failure symptoms within the observation window. Fig. 4 shows these two traces (CCHS (red) and Mayo (blue)) in the original and global latent spaces. Despite their similarity, their progression deviates in the original space while in the latent

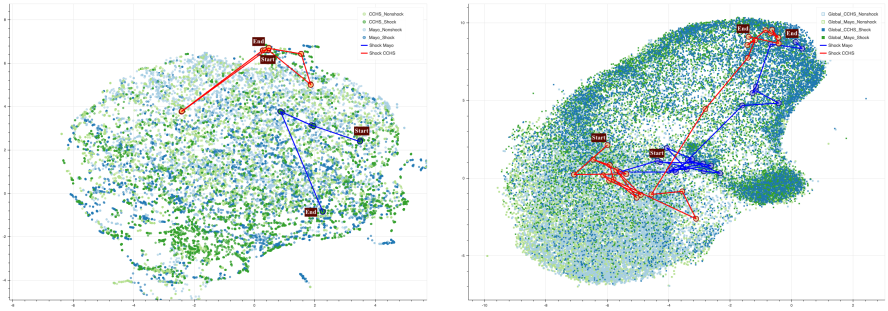


Fig. 4: Event-level t-SNE visualization of Original (left) vs. Global Latent (right) representation of MS-ADS on CCHS and Mayo. Red (CCHS) and Blue (Mayo) traces show sepsis progression of two similar patients.

Table 3: DG performance to unseen target domains for 24 hours early prediction.

Source	Unseen Target	Model	Accuracy	Precision	Recall	F_1 Score	AUC
MIMIC + Mayo	CCHS	VRNN	0.747(±0.02)	0.739‡(±0.041)	↑0.77(±0.028)	0.753(±0.011)	0.831(±0.015)
MIMIC + Mayo		VRADA	0.763‡(±0.021)	0.739(±0.043)	↑**0.826**(±0.048)	↑**0.778**(±0.013)	↑0.846‡(±0.017)
MIMIC + Mayo		MS-ADS	**0.764**(±0.017)	**0.74**(±0.027)	↑0.813‡(±0.023)	↑0.774‡(±0.013)	↑**0.851**(±0.008)
CCHS		VRNN	0.778(±0.008)	0.833(±0.022)	0.698(±0.034)	0.759(±0.014)	0.837(±0.012)
CCHS + MIMIC	Mayo	VRNN	0.698(±0.014)	0.725‡(±0.036)	**0.644**(±0.066)	0.678‡(±0.028)	0.763‡(±0.018)
CCHS + MIMIC		VRADA	0.678‡(±0.039)	0.702(±0.035)	0.608(±0.072)	0.65(±0.055)	0.739(±0.039)
CCHS + MIMIC		MS-ADS	↑**0.73**(±0.012)	↑**0.796**(±0.025)	0.62‡(±0.011)	**0.697**(±0.01)	↑**0.8**(±0.014)
Mayo		VRNN	0.731(±0.011)	0.732(±0.016)	0.729(±0.04)	0.73(±0.017)	0.795(±0.004)

· In each block, the best performance is in **bold**; Models that outperform the gold standard (bottom) are labeled with ↑.

representation their temporal progression is aligned. This further demonstrates the effectiveness of MS-ADS in addressing covariate shift at a temporal level.

5 Domain Generalization to Unseen Medical System

In the second task, MS-ADS is trained on EHRs from two medical systems and evaluated on an unseen target system: CCHS or Mayo. MS-ADS is compared against two baselines: a VRNN trained on the combination of two source domains and the original VRADA applied for DA across the two domains. Table 3 presents the generalization performance of all three models for 24 hours early prediction. Table 3 shows MS-ADS outperforms the two baselines for most metrics in both target domains. Finally, we also compared them against the gold standard supervised VRNN model trained on the target domain (last row in each section), Table 3 shows that MS-ADS outperforms the supervised VRNN for AUC metric in both target domains.

Table 4: VRNN performance trained and tested on different racial groups across medical systems for 24 hours early prediction.

Train Domain	Test Domain	Accuracy	Precision	Recall	F_1 Score	AUC
CCHS(WA)	CCHS(WA)	0.888(±0.014)	0.869(±0.025)	0.916(±0.017)	0.891(±0.013)	0.956(±0.008)
	CCHS(AA)	0.885(±0.011)	0.874(±0.018)	0.9(±0.004)	0.887(±0.01)	0.946(±0.007)
Mayo(WA)	Mayo(WA)	0.841(±0.038)	0.83(±0.043)	0.86(±0.039)	0.844(±0.036)	0.909(±0.03)
	Mayo(AA)	0.809(±0.025)	0.821(±0.042)	0.813(±0.038)	0.816(±0.024)	0.847(±0.038)
CCHS(AA)	Mayo(AA)	0.715(±0.031)	0.751(±0.031)	0.68(±0.061)	0.712(±0.038)	0.811(±0.037)
CCHS(WA+AA)	Mayo(AA)	0.792(±0.032)	0.834(±0.048)	0.753(±0.054)	0.79(±0.034)	0.872(±0.021)

Table 5: Generalization performance to unseen African American (AA) patients in Mayo using 2-domains and 3-domains.

Train Domains	Model	Accuracy	Precision	Recall	F_1 Score	AUC
CCHS(WA + AA), Mayo(WA)	VRNN(All)	0.844(±0.031)	0.895(±0.012)	0.793(±0.075)	0.839(±0.04)	0.913(±0.01)
	2-d VRADA	0.87(±0.025)	0.87(±0.034)	0.873(±0.063)	0.87(±0.029)	0.922(±0.021)
	2-d MS-ADS	0.854(±0.017)	**0.901**(±0.033)	0.813(±0.068)	0.852(±0.026)	0.914(±0.012)
CCHS(WA), CCHS(AA), Mayo(WA)	3-d VRADA	0.847(±0.016)	0.861(±0.033)	0.847(±0.049)	0.852(±0.017)	0.917(±0.034)
	3-d MS-ADS	**0.87**(±0.012)	0.876(±0.007)	**0.876**(±0.035)	**0.875**(±0.014)	**0.947**(±0.005)

· The best overall performance is in **bold**..

6 Unseen Racial Group across Medical Systems

In this task, we focus on two racial groups: White American (WA) and African American (AA). Table 4 compares the performance of models that are trained and tested on different racial groups across medical systems. Table 4 shows that while the model trained on CCHS(WA) performs equally well on CCHS(WA) and CCHS(AA); the model trained on Mayo(WA) performs much better on Mayo(WA) than Mayo(AA). This is probably because the percentages of WA and AA are more balanced than those of Mayo: 71% vs. 22.5% in CCHS while 91% vs. 3% in Mayo. The last block in Table 4 shows transfer across medical systems. The model trained on CCHS(AA) does not perform well on Mayo(AA) probably due to systematic bias across medical systems while adding CCHS(WA) to the training population can help predictions on Mayo(AA) probably because of more training data. As a result, our training domain settings will involve WA in Mayo and WA and AA in CCHS.

Table 5 compares the generalization performance on Mayo(AA) by using two training domains: CCHS (WA+AA) and Mayo (WA) (upper) vs. three training domains: CCHS (WA), CCHS(AA), and Mayo (WA) (bottom). For the two-domain generalization, MS-ADS is compared against the best non-DA baseline: VRNN(All) and VRADA. Table 5 shows that both VRADA and MS-ADS outperform the VRNN(All) and VRADA achieves the best F1 and AUC. When we conduct the same task by using three domains, the bottom block in Table 5 shows that the performance of VRADA suffered while the performance of

MS-ADS improved. Indeed, Table 5 shows that the F1 and AUC of 3-domain MS-ADS on Mayo(AA) are 0.875 and 0.947, catching up with the other three racial groups across the two systems. We argue the effectiveness of 3-domain MS-ADS over 2-domain MS-ADS is probably because the former can leverage 1) Mayo(WA) (same system different race), 2) CCHS(AA) (same race, different system), 3) the DA mechanism learned from unifying AA and WA in CCHS (addressing covariate shift within the same system), and 4) the DA learned from unifying WA between CCHS and Mayo (addressing systematic bias in the same racial group).

7 Related Work

Septic Shock Early Prediction: A variety of machine learning models have been developed to predict septic shock several hours before the onset. Among traditional approaches, multivariate logistic regression and survival analysis models have been proposed for early detection [14,31]. Moreover, sequential pattern mining approaches have shown to be effective for early prediction of septic shock while producing explainable patterns [12, 17]. Recently, various deep learning-based approaches have been proposed, especially variations of recurrent neural networks such as LSTM, and they have shown promising power in predicting septic shock several hours before the onset [22, 40, 41]. Despite the great power of LSTM models, they are not designed to address the high missing rate in EHR [18]. Variational recurrent neural network (VRNN) [5] is recently proposed to model complex temporal and conditional dependencies in sequential data and account for variabilities, like missing data, and has shown great promise [4,26,39].

Multi-source Domain Adaptation: The majority of existing DA work either addresses this problem by generating an invariant feature space for all pairs of source-target distributions [27, 43] or constructs the target distribution as a weighted combination of source distributions [23, 37]. For example, VRADA is a VRNN-based DA that has been applied to EHRs from different groups of patients and it has shown significant improvement in creating domain-invariant representations using adversarial learning [28]. In this work, we further expanded VRADA's architecture to address multi-source DA problems and use it as a baseline. These studies treat multiple medical systems as "source" domains to improve the prediction performance in a specific "target" medical system. By treating all domains equally, a group of DA studies aim at learning a unified minimal risk model from multiple domains [6, 8, 30]. While the majority of such DA research is conducted in computer vision and text classification, a few studies have proposed DA approaches to integrate EHRs across multiple medical systems and improve prediction in a target domain by addressing covariate shift and feature mismatch [36,38]. All existing approaches have shown great power in accounting for the covariate shift but not domain-specific characteristics or systematic bias that *should not* be unified across domains. Our MS-ADS model is capable of integrating multiple medical systems to build a robust unified model

that improves prediction across *all* systems while accounting for the covariate shift and systematic bias simultaneously but differently.

Domain Generalization aims to learn a model from an arbitrary number of source domains such that it can generalize to previously unseen target domains [9, 13, 29]. One class of approaches proposes to learn domain-invariant representations by minimizing domain mismatch across source domains, similar to DA approaches [11, 21, 25]. For example, Motiian et al. propose a unified DA and DG model exploiting Siamese architecture using a contrasting loss to minimize the distance between samples from the same class but in different domains [24]. Another type of DG method utilizes meta-learning techniques to synthesize domain shift and directly learn and optimize for generalization task [3, 7, 20]. Despite the critical application of DG in clinical deployment, especially in presence of limited data, this problem is still under-explored.

Further, to close the gap between performances among different groups of patients, previous studies have explored DA approaches to account for the covariate shift between groups within a medical system [2, 28]. For example, Zhang et al. proposed a time-aware adversarial LSTM network to transfer knowledge across different racial, age, and gender groups and improve prediction for minority groups [42]. As far as we know, this study is the first that investigates DG to simultaneously address the covariate shift across different racial groups and systematic bias across medical systems to generalize robustly and improve prediction among minority groups.

8 Conclusion

In this work, we propose a multi-source adversarial domain separation (MS-ADS) framework that unifies domain adaptation (DA) and domain generalization (DG) by accounting for systematic bias across medical systems and covariate shift among different patient groups to achieve a robust generalization. In specific, MS-ADS separates the global representation of each domain from the local ones to address systematic bias and leverage a multi-domain discriminator with Gradient Reversal Layer (GRL) to account for the covariate shift. We evaluate MS-ADS in three tasks for septic shock early prediction using EHR from three medical systems. First, on a task of DA across three medical systems, we show that the effective adaptation under MS-ADS leads to performance improvement in all three domains. Second, on a task of DG to an unseen medical system, we demonstrate the generalization power brought by MS-ADS architecture by comparing and showing its robustness against a gold standard supervised model on the target domain. Lastly, on a task of generalization to unseen racial groups across the medical system, we show that unifying DA and DG MS-ADS can significantly improve prediction among minority racial groups.

Acknowledgments This research was supported by the NSF Grants #2013502, #1726550, #1651909, and #1522107.

References

1. Agniel, D., et al.: Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ **361** (2018)
2. Alves, T., Laender, A., Veloso, A., Ziviani, N.: Dynamic prediction of icu mortality risk using domain adaptation. In: IEEE Big Data. pp. 1328–1336. IEEE (2018)
3. Balaji, Y., Sankaranarayanan, S., Chellappa, R.: Metareg: Towards domain generalization using meta-regularization. In: NeurIPS. pp. 998–1008 (2018)
4. Chien, J.T., Kuo, K.T., et al.: Variational recurrent neural networks for speech separation. In: Interspeech, VOLS 1-6: Situated Interaction. pp. 1193–1197 (2017)
5. Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A.C., Bengio, Y.: A recurrent latent variable model for sequential data. In: NeurIPS. pp. 2980–2988 (2015)
6. Ding, X., Shi, Q., Cai, B., Liu, T., Zhao, Y., Ye, Q.: Learning multi-domain adversarial neural networks for text classification. IEEE Access **7**, 40323–40332 (2019)
7. Dou, Q., Castro, D.C., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. arXiv:1910.13580 (2019)
8. Dredze, M., Kulesza, A., Crammer, K.: Multi-domain learning by confidence-weighted parameter combination. Machine Learning **79**(1-2), 123–149 (2010)
9. Du, Y., Xu, J., Xiong, H., Qiu, Q., Zhen, X., Snoek, C.G., Shao, L.: Learning to learn with variational information bottleneck for domain generalization. In: European Conference on Computer Vision. pp. 200–216. Springer (2020)
10. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. arXiv:1409.7495 (2014)
11. Ghifary, M., Bastiaan Kleijn, W., Zhang, M., Balduzzi, D.: Domain generalization for object recognition with multi-task autoencoders. In: CVPR
12. Ghosh, S., Li, J., Cao, L., Ramamohanarao, K.: Septic shock prediction for icu patients via coupled hmm walking on sequential contrast patterns. JBI **66**
13. Gong, R., Li, W., Chen, Y., Gool, L.V.: Dlow: Domain flow for adaptation and generalization. In: CVPR. pp. 2477–2486 (2019)
14. Henry, K.E., et. al: A targeted real-time early warning score (trewscore) for septic shock. Science translational medicine **7**(299) (2015)
15. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**, 160035 (2016)
16. Khoshnevisan, F., Chi, M.: An adversarial domain separation framework for septic shock early prediction across ehr systems. arXiv:2010.13952 (2020)
17. Khoshnevisan, F., Ivy, J., Capan, M., Arnold, R., Huddleston, J., Chi, M.: Recent temporal pattern mining for septic shock early prediction. In: IEEE ICHI. pp. 229–240. IEEE (2018)
18. Kim, Y.J., Chi, M.: Temporal belief memory: Imputing missing data during rnn training. In: IJCAI (2018)
19. Kumar, A., Roberts, D., Wood, K.E., Light, B., Parrillo, J.E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., et al.: Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine **34**(6), 1589–1596 (2006)
20. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Learning to generalize: Meta-learning for domain generalization. arXiv:1710.03463 (2017)
21. Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: ECCV

22. Lin, C., Zhangy, Y., Ivy, J., Capan, M., Arnold, R., Huddleston, J.M., Chi, M.: Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-lstm. In: IEEE ICHI. pp. 219–228. IEEE (2018)
23. Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain adaptation with multiple sources. In: NeurIPS. pp. 1041–1048 (2009)
24. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: CVPR. pp. 5715–5725 (2017)
25. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: ICML. pp. 10–18 (2013)
26. Mulyadi, A.W., Jun, E., Suk, H.I.: Uncertainty-aware variational-recurrent imputation network for clinical time series. arXiv:2003.00662 (2020)
27. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: CVPR. pp. 1406–1415 (2019)
28. Purushotham, S., Carvalho, W., Nilanon, T., Liu, Y.: Variational recurrent adversarial deep domain adaptation (2016)
29. Qiao, F., Zhao, L., Peng, X.: Learning to learn single domain generalization. In: Proceedings of the CVPR. pp. 12556–12565 (2020)
30. Schoenauer-Sebag, A., Heinrich, L., Schoenauer, M., Sebag, M., Wu, L.F., Altschuler, S.J.: Multi-domain adversarial learning. arXiv:1903.09239 (2019)
31. Shavdia, D.: Septic shock: Providing early warnings through multivariate logistic regression models. Ph.D. thesis, Massachusetts Institute of Technology (2007)
32. Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.D., Coopersmith, C.M., et al.: The third international consensus definitions for sepsis and septic shock (sepsis-3). Jama **315**(8), 801–810 (2016)
33. Tasar, O., Tarabalka, Y., Giros, A., Alliez, P., Clerc, S.: Standardgan: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization. In: CVPR Workshops. pp. 192–193 (2020)
34. Tato, A., Nkambou, R.: Improving adam optimizer (2018)
35. Tintinalli, J., J, S., O, J.M., D, C., R, C., G, M.: Tintinallis emergency medicine A comprehensive study guide, chap. 146: Septic Shock, pp. 1003–1014. McGraw-Hill Education, 7 edn. (2011)
36. Wiens, J., Guttag, J., Horvitz, E.: A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. Journal of the American Medical Informatics Association **21**(4), 699–706 (2014)
37. Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L.: Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In: CVPR
38. Yoon, J., Jordon, J., van der Schaar, M.: Radialgan: Leveraging multiple datasets to improve target-specific predictive models using generative adversarial networks. arXiv:1802.06403 (2018)
39. Zhang, S., Xie, P., Wang, D., Xing, E.P.: Medical diagnosis from laboratory tests by combining generative and discriminative learning. arXiv:1711.04329 (2017)
40. Zhang, Y., Lin, C., Chi, M., Ivy, J., Capan, M., Huddleston, J.M.: Lstm for septic shock: Adding unreliable labels to reliable predictions. In: IEEE Big Data. pp. 1233–1242. IEEE (2017)
41. Zhang, Y., Yang, X., Ivy, J., Chi, M.: Attain: attention-based time-aware lstm networks for disease progression modeling. In: IJCAI. pp. 10–16 (2019)
42. Zhang, Y., Yang, X., Ivy, J., Chi, M.: Time-aware adversarial networks for adapting disease progression modeling. In: IEEE ICHI. pp. 1–11. IEEE (2019)
43. Zhao, H., Zhang, S., Wu, G., Moura, J.M., Costeira, J.P., Gordon, G.J.: Adversarial multiple source domain adaptation. In: NeurIPS. pp. 8559–8570 (2018)