Semi-supervised Meta-learning for Multi-source Heterogeneity in Time-series Data

Lida Zhang LIDAZHANG@TAMU.EDU
Department of Computer Science & Engineering
Texas A&M University
College Station, Texas, USA

Bobak J. Mortazavi BOBAKM@EMAIL.EDU
Department of Computer Science & Engineering
Texas A&M University
College Station, Texas, USA

Abstract

Real-world time-series data is riddled with heterogeneity that is often present across a number of dataset dimensions: features, labels, and time-varying factors. The heterogeneity in time-series data may be raised by introducing new features, missing data, and domain shifts in the feature dimension, and the difficulty of collecting promising ground truth results in label uncertainty. In addition, the variation on the time manner further aggravates the complexity of data heterogeneity, since the features and labels may change on the same sequence of data over time. Many machine learning techniques have been proposed to address the data heterogeneity, including transfer learning, meta-learning, semi-supervised learning, recurrent networks, etc. However, each of these techniques is limited to one type of heterogeneity. In this study, we seek to create adaptable models for the multi-source heterogeneity in time-series data. We propose a semi-supervised-based meta-learning (SSML) with an adversarial training mechanism simultaneously addressing the heterogeneous features and labeling uncertainty, a time domain variation (TDV) framework to apply SSML and transfer learning for the third level of data heterogeneity. We test our models on two medical datasets, PhysioNet Challenge 2012 and MIMIC-III ICU dataset, and improve over all benchmark models. Our code is available at https://github.com/lidazhang/ssml-time-series-heterogeneity.git.

1. Introduction

Data heterogeneity is a natural attribute of many real-world applications and datasets in the time-series domain. Heterogeneity occurs frequently and can be complex across several dimensions: features, labels, and the time-varying nature of data. On the feature dimension, heterogeneity may come from the development of new sensors Wilson et al. (2020); Javeed et al. (2021), missing data Lipton et al. (2016), or different setups for data collection Macadam et al. (2019). The difficulty in observing ground truth Pereira and Silveira (2019); Yu and Sano (2022) and obtaining inconsistent user feedback Plötz and Guan (2018) may result in label uncertainty. In the time domain, the variation present in changing health conditions of patients Zhang et al. (2021), changes in seasons Kafy et al. (2021), cycles in the economy Brynjolfsson et al. (2018), or even the spreading of disease in

a pandemic can all lead to vastly different data representations and ranges. The different types of heterogeneity can occur not only individually but also simultaneously, and thus result in a problem of multi-source heterogeneity in time-series modeling and applications.

Machine learning (ML) techniques have been developed to address the challenges of data heterogeneity. However, they are usually limited to a certain type of heterogeneity. Often, the heterogeneous features are handled by training individual models for each subset of data Zhang et al. (2020); Chen et al. (2020), but this requires onerous training of multiple models and may result in poorly performing models if the same subsets have very limited data. Transfer learning and meta-learning are approaches used to aid this limitation across models Gupta et al. (2020); Desautels et al. (2017), and can significantly reduce the training time while maintaining performance. However, these techniques are not sufficient for multi-source heterogeneity. Semi-supervised learning algorithms including active learning are proposed to address the labeling challenge, and are applied in the time-series field Pereira and Silveira (2019); Yu and Sano (2022); Jiang et al. (2021); Gweon and Yu (2021), but lack consideration of various types or frequencies of data. Recurrent networks and attention-based transformer model Vaswani et al. (2017) are used to capture the time domain variation, but a generalized version of these models is static and restrictive across types of data Zhang et al. (2020). Methods that address the simultaneous multi-source heterogeneity occurring in time-series data are needed.

We seek to solve the multi-source data heterogeneity challenge in applications in medicine, one of the most complex time-series data types with all three types of data heterogeneity. First, medical data contain thousands of different observations, laboratory tests, medications, etc. from hospitals Johnson et al. (2016), and the frequency (and category) of these measurements comes from doctors' examinations and implies the potential health condition. Learning from the similar frequency of medical data can lead the model to be more specific for a type of patients, so that risk prediction tasks can be improved and aid in up-to-date clinical decision-making. Second, as a real-world time series dataset, medical data also has the challenge of obtaining labels. For example, the diagnosis from doctors is time-sensitive, and the development of patients' health conditions can cause changes in the labels. The development of patients' health conditions also raises the third heterogeneity, time domain variation. In addition, this variation can also be caused by other factors, such as receiving treatments in the hospital Webb et al. (2020), hospital transfer Cheng et al. (2020), ICU admission and release Raita et al. (2019), etc.

Facing these challenges, the goal of this paper is to build adaptive models to address the multi-source heterogeneity that can occur simultaneously in time-series data. We propose a semi-supervised meta-learning algorithm for the heterogeneous features and uncertainty in labels. Meta-learning, in the manner of few-shot learning, addresses the potential data limitation in certain types of feature space and the demand for fast adaptation in the future. A discriminator is introduced for adversarial training to improve the model generalization. Regarding the variation over time, we propose a time domain variation (TDV) framework applying transfer learning and our SSML. Our approach is a new connection between meta-learning, transfer learning, and semi-supervised learning. We test our approaches on two real-world medical datasets, PhysioNet Challenge 2012 and the MIMIC-III ICU dataset. To the best of our knowledge, we are the first the address this complex real-world simultaneous multi-source heterogeneity of feature space, time domain variation, and label uncertainty

Table 1: Overview of the ML techniques addressing various types of time-series heterogeneity

Algorithm	Heterogeneous features	Label uncertainty	Time domain variation
Recurrent network			✓
Transformer			✓
Transfer learning	✓		✓
Meta-learning	✓		
Semi-supervised learning		✓	
SSML (Ours)	✓	✓	
SSML-TDV (Ours)	✓	✓	✓

on time-series data (Table 1). Our proposed model is flexible to address all or part of the heterogeneity problem, and is also adaptive for future model update demands.

Generalizable Insights about Machine Learning in the Context of Healthcare

Heterogeneity is a common problem in real-world applications that impedes the development of modeling. Our SSML and SSML-TDV provide solutions to the complex multi-source heterogeneity situations in time-series data. Clinicians will benefit from the outcomes in this paper by obtaining models that are adaptable to the heterogeneous features, the uncertain labels, and the time domain variation across their patient population. The two versions also provide the flexibility of choosing appropriate models for different heterogeneity problems. Additionally, patients that are traditionally dissimilar from the average patient, perhaps ill-represented by models that work well on the general population, may see an advantage in this case.

2. Related Work

Meta-Learning. Meta-learning is designed to extract information about the optimization process on a few samples for various learning tasks Finn et al. (2017); Grant et al. (2018); Rajeswaran et al. (2019). Finn et al. Finn et al. (2017) propose MAML, which optimizes the model initialization as the meta-learner, and is widely applied to a large number of healthcare applications Hu et al. (2018); Banluesombatkul et al. (2020); Naren et al. (2021). Zhang et al. Zhang et al. (2019) apply MAML on EHR data to predict clinical risk for patients, and Zhang et al. Zhang et al. (2021) propose DynEHR based on MAML to model for the various duration of EHR data. Ren et al. Ren et al. (2018) first introduce semi-supervised learning to the few-shot learning algorithm Prototypical Network Snell et al. (2017); however, refining the prototype of each class without differentiating the domains cannot achieve the goal of building adaptive models for various EHR sequences. Our proposed model is also motivated by MAML and we compare SSML with MAML for heterogeneous EHR data modeling.

Semi-supervised Learning. The goal of semi-supervised learning is to make use of unlabeled data. Self-training uses the model prediction of unlabeled data as the produced label and is applied in many applications Rosenberg et al. (2005); Zou et al. (2019); Xie

et al. (2020); He and Zhou (2011). Pseudo-labeling further converts the confident prediction to hard labels Lee et al. (2013), but this may not be stable He and Zhou (2011); Arazo et al. (2020). Consistency regularization Bachman et al. (2014); Rasmus et al. (2015); Sajjadi et al. (2016) is then introduced to self-training Xie et al. (2020). Sohn et al. Sohn et al. (2020) propose FixMatch using augmentation French et al. (2017) as a consistency regularization into pseudo-labeling. Meta-learning is then applied in FixMatch as a new semi-supervised learning approach Wang et al. (2020); Xiao et al. (2021). However, meta-learning here is only applied between the labeled and unlabeled data of the same learning task, and these two papers, which are semi-supervised learning algorithms, cannot be used on multiple learning tasks and datasets, nor do they serve as an adaptive model for our data heterogeneity problem. Therefore, we do not directly compare them.

EHR clinical analysis. EHR has been studied in both medicine and machine learning since its wide use in hospitals. Harutyunyan et al. Harutyunyan et al. (2019) propose an LSTM-based multi-task model for clinical prediction with EHR variables, and Xu et al. Xu et al. (2018) introduce waveform data in their model. Transformer Vaswani et al. (2017) is first used in the EHR model as a replacement of LSTM by Song et al. Song et al. (2019). However, none of these works have considered the heterogeneity in EHR data. Shukla Shukla and Marlin (2021) addresses the irregularly-sampled data by mapping it to a regular space, but there is no specified analysis about each homogeneous set in the heterogeneous in EHRs. Zhang et al. Zhang et al. (2021) propose DynEHR as an adaptive model for EHRs, but the method is not flexible enough to be applied in other types of data heterogeneity other than the temporal source.

3. Methods

In this section, we present our solution for the multi-source heterogeneity in time-series data. We define the heterogeneous features challenge as a multi-domain problem, and each domain includes homogeneous examples. We use the meta-learning framework as a fast adaptive model for each domain, and propose the semi-supervised meta-learning algorithm (SSML) with adversarial training for the label uncertainty in the multi-domain setting, and SSML is then applied with transfer learning in a time domain variation (TDV) framework for the third level of heterogeneity.

3.1. Problem Setup

In this study, a set of domains represents the varied, heterogeneous feature space for the learning tasks. Each domain includes sequences with similar feature frequency distribution. Let $\mathcal{D}$ denote all domains, and $D_i \in \mathcal{D}$ represents the i-th domain. Let $\mathcal{X}_i$ and $\mathcal{U}_i$ denote the labeled and unlabeled data in domain D_i, and $\mathcal{Y}_i$ is the corresponding label of $\mathcal{X}_i$, then domain D_i has $D_i = \{\{\mathcal{X}_i, \mathcal{Y}_i\}, \mathcal{U}_i\}$.

Let $\mathcal{S}$ be a set of time-series data. Given a sequence example $s = x^{(1:T)}$ from $\mathcal{S}$ ($s \in \mathcal{S}$) containing T time stamps, and $x^{(t)}$ represents the feature vector at time point t ($1 \leq t \leq T$). Assume the time domain variation occurs on s (e.g., complication happening to a patient),

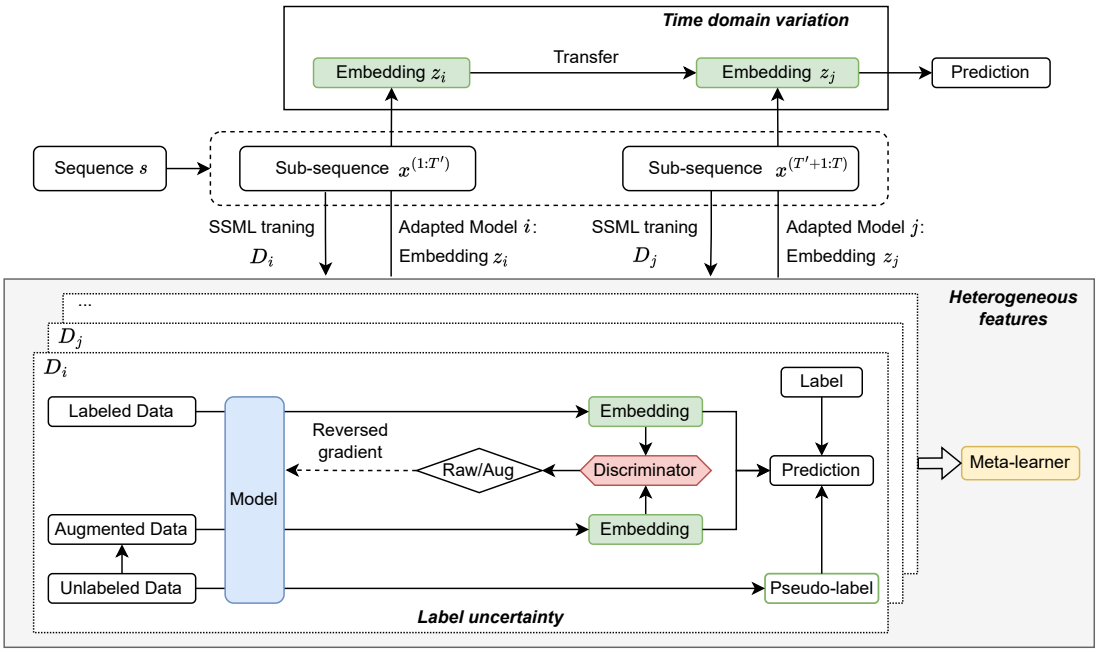


Figure 1: The framework of SSML-TDV for multi-source time-series heterogeneity. (***Bottom***) Semi-supervised meta-learning (SSML) with adversarial training for heterogeneous features and label uncertainty. (***Top***) The SSML-based time domain variation framework (SSML-TDV). Each sequence participates in SSML training, and applies the trained SSML with transfer learning for predictions.

which splits the sequence into sub-sequences at time point T':

$$s = \underbrace{x^{(1)}, ..., x^{(T')}}_{D_i}, \underbrace{x^{(T'+1)}, ..., x^{(T)}}_{D_j} \tag{1}$$

where sub-sequences $s^1 = x^{(1:T')}$ belongs to domain D_i and $s^2 = x^{(T'+1:T)}$ belongs to D_j. The time domain variation on long sequences also causes label uncertainty among the sub-sequences, such that $s^1 \in \mathcal{U}_i$ and $s^2 \in \mathcal{X}_j$ (the uncertain label may also come from unlabeled data). The goal of our work is to build adaptive models under a multi-domain setting respecting the potential shifts among different domains within each sequence s and the uncertain label problem.

3.2. Semi-supervised Meta-learning

In this section, we first address two heterogeneity problems: heterogeneous feature space and label uncertainty. Figure 1 ***Bottom*** box shows our proposed solution.

Supervised meta-learning

An underlying challenge of the heterogeneous feature space is the potential limitation of having sufficient training examples in each domain. In addition, standard supervised learning is also limited to future demands of model adaption in practice, for example, when

there is a new disease discovered but very limited patient examples are collected. Therefore, we address the data heterogeneity problem in a meta-learning setting.

Given a model $\mathcal{F}$ consisting a feature extractor $\mathcal{F}_\theta$ and a predictor $\mathcal{F}_\eta$, where θ and η represent their parameters correspondingly. The goal of meta-learning is to learn the optimization process of several domains and optimize the model initialization θ and η in $\mathcal{F}$, so that model $\mathcal{F}_{\theta;\eta}$ can be optimized to be very fast adapted to $\mathcal{F}_{\theta_k;\eta_k}$ for any domain D_k.

For a domain D_i from training domains $\mathcal{D}$, the model $\mathcal{F}_{\theta_i;\eta_i}$ is initialized with θ and η. Given the labeled data $\{x_i, y_i\} \subseteq \{\mathcal{X}_i, \mathcal{Y}_i\}$ in D_i, model $\mathcal{F}_{\theta_i;\eta_i}$ can be trained through supervised learning with cost

$$\mathcal{L}^l_{D_i}(\theta_i, \eta_i) = \mathcal{L}(\mathcal{F}_{\theta_i;\eta_i}(x_i), y_i), \tag{2}$$

where $\mathcal{L}$ represents the cost function (mean-squared error for a regression task or cross-entropy for a classification task). After N steps of training with gradient descent, $\mathcal{F}_{\theta_i;\eta_i}$ becomes the adapted model $\mathcal{F}_{\bar{\theta}_i;\bar{\eta}_i}$:

$$\bar{\theta}_i = \theta_i - \alpha \frac{\partial \mathcal{L}^l_{D_i}(\theta_i, \eta_i)}{\partial \theta_i}, \qquad \bar{\eta}_i = \eta_i - \alpha \frac{\partial \mathcal{L}^l_{D_i}(\theta_i, \eta_i)}{\partial \eta_i}, \tag{3}$$

where α is the step size.

For the purpose of fast adapting to any domain, model $\mathcal{F}_{\theta;\eta}$ needs to learn from several domains. In each training episode, we randomly generate a set of domains $D \subseteq \mathcal{D}$, and train their adapted model from Equation 2 and 3. After each domain $D_i \in D$ obtaining its adapted model $\mathcal{F}_{\bar{\theta}_i;\bar{\eta}_i}$, another set of data $\{\bar{x}_i, \bar{y}_i\} \subseteq \{\mathcal{X}_i, \mathcal{Y}_i\}$ (query set) is sampled and tested on the adapted model:

$$\bar{\mathcal{L}}^l_{D_i}(\bar{\theta}_i, \bar{\eta}_i) = \mathcal{L}(\mathcal{F}_{\bar{\theta}_i;\bar{\eta}_i}(\bar{x}_i), \bar{y}_i), \tag{4}$$

and θ, η is optimized with all domains in D:

$$\theta = \theta - \beta \frac{\partial \sum_{D_i}^{D} \bar{\mathcal{L}}^l_{D_i}(\bar{\theta}_i, \bar{\eta}_i)}{\partial \theta},$$

$$\eta = \eta - \beta \frac{\partial \sum_{D_i}^{D} \bar{\mathcal{L}}^l_{D_i}(\bar{\theta}_i, \bar{\eta}_i)}{\partial \eta},$$

where β is another step size.

Semi-supervised learning

Facing the challenge of label uncertainty in the multi-domain setting, we extend supervised-based meta-learning to become semi-supervised learning. Inspired by Lee et al. (2013), we convert the model prediction of the unlabeled data to be a hard label as their pseudo-label. Similar to the supervised-learning part, we randomly generate the unlabeled data $\{u_i\} \subset \{\mathcal{U}_i\}$ for each domain D_i. The pseudo-label $\hat{y}_i$ of u_i is produced from the outcome of the model. A problem with using the model outcome as the pseudo-label is that the produced label may include bias from a poorly-trained model in the early training stage. A threshold τ is then introduced to filter the maximum value of unlabeled data prediction, so that only high-confidence outcomes will be converted to hard labels as the produced pseudo-label:

$$\hat{y}_i = \mathbb{1}(\max(\mathcal{F}_{\theta_i;\eta_i}(u_i)) \geq \tau). \tag{5}$$

With pseudo-label, the unlabeled data then have goals to compare with. However, if we directly calculate the cost between the model prediction of u_i and its pseudo-label $\hat{y}_i$, the model will only be trained to maximize the maximum value of u_i, because both the prediction $\mathcal{F}_{\theta_i;\eta_i}(u_i)$ and pseudo-label $\hat{y}_i$ are functions of u_i. Therefore, we further introduce the augmentation from consistency regularization French et al. (2017); Sohn et al. (2020). Augmentation adds noise to the unlabeled data, playing a similar role as the activation function to prevent the prediction of the unlabeled data from being a linear function of u_i. More importantly, as a regularization method, augmentation increases the model generalization and stability: the model should predict the same outcome even with some noise. The semi-supervised part for domain D_i can be presented as

$$\mathcal{L}^u_{D_i}(\theta_i, \eta_i) = \mathcal{L}(\mathcal{F}_{\theta_i;\eta_i}(\mathcal{A}(u_i)), \hat{y}_i), \tag{6}$$

where $\mathcal{A}(\cdot)$ denotes the augmentation function for unlabeled data, for example, cropping, flipping, and noise injection techniques Wen et al. (2020); Iwana and Uchida (2021). In our study, each feature represents a measurement taken in-hospital, and we augment the data with random feature removal, with the assumption that the model should produce similar output even if some measurements are missing. [1]

Adversarial training

By augmenting the unlabeled data for consistency regularization, noise is introduced in training. In order to minimize the side effect of augmentation in training process, we further modify the semi-supervised domain-adapted model training to be adversarial training Ganin et al. (2016). We design the adversarial training between the labeled data and the augmentation of unlabeled data by classifying the source of the latent space from $\mathcal{F}_{\theta_i}$. On the one hand, adversarial training can improve from introducing augmentation, and on the other hand, the potential data shift between labeled and unlabeled data can be addressed too. A discriminator $\mathcal{F}_\phi$ is introduced for the data source classification in each domain D_i:

$$\mathcal{L}^d_{D_i}(\theta_i, \phi_i) = \log(\mathcal{F}_{\theta_i;\phi_i}(x_i)) + \log(1 - \mathcal{F}_{\theta_i;\phi_i}(\mathcal{A}(u_i))) \tag{7}$$

where ϕ represents the parameters of the discriminator.

During the model adaptation process of each domain, the feature extractor and predictor $\mathcal{F}_{\theta_i}$ are trained against the the discriminator $\mathcal{F}_{\theta_i;\phi_i}$:

$$\mathcal{L}_{D_i}(\theta_i, \eta_i, \phi_i) = \mathcal{L}^l_{D_i}(\theta_i, \eta_i) + \mathcal{L}^u_{D_i}(\theta_i, \eta_i) - \lambda\mathcal{L}^d_{D_i}(\theta_i, \phi_i) \tag{8}$$

where λ is a weighting hyper-parameter. The adversarial training aims finding a balanced point $\mathcal{F}_{\bar{\theta}_i;\bar{\eta}_i;\bar{\phi}_i}$ between the feature extractor $\mathcal{F}_\theta$ and discriminator $\mathcal{F}_\phi$ such that

$$\bar{\theta}_i, \bar{\eta}_i = \arg\min_{\theta_i,\eta_i} \mathcal{L}_{D_i}(\theta_i, \eta_i, \bar{\phi}_i) \tag{9}$$

$$\bar{\phi}_i = \arg\max_{\phi_i} \mathcal{L}_{D_i}(\bar{\theta}_i, \bar{\eta}_i, \phi_i) \tag{10}$$

1. Flipping is not an ideal augmentation because the scales of measurements vary, but it could be an option for other time-series data such as ECG. We also tried augmenting the data by adding noise and found that data removal (cropping) is a better solution.

Algorithm 1 SSML

Randomly initialize θ, η
while not done **do**
  Sample domain subset $D' \subseteq \mathcal{D}$
  for $i \in D'$ **do**
    for $m \in [1, M]$ **do**
      Initialize domain network $\mathcal{F}_{\theta_i,\eta_i} \leftarrow (\theta, \eta)$
      Randomly sample support set $\{\{x_i, y_i\}, u_i\}$ and query set $\{\{\bar{x}_i, \bar{y}_i\}, \bar{u}_i\}$
      Compute cost $\mathcal{L}^l_{D_i}(\theta_i, \eta_i)$ from $\{x_i, y_i\}$ in Equation 2 ▷ *Supervised learning*
      Produce pseudo label $\hat{y}_i$ from $\{u_i\}$ in Equation 5 ▷ *Pseudo-labeling*
      Compute classification cost $\mathcal{L}^d_{D_i}(\theta_i, \phi_i)$ in Equation 7 ▷ *Discriminator*
      Adapt parameters $\bar{\theta}_i, \bar{\eta}_i, \bar{\phi}_i$ with gradient descent in Equations 11 12 13 ▷ *Adversarial training*
    end for
    Compute cost $\bar{\mathcal{L}}_{D_i}(\bar{\theta}_i, \bar{\eta}_i)$ from $\{\{\bar{x}_i, \bar{y}_i\}, \bar{u}_i\}$ in Equation 14
  end for
  Update θ and η with domains in D in Equation 15 and 16 ▷ *Meta-learning*
end while

By adversarial training, $\mathcal{F}_\phi$ is trained to determine the source of an example (from labeled data or augmented unlabeled data), but $\mathcal{F}_\theta$ is trained to not recognize them, so that the extracted latent space include the information which is only related to the prediction from $\mathcal{F}_\eta$ without any biased information from augmentation or the domain shift between labeled and unlabeled data. The parameters in predictor $\mathcal{F}_{\eta_i}$ and discriminator $\mathcal{F}_{\phi_i}$ are updated by gradient descent:

$$\bar{\eta}_i = \eta_i - \alpha \cdot \frac{\partial(\mathcal{L}^l_{D_i}(\theta_i, \eta_i) + \mathcal{L}^u_{D_i}(\theta_i, \eta_i))}{\partial \eta_i} \tag{11}$$

$$\bar{\phi}_i = \phi_i - \alpha\lambda \cdot \frac{\partial \mathcal{L}^d_{D_i}(\theta_i, \phi_i)}{\partial \phi_i} \tag{12}$$

The gradient of feature extractor $\mathcal{F}_{\theta_i}$ is reversed in data source classification $\mathcal{L}^d_{D_i}(\theta_i, \phi_i)$, so that the feature extractor is trained toward two parallel directions: the decrease of prediction cost and increase of discrimination cost:

$$\begin{aligned} \bar{\theta}_i = \theta_i - \alpha(& - \lambda \cdot \frac{\partial \mathcal{L}^d_{D_i}(\theta_i, \phi_i)}{\partial \theta_i} \\ & + \frac{\partial(\mathcal{L}^l_{D_i}(\theta_i, \eta_i) + \mathcal{L}^u_{D_i}(\theta_i, \eta_i))}{\partial \theta_i}) \end{aligned} \tag{13}$$

This way, the feature extractor is trained to not be able to recognize if an example is from the labeled data $\{x_i\}$ or the augmented unlabeled data $\{\mathcal{A}(u_i)\}$, and the extracted information is optimized to be prediction-related regardless the bias from adding noise in augmentation.

Semi-supervised meta-learning

Similar to supervised meta-learning in Equation 4, after all the domains in D obtained their adapted model with N steps of training, a query set with unlabeled data for each domain D_i is sampled $\{\{\bar{x}_i, \bar{y}_i\}, \bar{u}_i\} \subseteq \{\{\bar{\mathcal{X}}_i, \bar{\mathcal{Y}}_i\}, \bar{\mathcal{U}}_i\}$ and tested on its adapted model, and adversarial training does not participate in meta-learning

$$\begin{aligned}\bar{\mathcal{L}}_{D_i}(\bar{\theta}_i, \bar{\eta}_i) = &\mathcal{L}(\mathcal{F}_{\bar{\theta}_i;\bar{\eta}_i}(\bar{x}_i), \bar{y}_i) \\ &+ \mathcal{L}(\mathcal{F}_{\bar{\theta}_i;\bar{\eta}_i}(\mathcal{A}(\bar{u}_i)), \mathcal{F}_{\theta_i,\eta_i}(\bar{u}_i)),\end{aligned} \tag{14}$$

and model initialization θ and η is then updated with gradient descent:

$$\theta = \theta - \beta \cdot \frac{\partial \sum_{D_i}^{D} \bar{\mathcal{L}}_{D_i}(\bar{\theta}_i, \bar{\eta}_i)}{\partial \theta}, \tag{15}$$

$$\eta = \eta - \beta \cdot \frac{\partial \sum_{D_i}^{D} \bar{\mathcal{L}}_{D_i}(\bar{\theta}_i, \bar{\eta}_i)}{\partial \eta}, \tag{16}$$

The updated θ can then be used as model initialization in the next training episode. Algorithm 1 is the pseudo-code for our proposed SSML. Section A.1 is the optimization of SSML training.

3.3. Time Domain Variation with SSML

In addition to heterogeneous features and label uncertainty, time-series data also has time domain variation, such as the health condition change when taking treatment, hospital transmission, etc. We propose a time domain variation framework (TDV) based on our proposed SSML and transfer learning. Equation 1 defines the time domain variation in a sequence $s = x^{(1:T)}$. The variation on each sequence s participates in training SSML, and the trained SSML is applied to the domain shift on s with transfer learning. According to SSML, domain D_i for sub-sequence $x^{(1:t_)}$ can obtain their adapted models $\mathcal{F}_{\theta_i;\eta_i}$, so that sub-sequence $x^{(1:t_1)}$ can be encoded and obtain its latent space $h^{(t_1)}$:

$$h^{(t_1)} = \mathcal{F}_{\theta_i}(x^{(1:t_1)}),$$

and the prediction at time t_1 is $\mathcal{F}_{\eta_i}(h^{(t_1)})$.

Assuming the domain is shifted to domain D_j for sub-sequence $x^{(t_1+1:t_2)}$ $(D_i \neq D_j)$, the encoded latent space h_{t_1} from sub-sequence $x^{(1:t_1)}$ is transmitted to domain D_j feature extractor $\mathcal{F}_{\theta_j}$:

$$h^{(t_2)} = \mathcal{F}_{\theta_j}(x^{(t_1+1:t_2)}|h^{(t_1)}).$$

By applying SSML, the homogeneous data on each sequence can be addressed independently through each domain's corresponding model, and transfer learning in the TDV framework connects the time domain variation and includes the historical information which prevents information loss. The representation of the entire sequence $x_{1:T}$ with a series of information transmissions can then be presented as

$$h^{(T)} = \mathcal{F}_{\theta_{\{D\}}}(x^{(1:T)}|h^{(t_1)}, h^{(t_2)}, ...).$$

Figure 1 is the framework of SSML-TDV. SSML can be trained from the sequences with time domain variation and Figure 1 ***Top*** box shows how a trained SSML is used to address the time domain variation.

4. Experiment

Datasets

The PhysioNet Challenge 2012 dataset collects the first 48 hours of measurements after patients are admitted to the intensive-care unit (ICU) Silva et al. (2012). PhysioNet collects 41 variables, including 36 time-series features and five general descriptors: age, gender, height, ICU type, and initial weight. There are 4,000 labeled sequences of mortality, with 13.8 % positive cases, and another 4,000 unlabeled sequences. The hourly average value for each feature is computed, and missing data are imputed with the previous existing values. The mask of data missing is also included as extra features Harutyunyan et al. (2019), and at the same time is used to analyze the frequency of features and determine the domains for heterogeneous features.

MIMIC-III (Medical Information Mart for Intensive Care) is a large EHR dataset collected from the intensive-care unit (ICU) Johnson et al. (2016). MIMIC-III contains the ICU stays of over 38,000 adult patients, which includes a great number of heterogeneous EHR records. We select 17 features and discretize them to be hourly-sampled Harutyunyan et al. (2019). Similar to PhysioNet, the missing data is imputed with previous values. We have three classification tasks for risk prediction in MIMIC-III: physiologic decompensation (whether a patient's health will rapidly deteriorate, binary classification with 2.1 % positive), length of stay in the ICU (multi-class classification, 10 classes/buckets), and in-hospital mortality (binary classification with 8.8 % positive). For length of stay, we evaluate the models using Cohen's kappa coefficient for the inter-annotator agreement, and the mean absolute deviation (MAD) between the predicted length of stay and their reference. For the unbalanced classification tasks decompensation and in-hospital mortality, we introduce both AUROC and AUPRC for evaluation.

Data preprocessing and learning domains

Feature space is an important aspect of data heterogeneity, stemming from potential diagnoses and clinical observations. For example, patients with cardiovascular diseases require more frequent monitoring of blood pressure, and oxygen saturation is more important to anemia or pulmonary patients. Therefore, the distribution of features, including the presence and frequency of condition-specific features, is valuable. In order to analyze feature space with the challenge of multi-dimension data heterogeneity, we calculate the frequency of each feature and use K-means to cluster the sequences based on the combination of frequencies of all features. Each cluster then includes homogeneous sequences with similar feature frequencies and missingness, which indicates the potential similar health conditions. In medicine, a hierarchical clustering method has been applied to cluster patients Ahmad et al. (2014), however, in this study we only cluster feature frequency instead of the raw values, and a comparison shows similar results between K-means and hierarchical clustering (see A.3), therefore, we apply the simpler method K-means to lighten the data preprocessing. To address the problem of the uncertain labels with our proposed SSML, we randomly remove a feature as the augmentation method in Equations 6 and 7. The hourly-average values are computed and the missing data is imputed with the previous value.

Implementations and experimental details

For the multi-source heterogeneity in time-series data, we first test our SSML on a simpler situation of heterogeneity: feature space and label uncertainty (SSML), and later

include the time domain variation into the experiment (SSML-TDV). PhysioNet has both labeled data and unlabeled data. The labeled data is randomly split into 80% training data (20% as validation) and 20% test data for 10 rounds of experiments. Domains of heterogeneous feature frequencies are clustered separately each time for the training set and test set including labeled and unlabeled data. On MIMIC-III, considering the various length of sequences and the variation in the time domain, we build the model based on multiple 24-hour windows on each sequence. Due to the variation and uncertainty in the time domain (e.g., decompensation may happen at multiple random time points during an ICU stay), the early windows are used as the unlabeled data in SSML. PhysioNet only includes sequences with a length of 48-hour which limits the time domain variation, therefore, we only test SSML-TDM on MIMIC-III.

The SSML and SSML-TDV models are implemented on top of an LSTM model with a hidden size of 128. The sequences with heterogeneous feature spaces are clustered into 8 clusters (domains) for PhysioNet and 18 for MIMIC-III (obtained from hyperparameters tuning). In each training episode, five optimization steps are applied on a support set (with labeled and unlabeled data) for each domain with a learning rate of 0.005, and the optimized model for each domain is then tested on another query set. The loss on the query sets from all the randomly sampled domains in this episode is collected to for meta-training with a learning rate of 0.0005. In validation and test sets, we only apply labeled data to evaluate the model performance. Please see section A.2 for details of hyperparameter tuning for pseudo-labeling threshold τ, number clusters, and optimization steps. This work is implemented in Python 3.6 with PyTorch 1.3.1, Numpy 1.18, sklearn 0.21 on our server of 2 Xeon 2.2GHz CPUs, 8 GTX 1080ti GPUs, and 528 GB RAM.

Baseline models

We test our SSML and SSML-TDV against:

- **LogsticRegression**: a logistic regression model with grid search among penalty and the regularization strength.
- **Transformer**: an attention-based model for sequential data without recurrent or convolutional mechanism Vaswani et al. (2017).
- **LSTM**: an LSTM model on hourly time-series medical data with missing data imputed Harutyunyan et al. (2019).
- **P-LSTM**: a phased LSTM model applying a time gate to regulate the access of hidden and cell state of LSTM which captures the time-series irregularity Neil et al. (2016).
- **FixMatch**: a semi-supervised learning method producing confident pseudo-label for unlabeled data and compare with its augmentation Sohn et al. (2020).
- **MAML**: a few-shot-based meta-learning method optimizing global initialization for various tasks and rapidly adapting to any new task Finn et al. (2017).
- **DynEHR**: a meta-learning based model for various lengths of medical data Zhang et al. (2021) (only compare with SSML-TDV for time domain variation).

4.1. Experiments on Heterogeneous Features and Label Uncertainty

PhysioNet

Table 2 represents the experimental results on PhysioNet mortality prediction task. Our proposed SSML shows great improvement over all the baseline models on both AUCOC

Table 2: Average performance (and standard deviations) on PhysioNet.

Evaluation	AUCROC	AUCPRC
LogisticReg	0.711 (0.003)	0.343 (0.005)
Transformer	0.770 (0.009)	0.405 (0.008)
LSTM	0.784 (0.010)	0.399 (0.007)
P-LSTM	0.756 (0.015)	0.368 (0.009)
FixMatch	0.789 (0.013)	0.401 (0.010)
MAML	0.809 (0.007)	0.431 (0.008)
SSML (Ours)	**0.826 (0.008)**	**0.462 (0.007)**

Table 3: Average performance (and standard deviations) on MIMIC-III for heterogeneous features and label uncertainty.

Task	Decompensation		Length-of-stay		In-hospital Mortality	
Evaluation	AUCROC	AUCPRC	Kappa	MAD	AUCROC	AUCPRC
LogisticReg	0.816 (0.016)	0.231 (0.026)	0.346 (0.008)	163.8(10.9)	0.795 (0.011)	0.492 (0.019)
Transformer	0.837 (0.012)	0.241 (0.019)	0.371 (0.019)	160.0(6.9)	0.829 (0.012)	0.497 (0.013)
LSTM	0.848 (0.009)	0.278 (0.012)	0.405 (0.013)	156.2(6.4)	0.835 (0.011)	0.500 (0.010)
P-LSTM	0.836 (0.007)	0.207 (0.014)	0.382 (0.008)	152.4(7.8)	0.834 (0.006)	0.504 (0.009)
FixMatch	0.856 (0.008)	0.282 (0.016)	0.413 (0.016)	157.4 (7.5)	0.840 (0.004)	0.507 (0.008)
MAML	0.868 (0.009)	0.292 (0.007)	0.400 (0.009)	151.5 (4.1)	0.840 (0.008)	0.552 (0.011)
SSML	**0.875(0.010)**	**0.330(0.008)**	**0.422(0.007)**	**148.6(4.7)**	**0.851(0.009)**	**0.575(0.008)**

and AUCPRC. For the models not considering data heterogeneity, LSTM performs the best (compared to LogisticReg, Transformer, and P-LSTM). The comparison between MAML and LSTM shows the benefits of addressing the heterogeneous feature space, and by introducing unlabeled data, FixMatch also has an improvement to LSTM. However, both FixMatch and MAML only address a single type of data heterogeneity. For a multi-source heterogeneity situation in PhysioNet, SSML handles both the heterogeneous features and the label uncertainty, and further improves over FixMatch and MAML.

MIMIC-III

Compared to PhysioNet, MIMIC-III is a more complex dataset with various lengths of sequences. In Table 3, we first focus on the heterogeneous features and label uncertainty in MIMIC-III by simplifying it using the latest 24-hour data. We test MIMIC-III on three learning tasks decompensation, length-of-stay, and in-hospital mortality, and SSML performs the best for all three tasks. Compared to MAML, the improvements of SSML on decompensation and length-of-stay indicate that valuable information from introducing the unlabeled data, and the results on in-hospital mortality further show a better performed meta-learning algorithm SSML with better noise tolerance from the augmented data. When comparing SSML with LSTM and FixMatch, the improvements on SSML further show that specializing the medical sequences from the feature distributions obtain better models on

Table 4: Average performance (and standard deviations) on MIMIC-III full sequences with time domain variation.

Task	Decompensation		Length-of-stay		In-hospital Mortality	
Evaluation	AUCROC	AUCPRC	Kappa	MAD	AUCROC	AUCPRC
LogisticReg	0.839 (0.015)	0.246 (0.017)	0.378 (0.009)	161.2 (8.7)	0.825 (0.011)	0.499 (0.019)
Transformer	0.842 (0.012)	0.260 (0.019)	0.384 (0.014)	147.2 (7.5)	0.836 (0.009)	0.504 (0.010)
LSTM	0.856 (0.011)	0.313 (0.015)	0.423 (0.010)	152.4 (4.2)	0.847 (0.008)	0.515 (0.012)
P-LSTM	0.838 (0.009)	0.237 (0.013)	0.426 (0.012)	145.6 (4.9)	0.848 (0.006)	0.505 (0.008)
FixMatch	0.876 (0.004)	0.317 (0.018)	0.425 (0.006)	151.7 (5.4)	0.854 (0.007)	0.519 (0.009)
MAML	0.879 (0.008)	0.320 (0.011)	0.428 (0.011)	149.5 (4.7)	0.858 (0.009)	0.540 (0.014)
DynEHR	0.863 (0.008)	0.345 (0.009)	0.415 (0.016)	137.4 (7.5)	0.847 (0.006)	0.556 (0.005)
SSML-TDV	**0.906(0.007)**	**0.359(0.006)**	**0.443(0.009)**	**132.6(3.6)**	**0.869(0.007)**	**0.566(0.009)**

each homogeneous set of data, especially with unbalance dataset, obtaining higher AUCPRC values.

4.2. Experiments on Three-source Heterogeneity (including Time Domain Variation)

MIMIC-III

Table 4 represents the results of the three-source heterogeneity in MIMIC-III: heterogeneous features, label uncertainty, and time domain variation. Similar to Table 3, we also test three tasks and provide the averaged performances and their standard deviation. From the table, SSML-TDV performs better than all the baseline models on all the tasks. For example, SSML-TDV improves AUCPRC on decompensation by 13.2 % (0.042) over FixMatch and 11.1 % (0.036) compared to the best baseline model MAML. SSML-TDV and MAML are both meta-learning algorithms, and SSML-TDV has an additional consistency regularization mechanism from the label uncertainty. The improvements of SSML-TDV over MAML indicate a more reliable stable model with higher noise tolerance obtained from applying this consistency regularization method to the augmented data. When compared to FixMatch which also has consistency regularization, the benefits of SSML-TDV then imply that the EHR feature distribution is a valuable aspect of heterogeneity to analyze, and modeling it can help the model better concentrate on each homogeneous set of data.

From Table 4, logistic regression performs the worse in all the models, and LSTM is slightly better than transformer and P-LSTM. When comparing these four static models with SSML-TDV and MAML, we observe that both SSML-TDV and MAML have great improvements, especially on the AUCPRC for decompensation and in-hospital mortality, meaning a better performance on the imbalanced dataset. In addition, the improvement of SSML-TDV over SSML (in Table 3) shows the benefit of the transfer learning mechanism in SSML-TDV by handling the time domain variation in time-series sequences.

5. Discussion

Facing the multi-source heterogeneity problem, we propose SSML for the heterogeneous feature space and label uncertainty and SSML-TDV which also accounts for the time domain variation simultaneously. These two versions offer flexible solutions for researchers to choose appropriate models giving any new problems of data heterogeneity. More importantly, we provide an example of how these complex real-world problems can be effectively solved. From experiments, SSML outperforms all baseline models, including a semi-supervised model for unlabeled data and a meta-learning-based model for heterogeneous features. This demonstrates the advantages of addressing data heterogeneity simultaneously. Furthermore, the robustness of our model is shown through experiments conducted on two different datasets. When extending to time domain variation, our SSML-TDV also outperforms DynEHR, another adaptive model for time domain variation. This indicates that other types of data heterogeneity (heterogeneous features and uncertain labels) can influence the model's adaptation to the time domain, and how our solution can address the problem. Additionally, the higher performance of SSML-TDV compared to SSML alone indicates the value of incorporating historical information in predictions and the importance of adapting models over time.

Limitations A challenge for addressing the heterogeneity in time-series data is the definition of heterogeneity. Our proposed models require pre-defined domains of heterogeneous data. In our experiment, we process the heterogeneity by computing the frequency of each medical measurement and applying an unsupervised clustering method to obtain the groups of patients with similar feature distributions. However, the number of clusters is manually chosen as a hyperparameter, causing the tedious work of searching for the optimal setting. In addition, clustering with a given number of clusters has difficulties handling new activities in practice, for example, a newly discovered disease (e.g., COVID-19) will all be clustered in the existing clusters. In the future, we plan to extend our SSML and SSML-TDV to a flexible number of domains. We look to apply a growing clustering method so that our model can address any new coming data.

6. Conclusion

Time-series data faces the challenge of multi-source heterogeneity, including heterogeneous features, uncertain labels, and time-varying factors. Traditional machine learning techniques have difficulty addressing these heterogeneities simultaneously. In this paper, we propose a semi-supervised meta-learning (SSML) algorithm with adversarial training mechanism for the multi-source heterogeneity challenge in time-series data. Our SSML can address the heterogeneous features and label uncertainty at the same time. In addition, for the time-varying factor, we further introduce a time domain variation framework based on our proposed SSML and transfer learning. We test our proposed models on two real-world medical datasets: PhysioNet Challenge 2012 and MIMIC-III ICU dataset, and over-perform all the baseline models.

References

Tariq Ahmad, Michael J Pencina, Phillip J Schulte, Emily O'Brien, David J Whellan, Ileana L Piña, Dalane W Kitzman, Kerry L Lee, Christopher M O'Connor, and G Michael Felker. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *Journal of the American College of Cardiology*, 64(17):1765–1774, 2014.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27:3365–3373, 2014.

Nannapas Banluesombatkul, Pichayoot Ouppaphan, Pitshaporn Leelaarporn, Payongkit Lakhan, Busarakum Chaitusaney, Nattapong Jaimchariya, Ekapol Chuangsuwanich, Wei Chen, Huy Phan, Nat Dilokthanakul, et al. Metasleeplearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning. *IEEE Journal of Biomedical and Health Informatics*, 2020.

Erik Brynjolfsson, Tom Mitchell, and Daniel Rock. What can machines learn, and what does it mean for occupations and the economy? In *AEA Papers and Proceedings*, volume 108, pages 43–47, 2018.

Yiqiang Chen, Xin Qin, Jindong Wang, Chaohui Yu, and Wen Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4): 83–93, 2020.

Fu-Yuan Cheng, Himanshu Joshi, Pranai Tandon, Robert Freeman, David L Reich, Madhu Mazumdar, Roopa Kohli-Seth, Matthew A Levin, Prem Timsina, and Arash Kia. Using machine learning to predict icu transfer in hospitalized covid-19 patients. *Journal of clinical medicine*, 9(6):1668, 2020.

Thomas Desautels, Jacob Calvert, Jana Hoffman, Qingqing Mao, Melissa Jay, Grant Fletcher, Chris Barton, Uli Chettipally, Yaniv Kerem, and Ritankar Das. Using transfer learning for improved mortality prediction in a data-scarce hospital setting. *Biomedical informatics insights*, 9:1178222617712994, 2017.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

Priyanka Gupta, Pankaj Malhotra, Jyoti Narwariya, Lovekesh Vig, and Gautam Shroff. Transfer learning for clinical time series analysis using deep neural networks. *Journal of Healthcare Informatics Research*, 4(2):112–137, 2020.

Hyukjun Gweon and Hao Yu. A nearest neighbor-based active learning method and its application to time series classification. *Pattern Recognition Letters*, 146:230–236, 2021.

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

Yulan He and Deyu Zhou. Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47(4):606–616, 2011.

Shi Hu, Jakub Tomczak, and Max Welling. Meta-learning for medical image classification. 2018.

Brian Kenji Iwana and Seiichi Uchida. An empirical survey of data augmentation for time series classification with neural networks. *Plos one*, 16(7):e0254841, 2021.

Madiha Javeed, Ahmad Jalal, and Kibum Kim. Wearable sensors based exertion recognition using statistical features and random forest for physical healthcare monitoring. In *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*, pages 512–517. IEEE, 2021.

Jehn-Ruey Jiang, Jian-Bin Kao, and Yu-Lin Li. Semi-supervised time series anomaly detection based on statistics and deep learning. *Applied Sciences*, 11(15):6698, 2021.

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

Abdulla-Al Kafy, Ragib Mahmood Shuvo, Md Nazmul Huda Naim, Md Soumik Sikdar, Radwan Rahman Chowdhury, Md Arshadul Islam, Md Hasnan Sakin Sarker, Md Hasib Hasan Khan, Marium Akter Kona, et al. Remote sensing approach to simulate the land use/land cover and seasonal land surface temperature change using machine learning algorithms in a fastest-growing megacity of bangladesh. *Remote Sensing Applications: Society and Environment*, 21:100463, 2021.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

Zachary C Lipton, David C Kale, Randall Wetzel, et al. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare*, 56, 2016.

Paul Macadam, John B Cronin, Aaron M Uthoff, and Erin H Feser. Effects of different wearable resistance placements on sprint-running performance: A review and practical applications. *Strength & Conditioning Journal*, 41(3):79–96, 2019.

Tarun Naren, Yuanda Zhu, and May Dongmei Wang. Covid-19 diagnosis using model agnostic meta-learning on limited chest x-ray images. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9, 2021.

Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. *Advances in neural information processing systems*, 29, 2016.

Joao Pereira and Margarida Silveira. Learning representations from healthcare time series data for unsupervised anomaly detection. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–7. IEEE, 2019.

Thomas Plötz and Yu Guan. Deep learning for human activity recognition in mobile computing. *Computer*, 51(5):50–59, 2018.

Yoshihiko Raita, Tadahiro Goto, Mohammad Kamal Faridi, David FM Brown, Carlos A Camargo, and Kohei Hasegawa. Emergency department triage prediction of clinical outcomes using machine learning models. *Critical care*, 23(1):1–13, 2019.

Aravind Rajeswaran, Chelsea Finn, Sham Kakade, and Sergey Levine. Meta-learning with implicit gradients. 2019.

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder networks. *arXiv preprint arXiv:1507.02672*, 2015.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. 2005.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171, 2016.

Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*, pages 245–248. IEEE, 2012.

Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Xingyou Song, Wenbo Gao, Yuxiang Yang, Krzysztof Choromanski, Aldo Pacchiano, and Yunhao Tang. Es-maml: Simple hessian-free meta learning. *arXiv preprint arXiv:1910.01215*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Yulin Wang, Jiayi Guo, Shiji Song, and Gao Huang. Meta-semi: A meta-learning approach for semi-supervised learning. *arXiv preprint arXiv:2007.02394*, 2020.

Christian A Webb, Zachary D Cohen, Courtney Beard, Marie Forgeard, Andrew D Peckham, and Thröstur Björgvinsson. Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology*, 88(1):25, 2020.

Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*, 2020.

Garrett Wilson, Janardhan Rao Doppa, and Diane J Cook. Multi-source deep domain adaptation with weak supervision for time-series sensor data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1768–1778, 2020.

Taihong Xiao, Xin-Yu Zhang, Haolin Jia, Ming-Ming Cheng, and Ming-Hsuan Yang. Semi-supervised learning with meta-gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 73–81. PMLR, 2021.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2565–2573, 2018.

Han Yu and Akane Sano. Semi-supervised learning and data augmentation in wearable-based momentary stress detection in the wild. *arXiv preprint arXiv:2202.12935*, 2022.

Lida Zhang, Nathan C Hurley, Bassem Ibrahim, Erica Spatz, Harlan M Krumholz, Roozbeh Jafari, and Mortazavi J Bobak. Developing personalized models of blood pressure estimation from wearable sensors data using minimally-trained domain adversarial neural networks. In *Machine Learning for Healthcare Conference*, pages 97–120. PMLR, 2020.

Lida Zhang, Xiaohan Chen, Tianlong Chen, Zhangyang Wang, and Bobak J Mortazavi. Dynehr: Dynamic adaptation of models with data heterogeneity in electronic health records. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE, 2021.

Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2487–2495, 2019.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.

Appendix A.

A.1. Optimization for SSML Training

We now explain the optimization process of training our proposed SSML algorithm. The objective of SSML includes the feature extraction θ, prediction network η, and the discriminator ϕ:

$$\underset{\{\theta\},\{\eta\}}{\text{minimize}}\, \mathcal{L}(\theta, \eta, \phi), \quad \underset{\{\phi\}}{\text{maximize}}\, \mathcal{L}(\theta, \eta, \phi)$$

such that

$$\theta_n = \Theta_n(\theta_{n-1}),\ \eta_n = \Psi_n(\eta_{n-1}),\ \phi_n = \Phi_n(\phi_{n-1}) \quad (n \in [1, N])$$

where Θ_n, Ψ_n, Φ_n represents the gradient step of parameter optimizations at step n of the SSML adversarial training. The Lagrangian is this:

$$\mathcal{L}(\{\theta\}, \{\eta\}, \{\phi\}, \delta, \epsilon, \sigma) = \ell(\theta, \eta, \phi) + \sum_{n}^{N} \delta_n(\Theta(\theta_{n-1}) - \theta_n)$$
$$+ \sum_{n}^{N} \epsilon_n(\Psi_n(\eta_{n-1}) - \eta_n) - \sum_{n}^{N} \sigma_n(\Phi_n(\phi_{n-1}) - \phi_n)$$

where δ_n, ϵ_n and σ_n are the associated Lagrangian multipliers of step n of Θ, Ψ, and Φ. The derivatives of the last step of SSML inner loop are given as:

$$\nabla_{\theta_N} \mathcal{L} = \nabla_{\theta_N} \ell(\theta_N, \eta_N) - \nabla_{\theta_N} \ell(\theta_N, \phi_N) - \delta_N$$

$$\nabla_{\eta_N} \mathcal{L} = \nabla_{\eta_N} \ell(\theta_N, \eta_N) - \epsilon_N$$

$$\nabla_{\phi_N} \mathcal{L} = \nabla_{\phi_N} \ell(\theta_N, \phi_N) - \sigma_N$$

At each intermediate step n of SSML, the derivatives are:

$$\nabla_{\theta_n} \mathcal{L} = -\delta_n + \delta_n \nabla_{\theta_n} \Theta_{n+1}(\theta_n|\eta_N) - \delta_n \nabla_{\theta_n} \Theta_{n+1}(\theta_n|\phi_N), \quad n \in [1, N-1]$$

$$\nabla_{\eta_n} \mathcal{L} = -\epsilon_n + \epsilon_n \nabla_{\eta_k} \Psi_{n+1}(\eta_n), \quad n \in [1, N-1]$$

$$\nabla_{\phi_n} \mathcal{L} = -\sigma_n + \sigma_n \nabla_{\phi_k} \Phi_{n+1}(\phi_n), \quad n \in [1, N-1]$$

Each derivative is set to zero to optimize the model:

$$\epsilon_N = \nabla_{\eta_N} \ell(\eta_N)$$
$$\epsilon_n = \epsilon_{n+1} + \nabla_{\eta_n} \Psi_{n+1}(\eta_n), \quad n \in [1, N-1]$$
$$\sigma_N = \nabla_{\phi_N} \ell(\phi_N)$$
$$\sigma_n = \sigma_{n+1} + \nabla_{\phi_n} \Phi_{n+1}(\phi_n), \quad n \in [1, N-1]$$
$$\delta_N = \nabla_{\theta_N} \ell(\theta_N)$$
$$\delta_n = \delta_{n+1} + \nabla_{\theta_n} \Theta_{n+1}(\theta_n|\eta_n) - \nabla_{\theta_n} \Theta_{n+1}(\theta_n|\phi_n), \quad n \in [1, N-1]$$

A.2. Hyperparameters Study

One important hyperparameter in our proposed SSML is the threshold τ in pseudo-labeling (Equation 6). We test the different settings for hyperparameter τ of 0.5, 0.6, 0.7, 0.8, 0.9, as well as 0 (using all produced pseudo-labels) for both PhysioNet Challenge 2012 and MIMIC-III datasets, and compare with the baseline models FixMatch (with different settings of τ) and MAML (i.e. $\tau = 1$).

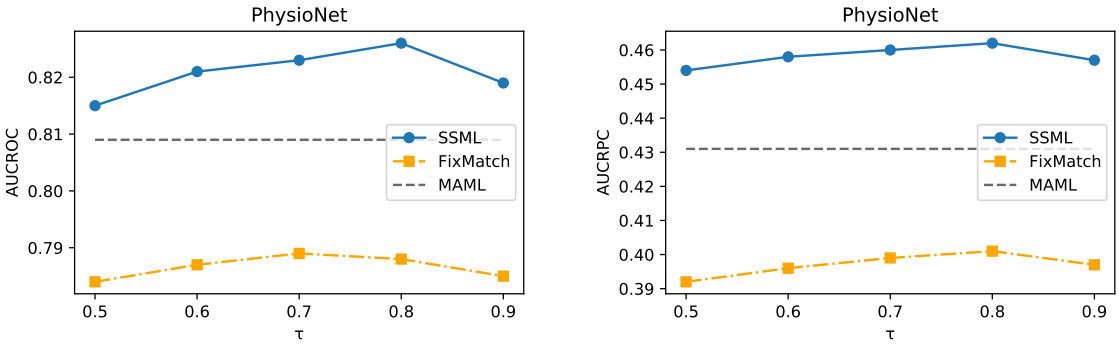
Figure 2 shows the experiments on PhysioNet. From the figure, $\tau = 0.8$ is the optimal setting for SSML, and for FixMatch, the optimal τ is around 0.7 to 0.8. When τ is 0.5 or 0.9, the performance decreases for both SSML and FixMatch, and there is a further decrease when τ is 0. This result indicates that the threshold τ can filter out the samples with low-confidence pseudo labels, and can improve the model performance by providing high-confidence samples in consistence generalization. However, a very high value of τ can cause a decrease because too few samples are kept and the model only gets limited benefits from the very little pseudo-labeling. The performance of SSML varies between different values of τ, but all are better than MAML and FixMatch. MAML does not have the hyperparameter τ, so we only compare with its average performance.

Figure 3 includes the experiments for hyperparameters τ on all tasks of MIMIC-III: Figures 3(a)subfigure, 3(b)subfigure are the performance comparison of AUCROC and AUCPRC for Decompensation, 3(c)subfigure and 3(d)subfigure are the comparison for In-hospital Mortality, and 3(e)subfigure and 3(f)subfigure are Kappa score and MAD for Length-of-stay. Note that the higher values of AUCROC, AUCPRC, Cohen's Kappa, and lower MAD represent better performance. The best performing τ is around 0.7. Similar to the experiment on PhysioNet, SSML and FixMatch perform the worst when τ is 0, and there is also a decrease when τ is a large value. For Decompensation, In-hospital Mortality, and Cohen's Kappa score of Length-of-stay, SSML performs better than both MAML and FixMatch for all the settings of τ. However, for MAD of Length-of-Stay, SSML is only better than MAML when τ is between 0.5 to 0.8.

In addition to the hyperparameter τ, we also test the different number of clusters in data preprocessing, and the meta-learning steps (inner loop). We test the number of clusters between 5 to 40, and observe that the optimal setting is eight clusters for PhysioNet and 18 for MIMIC-III. The reason may come from the size of the dataset. PhysioNet only includes 4,000 labeled data and 4,000 unlabeled data, and MIMIC-III has over 38,000 patients recorded, and the bigger dataset needs more clusters. We also run experiments for the steps of inner loop optimization from 1 to 15, and the optimal step is 5 for both PhysioNet and MIMIC-III.

A.3. Clustering Results Analysis

In the study, we use clustering to address the heterogeneous feature space. Instead of using the actual feature values, we apply the clustering on the feature frequency, so that the samples in each cluster have similar feature occurrences. In medicine, hierarchical clustering is widely applied. Here we compare the statistical analysis of the two different clustering methods - K-means and hierarchical clustering. We use 18 clusters for both methods, the optimal setting obtained from the prediction tasks. For each cluster, we calculate the percentage of positive samples for the two binary classification tasks mortality and decompensation,



(*a*) PhysioNet AUCROC (*b*) PhysioNet AUCPRC

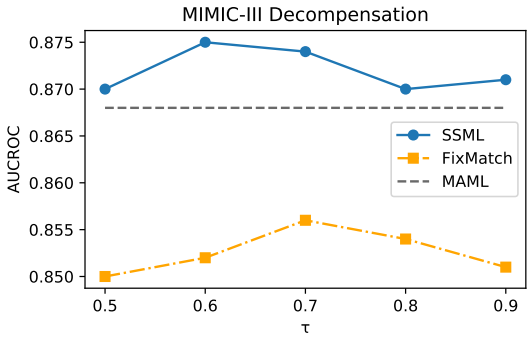
Figure 2: Hyperparameters comparison on PhysioNet: Blue, gray, orange represent our proposed SSML, MAML, and FixMatch repectively. X-axis is the hyperparameter τ and y-axis are the AUCROC in (a) and AUCPRC in (b). The optimal τ is around 0.8 on PhysioNet.

and the average length of hospital stay for length-of-stay. Then, the results from all the clusters are used to obtain the mean and standard deviation, maximum, and minimum values. Through these statistical data, we can learn if the two clustering methods have a significant difference, and also if any of the clustering methods have a serious bias, for example, separating the very sick patients from others.

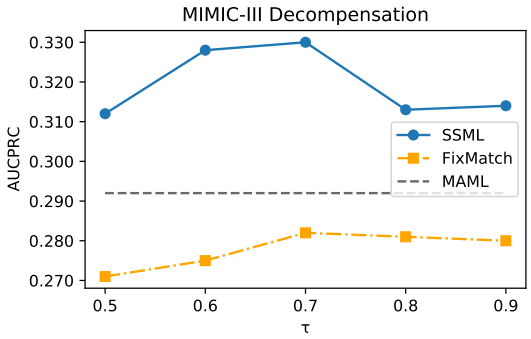
Table 5 shows our analyzing results. When comparing the two clustering methods K-means and hierarchical clustering, we learn that the two methods do not have a significant difference. They have very similar average, standard deviation, maximum, and minimum values for all three tasks. To lighten the data preprocessing and focus on addressing the multi-source heterogeneity problem, we use the simpler method K-means in the paper to obtain the learning domains for the heterogeneous feature space problem. On the other hand, both clustering methods have low standard deviation values for all three tasks, indicating that both methods do not cause serious bias in the clustering results.

Table 5: The label distributions with K-means and hierarchical clustering methods.

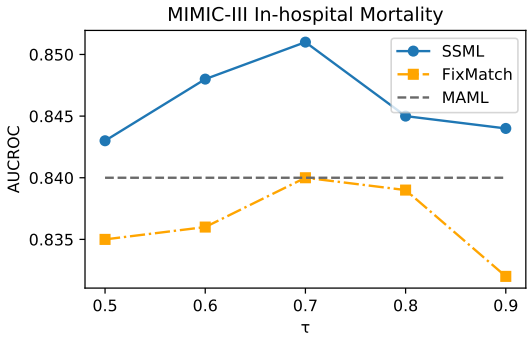
	K-means			Hierarchical clustering		
	Avg (stdev)	Max	Min	Avg (stdev)	Max	Min
In-hospital mortality	0.146 (0.039)	0.222	0.082	0.144 (0.033)	0.207	0.073
Decompensation	0.027 (0.010)	0.052	0.014	0.026 (0.011)	0.049	0.011
Length-of-stay	153.6(35.9)	225.2	106.9	159.3 (38.8)	230.7	107.4



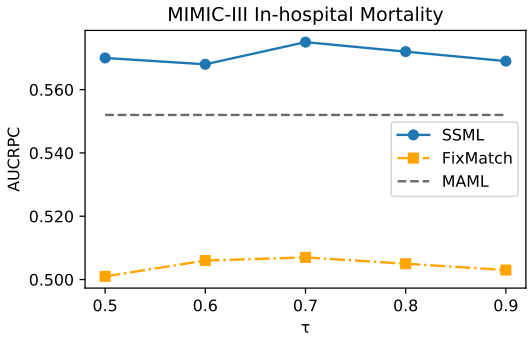
(a) Decompensation AUCROC



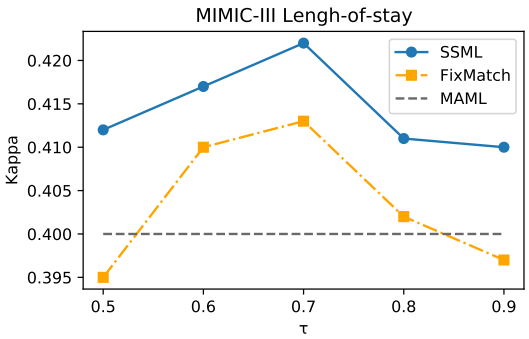
(b) Decompensation AUCPRC



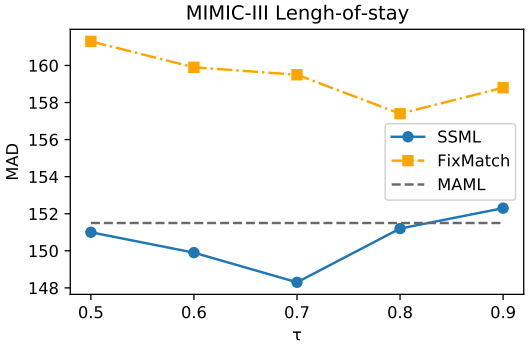
(c) In-hospital Mortality AUCROC



(d) In-hospital Mortality AUCPRC



(e) Lengh-of-stay Kappa



(f) Lengh-of-stay MAD

Figure 3: Hyperparameters comparison on MIMIC-III: Blue, gray, orange represent our proposed SSML, MAML, and FixMatch repectively. X-axis is the hyperparameter τ and the optimal τ is around 0.7.