UDAMA: Unsupervised Domain Adaptation through Multi-discriminator Adversarial Training with Noisy Labels Improves Cardio-fitness Prediction

Yu Wu   YW573@CAM.AC.UK
Department of Computer Science and Technology, University of Cambridge, UK

Dimitris Spathis   DS806@CAM.AC.UK
Nokia Bell Labs, Cambridge, UK
Department of Computer Science and Technology, University of Cambridge, UK

Hong Jia   HJ359@CAM.AC.UK
Department of Computer Science and Technology, University of Cambridge, UK

Ignacio Perez-Pozuelo   IP325@CAM.AC.UK

Tomas I. Gonzales   TOMAS.GONZALES@MRC-EPID.CAM.AC.UK

Soren Brage   SOREN.BRAGE@MRC-EPID.CAM.AC.UK

Nicholas Wareham   NICK.WAREHAM@MRC-EPID.CAM.AC.UK
MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, UK

Cecilia Mascolo   CM542@CAM.AC.UK
Department of Computer Science and Technology, University of Cambridge, UK

Abstract

Deep learning models have shown great promise in various healthcare monitoring applications. However, most healthcare datasets with high-quality (gold-standard) labels are small-scale, as directly collecting ground truth is often costly and time-consuming. As a result, models developed and validated on small-scale datasets often suffer from overfitting and do not generalize well to unseen scenarios. At the same time, large amounts of imprecise (silver-standard) labeled data, annotated by approximate methods with the help of modern wearables and in the absence of ground truth validation, are starting to emerge. However, due to measurement differences, this data displays significant label distribution shifts, which motivates the use of domain adaptation. To this end, we introduce **UDAMA**, a method with two key components: **Unsupervised Domain Adaptation** and **Multi-discriminator Adversarial Training**, where we pre-train on the silver-standard data and employ adversarial adaptation with the gold-standard data along with two domain discriminators. In particular, we showcase the practical potential of UDAMA by applying it to Cardio-respiratory fitness (CRF) prediction. CRF is a crucial determinant of metabolic disease and mortality, and it presents labels with various levels of noise (gold-and silver-standard), making it challenging to establish an accurate prediction model. Our results show promising performance by alleviating distribution shifts in various label shift settings. Additionally, by using data from two free-living cohort studies (Fenland and BBVS), we show that UDAMA consistently outperforms up to 12% compared to competitive transfer learning and state-of-the-art domain adaptation models, paving the way for leveraging noisy labeled data to improve fitness estimation at scale.

1. Introduction

Deep learning (DL) has been widely applied to many healthcare applications, such as sleep stage classification (Yildirim et al., 2019), stress detection (Jaques et al., 2017), and fitness prediction such as cardiorespiratory fitness and exercise adherence prediction (Sakr et al., 2018; Zhou et al., 2019). However, the significant breakthroughs and encouraging results brought by DL are often accompanied by the need for accurate data collection under laboratory-controlled experiments and clinically verified labeling, termed as *gold-standard*.

Collecting high-quality labels (i.e., gold-standard) for healthcare applications may require extensive effort and can be particularly time-consuming. For example, developing a precise epileptic seizure diagnosis model requires electroencephalography (EEG) during ambulatory screening (Shoeibi et al., 2020) for accurate brain function detection and a diagnosis from physicians or neurologists to label seizure occurrence. Consequently, most existing datasets tend to be small-scale, leading to poor performance and model generalization (Raschka, 2018) when developing DL models for different cohorts.

In comparison, with the widespread use of mobile and wearable devices, less accurate yet large-scale labels, referred to as *silver-standard*, are available. For example, Heart rate or Oxygen saturation (SpO_2) can be easily gathered from smartwatches in daily life without clinical visits. However, these silver-standard labels, which are derived from a less accurate estimation scheme, often lead to predictions with lower accuracy, as they are often noisy and display distribution shifts compared to gold-standard labels (Karimi et al., 2020).

As such, gwojuedold-standard labels are crucial for the development and validation of robust clinical models. Although silver-standard labels with extensive labeling are easy to access, they usually contain noise due to less accurate collection schemes and are characterized by distribution shifts, which makes validation against gold-standard data difficult. In this paper, we answer the following question: *Can we leverage large-scale noisy silver-standard datasets to improve deep learning model validation on gold-standard datasets for healthcare applications?*

Domain adaptation (DA) is the natural candidate to solve the problem of distribution mismatch between the source (large-scale data) and target domains (small-scale data) (Patricia and Caputo, 2014). Specifically, discrepancy-based and adversarial-based techniques are prevalent among recent DA methods and achieve state-of-the-art performance in image and language tasks (Sun and Saenko, 2016; Shen et al., 2017; Ganin et al., 2015; Bousmalis et al., 2016; Du et al., 2020). Discrepancy-based DAs seek to reduce the divergence between domains, especially the distance between two domains in the feature space, and demonstrate promising results (Gretton et al., 2012; Wang and Deng, 2018b). On the other hand, adversarial-based DAs employ domain discriminators to encourage domain confusion through an adversarial objective (Wang and Deng, 2018b; Ganin et al., 2015). However, these approaches mainly concentrate on large-scale and gold-standard source labels to adapt to unlabeled target data, while neglecting the fact that the source domain might contain noisy silver-standard labels.

In this paper, we present UDAMA and target the *label distribution* shift problem from noisy labeled but large-scale source domains caused by less accurate data collection schemes to target domains with small-scale gold-standard labels. Our proposed framework is inspired by adversarial-based DA methods, which often contain a specific discriminator that

categorizes the source or target domains from which samples originate. However, compared to existing approaches that mainly focus on discriminating domains as a binary classification task, UDAMA captures the fine-grained information that resides within the label distribution shifts. To achieve this, we introduce a fine-grained discriminator that attempts to discriminate the distribution of domain labels, thereby capturing domain-invariant feature representation learning. As a result, through a multi-discriminator learning scheme, UDAMA can effectively learn cross-domain representation by integrating both coarse-grained and fine-grained domain information, resulting in promising performance on small-scale datasets.

We practically demonstrate the potential of UDAMA in the context of healthcare data by applying it to cardiorespiratory fitness (CRF) prediction. CRF is a significant predictor of cardiovascular disease (CVD) (Laukkanen et al., 2001), one of the leading causes of death globally (Kaptoge et al., 2019). CRF is directly measured by maximal oxygen consumption (VO_2max), which is assessed using heart rate responses to standard maximal exercises test (i.e., gold-standard). However, collecting VO_2max labels through such tests is difficult and expensive and needs extra equipment and clinical monitoring. Alternatively, submaximal VO_2max tests (Gonzales et al., 2020b) have been proposed to capture fitness levels. Despite their potential, such measurements have been shown to provide silver-standard labels with lower accuracy and a shift in distribution compared with gold-standard VO_2max.

In this paper, we validate the power of silver-standard VO_2max to enhance the precision of fitness predictions by mitigating distribution shifts using the proposed UDAMA. Specifically, we apply UDAMA on a large CRF datasets (Lindsay et al., 2019; Gonzales et al., 2021) and show that UDAMA achieves the SOTA performance. Our contributions are summarized as follows:

- We propose a novel domain adaptation framework via multi-discriminator adversarial training, which incorporates samples from different distributions and allows us to learn better feature representations for (often small) gold-standard datasets.

- In the CRF prediction task, gold-standard VO_2max are hard to get and silver-standard labels display domain shift problems. We are able to address the domain shift problem using UDAMA, significantly improving the prediction accuracy in the target domain. Furthermore, we stress-test our models with semi-synthetic data under various label shifts to show and test for robustness.

- Through a set of extensive experiments, we show that UDAMA achieves strong results (corr = 0.701 ± 0.032) and improves model performance up to 12.0% compared to baselines on the CRF tasks. These results show that the proposed model is able to improve fitness prediction by only leveraging large-scale silver-standard labels.

Generalizable Insights about Machine Learning in the Context of Healthcare

In healthcare, high-quality datasets with gold-standard labels are expensive and hard to collect, leading to sparse and small-scale datasets that make it difficult to generalize to other unseen cohorts and applications. In contrast, collecting silver-standard labels from less accurate collection schemes with modern wearables is more affordable. However, these extensive less-accurate labels exhibit distribution discrepancies when compared to ground

truth and cannot be directly leveraged for model deployment. Our work addresses this by introducing a multi-discriminator domain adaptation method for cross-domain representation learning, which transfers knowledge from large-scale weakly labeled data to small-scale health datasets with gold-standard labels. Specifically, in the context of CRF prediction, our results demonstrate that leveraging large-scale noisy VO2max labels using UDAMA not only achieves improved fitness prediction but also effectively mitigates label distribution shifts. This paves the way for the practical application of machine learning for real-world health outcomes. Furthermore, our approach can be easily adapted to various health-related tasks, particularly those involving high-dimensional time-series data and changes in label distribution for regression tasks.

2. Related Work

2.1. Cardio-fitness estimation

Numerous prediction models have been developed recently using different testing schemes (submaximal exercise / non-exercise tests) and a variety of machine learning methods (Jensen et al., 2021; Abut et al., 2016) to substitute direct measurements of VO_2max, the direct indicator of CRF. Specifically, with the help of modern wearable technology, which can track physical activity, resting heart rate (RHR), and other biosignals, various silver-standard methods have emerged for more convenient VO_2max calculation without maximal exercise testing (Plasqui and Westerterp, 2006; Esco et al., 2011; Shcherbina et al., 2017; Henriksen et al., 2018; Perez-Pozuelo et al., 2021). For example, deep learning models are utilized to predict fitness prediction converting raw wearable sensor data (Spathis et al., 2022). Nonetheless, these approaches overlook the impact of silver-standard labels generated from imprecise testing schemes, which can result in diminished model performance and untrustworthy fitness forecasts. Recently, a few works applying machine learning models on CRF predicting using gold-standard VO_2max labels have been proposed (Abut and Akay, 2015). Nonetheless, they mainly test on small cohorts, which might lead to poor generalization performance. This paper proposes a novel DL method that aims to alleviate the distribution difference between imprecise silver-standard and gold-standard data for better fitness prediction (Results are discussed in §6).

2.2. Domain adaptation

A domain combines the input population with the output following a certain probability distribution (Kouw and Loog, 2018). DA is one of the state-of-the-art solutions (Hoffman et al., 2017) for learning information from an abundant labeled source domain and applying it to a new and unseen target domain with a different distribution. DA trains a feature extractor to learn the shared information across domains, so the model can generalize to new settings and thus mitigate the domain shift (Kouw and Loog, 2018). Discrepancy-based and adversarial-based DA approaches are effective among DA methods.

2.2.1. Discrepancy-based method

For discrepancy-based approaches, the network typically shares or reuses the initial layers between the source and target domains and aims to reduce feature space divergence (Wang

and Deng, 2018b). The maximum mean discrepancy (MMD) (Gretton et al., 2012) is an effective distance metric that measures the distribution divergence between the mean embeddings of two distributions in the reproducing kernel Hilbert space (RKHS) to minimize the difference between two distributions. Many methods (Tzeng et al., 2014; Long et al., 2015) utilize the MMD metric within the network to learn domain-invariant and discriminative representations. On the other hand, the correlation alignment (CORAL) (Sun and Saenko, 2016) method is proposed to align the second-order statistics of the source and target distribution with a linear transformation. Deep-Coral (Sun and Saenko, 2016) is an extension of CORAL that can train a non-linear transformation to align the correlations of the representation embedding in deep neural networks from the source domain to the destination domain. Besides, some works minimize empirical Wasserstein distance between source and target feature representations and also showed good results in domain-invariant representation learning approaches (Shen et al., 2017). Although these methods are effective and easily incorporated into deep neural networks, they are mainly designed for feature-based distribution shift alignment. Unlike most discrepancy-based methods, our method mainly addresses label distribution shift problems when the source domain has large-scale noisy labels.

2.2.2. Adversarial-based DA

Most adversarial-based methods are motivated by theory suggesting that a good cross-domain representation contains no discriminative information about the origin of the input and shows good performance to reduce domain discrepancy (Wang and Deng, 2018a). Among these methods, Domain-Adversarial Neural Network (DANN) (Ganin et al., 2015) first introduces adversarial training to domain adaptation. DANN utilizes a shared feature extractor to learn feature embedding and a discriminator to maximize the domain difference. After the convergence of the whole training, general features are learned for each input while their domain information cannot be discriminated. Followed by this, Domain Separation Networks (DSN) (Bousmalis et al., 2016) and Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al., 2017) propose more sophisticated feature extraction methods using both shared and private feature extractors for the task classifier and domain discriminator. Moreover, Dual Adversarial Domain Adaptation (DADA) (Du et al., 2020) uses two discriminators to pit against each other to aid the discriminative training.

The aforementioned methods mainly focus on discriminating domains as a coarse binary classification task while neglecting the rich information of the domain distribution. Moreover, the majority of applications of DA mainly focus on medical image segmentation and classification tasks (Perone et al., 2018; Mei et al., 2020; Venkataramani et al., 2018). Our work aims to resolve the label distribution shift while keeping the input data constant for regression targets. We propose a novel DA framework (UDAMA) with multi-discriminators to learn the coarse-grained and fine-grained domain information and validate it on a CRF prediction task.

3. Cardio-respiratory Fitness Prediction

CRF is one of the strongest predictors of CVD compared with other risk factors like hypertension and type 2 diabetes (Laukkanen et al., 2001). Routinely assessing CRF through
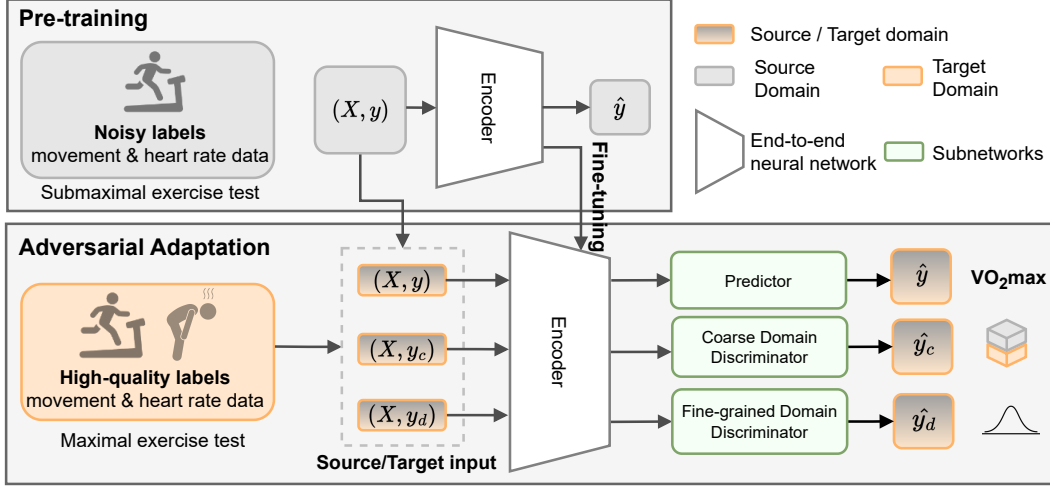
Figure 1: **UDAMA architecture and data pipeline.**

VO$_2$max, which is considered the benchmark measurement, provides valuable insights into a person's overall fitness. However, obtaining gold-standard VO$_2$max measurements, as shown in Figure 1, is time-consuming and thus rarely performed in clinical settings. In particular, it requires participants to undergo a maximal exercise test to reach exhaustion on a treadmill, while wearing a face mask with a computerized gas analysis system to monitor ventilation and expired gas fractions.

Recently, less-accurate measurement schemes such as sub-maximal exercise tests (silver-standard) utilizing modern wearables embedded with accelerometers and ECG sensors have started to provide opportunities for population-level fitness prediction. However, this alternative measurement method has been shown to demonstrate a measurement bias ranging from -3.0 to -1.6 ml O2/min/kg and a Pearson's r ranging from 0.57 to 0.79 (Gonzales et al., 2020b) compared to gold-standard. Apart from producing less accurate VO$_2$max values, these measurements also exhibit distribution mismatches, making it difficult to integrate into clinical practice.

Similar to other healthcare applications, the distribution shift between silver- and gold-standard labels in the CRF prediction task is often ill-defined. To tackle this issue, this paper aims to adapt the source domain, characterized by noisy yet large-scale silver-standard labels, to the target domain with small-scale gold-standard datasets. Specifically, we introduce a novel adversarial-based unsupervised domain adaptation framework with multiple domain discriminators, i.e., UDAMA to learn domain-invariant features and improve model validation on gold-standard VO$_2$max prediction.

4. Methods

This work introduces a novel unsupervised domain adaptation framework that utilizes multi-discriminator during adversarial training. The overall model architecture and multi-discriminator training scheme are shown in Figure 1. Herein, in this section, we formally

discuss the problem formulation (§4.1) and details of our framework, which includes the first-step pre-training and second-step multi-discriminator domain adaptation training (§4.2).

4.1. Problem Formulation and Notation

Here, we denote D_s as the source domain containing silver-standard labels and D_t as the target domain with gold-standard labels, as shown in Figure 1. For each domain, we assume the data as $X = (x_1, ..., x_n) \in \mathbb{R}^{N \times T \times F}$ corresponds to the accelerometer and Electro-cardiogram (ECG) data from a chest ECG device and a target regression VO_2max labels $y = (y_1, ..., y_n) \in \mathbb{R}^N$. Additionally, we take into account contextual information such as the height or weight as metadata $M = (m_1, ..., m_n) \in \mathbb{R}^{N \times F}$. For the input data X and M, N represents the number of samples/subjects, T represents the length of input sequences, and F represents the number of input features. Besides, We use coarse- and fine-grained domain labels to train our model during adaptation and utilize multiple discriminators to differentiate between them. In particular, $y_c = (y_c[1], ..., (y_c[n])$ is the categorical value representing the coarse-grained binary domain label. $y_d = (y_d[1], ..., (y_d[n])$ is the numerical value that denotes the fine-grained domain distribution label. The coarse-grained domain discriminator is D_c and the fine-grained domain discriminator is D_f. Also, for the training process, we denote the feature encoder with E and the regression predictor with G_y. The overall networks thus can be represented as $\hat{y}_c = D_c \cdot E$, $\hat{y}_d = D_f \cdot E$ and $\hat{y} = G_y \cdot E$. The full table is shown in Table 4.

4.2. Unsupervised Domain Adaptation and Multi-discriminator Adversarial Training

Domain adaptation is a method for learning a mapping between domains with distinct distributions, including data distribution shifts such as covariate shift, conditional shift, and label distribution (Wang and Deng, 2018b). In this paper, we propose UDAMA, the unsupervised domain adversarial training, to address the label distribution shift problem, particularly when the source domain contains numerous noisy labels.

As shown in Figure 1, after pre-training on source domain with large-scale silver-standard labels, we first incorporate part of prior knowledge from the D_s to create the adversarial training environment. Then we use the mixed silver-standard and gold-standard data to train the predictors and discriminator during the adaptation phase.

In particular, the adversarial training process consists of an encoder (E), a VO2max label predictor(G_y), and two domain classifiers/discriminators (D) designed for label shift problems by distinguishing both the domain and domain distribution information. During training, the fine- (D_f)and coarse-grained (D_c) discriminators are first optimized to identify the domain of each sample (i.e., $\max D_f, D_c$). In adversarial, the label predictor and encoder are then optimized to predict continuous fitness values from encoding(i.e., $\min E, G_y$). The above-mentioned adversarial process will finally achieve the trade-off (i.e., the best prediction result in the most difficult-to-distinguish domain).

4.2.1. COARSE-GRAINED DISCRIMINATOR

The coarse-grained discriminator (D_c) is similar to other DAs (de Mathelin et al., 2020; Zhao et al., 2018) and follows the DANN (Ganin et al., 2015) style. In other words, D_c

aims to discriminate the source of each data point in the mixture of pre-training and target labeled data as a binary classification task, where 0 represents the data comes from the $\boldsymbol{D_s}$ and 1 from $\boldsymbol{D_t}$. In specific, After getting the representation matrix by fine-tuned feature extractor, two fully connected layers with the corresponding activation in $\boldsymbol{D_c}$ are used to discriminate the rough binary domains labels and predict a probability vector $\hat{\boldsymbol{y_c}}$. Let $\hat{\boldsymbol{Y_c}} = \hat{\boldsymbol{y_c}}[\boldsymbol{n}]$ denote the predicted probability vectors for all the data points in $(\boldsymbol{D_t})$. The classification loss of the coarse-grained discriminator is defined as:

$$L_{CSE} = \sum_N l_c(y_c[n], \hat{y_c}[n]) \tag{1}$$

where l_c is the cross entropy loss of a single data point, and by optimizing $\boldsymbol{L_{CSE}}$ for $\boldsymbol{D_c}$, we can force the extractor to learn a general feature by maximizing such divergence.

4.2.2. FINE-GRAINED DISCRIMINATOR

However, a simple binary classification task cannot properly represent the domain label distribution. Therefore, we augment adversarial training with a fine-grained discriminator $(\boldsymbol{D_f})$ to discriminate distribution differences. Specifically, instead of generating the binary domain labels (0 or 1) for the source or target domain for each sample, we construct a more complex pseudo-label $(\boldsymbol{y_d})$ to represent its domain label distribution. Based on our observation of the health outcome labels, which conform to the Gaussian distribution, as shown in Figure 2, we then assign the $(\boldsymbol{y_d})$ for each training sample during adaptation. Specifically, $\boldsymbol{y_d}$ represents the mean and variance of the regression label distribution. This label is based on whether the sample is from the pre-training or target domain. Therefore, after generating the feature matrix using $\boldsymbol{E}$, $\boldsymbol{D_f}$ is designed to distinguish the mean and variance of the label distribution using two fully connected layers with the corresponding activation. Let $\hat{\boldsymbol{Y_d}} = \hat{\boldsymbol{y_d}}[\boldsymbol{n}]$ denote the predicted probability vectors for all the data points in $(\boldsymbol{D_t})$. Then, the loss of the fine-grained discriminator is defined as:

$$L_{GLL} = \sum_N l_g(y_d[n], \hat{y_d}[n]) \tag{2}$$

where l_g is each data point's Gaussian Negative Log-Likelihood (GLL) loss. In particular, GLL optimizes the mean and variance of a distribution and thus further maximizes the nuance changes among the sample and updates the discriminator.

After that, D_c and D_f can maximize the difference between the source and target domains for the multiple-domain discriminator training scheme. Meanwhile, the encoder and the predictor try to maximize the correct prediction of y. These two modules play two games and finally reach a balance during the training. As a result, the encoder and predictor can learn a representation that cannot tell the difference between the source and target domains after the training is converged.

4.2.3. OBJECTIVE FUNCTIONS AND TRAINING

For adversarial training, the shared encoder first leverages the pre-trained model and extract general feature. Then discriminators are trained simultaneously to differentiate the domain labels. The predictor is used for the regression healthcare outcomes prediction task, and

a mean squared error loss L_{MSE} is applied to optimize the G_y. The detailed training overflow and input for each module are shown in Figure 4. Finally, the whole framework can be optimized by the total loss L, which is defined as:

$$L = \alpha L_{MSE} - \lambda_1 L_{CSE} - \lambda_2 L_{GLL} \tag{3}$$

We optimize the overall loss L to minimize the predictor loss while maximizing the loss of domain discriminators. In detail, α is used to scale down the predictor loss to the same level as predictors, λ_1 and λ_2 control the relative weight of the discriminator loss, and $\lambda_1 + \lambda_2 = 1$. Hyperparameter choice details are discussed in Section §5.

5. Experiments

We conduct various qualitative and quantitative evaluations to validate our models on the cardio-respiratory fitness prediction task with wearable sensing (§5.1). We outline the model's architecture and how it differs from other baselines (§5.2 and §5.3). Next, we further quantify the impact of incorporating source domain knowledge to UDAMA (§5.4). The overall workflow is depicted in Figure 1. Code is available at: `https://github.com/yvonneywu/UDAMA-CRF-Prediction`.

5.1. Datasets and Training Strategy

Datasets. The source domain comes from the silver-standard measurement study (Fenland), including 11,059 participants. It contains a combined heart rate and movement signal from chest ECG sensor Actiheart and noisy VO_2max labels collected from submaximal exercise tests (Lindsay et al., 2019). The target domain D_t represents the gold-standard measurement dataset BBVS, which is a subset of 181 participants from the Fenland study with directly measured gold-standard VO_2max (Gonzales et al., 2021) during the maximal exercise tests. In the BBVS study, participants need to wear a face mask to measure respiratory gas measurements (Rietjens et al., 2001) to experience exhaustion tests. Further, movement and heart rate are collected using the Actiheart (CamNtech, Papworth, UK) sensor. The details of the two datasets are in Appendix (3)

Training Strategy. We evaluate UDAMA on these two datasets by first pre-training a model on the D_s Fenland. Second, we develop the adversarial training framework with multi-discriminators on the BBVS (D_t), with the help of incorporated prior domain knowledge (i.e., injecting random samples from the source domain). For each domain, feature choices are listed in Appendix C. After the adversarial adaptation, we predict VO_2max on the held-out test set of BBVS using 3-fold cross-validation using UDAMA. Within each fold, the dataset is split into 70% training and 30% testing consisting only of target domain samples.

5.2. Model architecture and tuning

5.2.1. MODEL ARCHITECTURE

This section discusses the details of the neural networks used in this study. For the encoder network, two modules are integrated within the network to extract temporal and metadata

information. In particular, we use two Bidirectional GRU layers of 32 units for the time series data module, followed by one 1D-global averaging pooling layer. On the other hand, in the meta-data module, MLP layers of dimensionality 128 are constructed to extract associated metadata representation after a Batch Normalization layer. Following this, the time and metadata module outputs are concatenated together to generate a complete embedding matrix for the subsequent regression or classification tasks. The training architecture is shown in Figure 4.

The training pipeline consists of two phases using the pre-training and fine-tuning learning scheme. First, after comparing different parameter-sharing techniques for fine-tuning the encoder utilized in the source domain, we freeze the first GRU and MLP layers and fine-tuned the remaining network. Compared with freezing all layers except re-training the output layer, the fine-tuning scheme in our network could capture the general features from the lower layers and extract problem-specific characteristics from higher layers. Lastly, the representation embedding produced from the fine-tuned encoder is transmitted to distinct tasks with a linear activation layer appropriate for predicting the fitness level or classifying the domain labels.

5.2.2. Hyper-paremeter tuning

All network blocks in the framework are trained using Adam with a learning rate tuned over {1e-2, 1e-3}. The dropout rate is tuned over the following ranges {0.2, 0.3}. Moreover, we tune the batch size between 8, 16, and 32 based on the efficiency and stability of the training process. To tune the UDAMA total loss, we conduct a grid search {0.01, 0.02, 0.03} for the α and {(0.9, 0.1), (0.8, 0.2), (0.7, 0.3), (0.6, 0.4), (0.5, 0.5)} for the combination of λ_1 and λ_2. We perform early stopping to combat overfitting until the validation loss stops improving after ten epochs. The details of hyperparameter selection are listed in Table 7.

5.3. Baselines

To verify the effectiveness of our proposed network, we compare the UDAMA against several baselines:

- **In-domain supervised model.** A multi-model network with training on the same domain train and test set.

- **Out-of-domain supervised model**. A network with the same structure as the pre-training model, using wearable data and common biomarkers in D_t as input to predict VO$_2$max.

- **Transfer learning.** A pre-trained model trained on D_s is reused and fine-tuned on the target domain.

- **Autoencoder (Srivastava et al., 2015).** Pre-train a model with stacked recurrent autoencoders on D_s and fine-tune the representation from the encoder to D_t.

- **Deep-Coral (Sun and Saenko, 2016).** A widely employed discrepancy-based domain adaptation minimizes the divergence between the source and target in feature space. In particular, it seeks to align the second-order statistics of the source and target distributions.

- **WDGRL (Shen et al., 2017)**. Wasserstein Distance is used to minimize the disparity between the source and target representations in the feature space. Specifically, the distance will be optimized in an adversarial way for the feature extractor, and domain-invariant features will be learned.

- **Domain Adversarial Neural Networks (DANN) (Ganin et al., 2015)**. A benchmark domain adaption method that uses adversarial training for binary domain classification.

Although most recent domain adaptation methods have shown enhanced performance in dealing with distribution shift through self-training or sophisticated adversarial training scheme (Du et al., 2020; Liu et al., 2021), they do not specifically tackle the regression tasks or healthcare datasets that contain both noisy and gold-standard labels. Therefore, their relevance in the cardio-fitness prediction task is limited.

5.4. Effect of Injected Source Domain Samples

Our proposed approach incorporates prior knowledge from the source domain to create the adversarial training environment. To assess the degree to which the incorporated source domain knowledge impacts our model, we evaluate UDAMA on D_t with different levels of injected samples from (D_s). As such, we put {0.1%, 0.2%, 0.4%, 1%, 2%, 4% } from D_s, which equals to {1%, 5%, 10%, 30%, 50% and 100%} of training data of the target domain (D_t). We set the maximum amount to 100% because the two domains are highly different, and for any larger amount, the model would be trying to predict the silver data rather than the gold data. Specifically, the number of samples differs between source and target domain samples (source/target = 25:1), which means 0.2% of source data equals 5% of target data. If we continually add more source data, for example, 4% of the source (which equals 100% of the target data), we will ultimately train on the source data instead of the target data. Finally, we run the experiments 15 times with different seeds to assess the impact of the injected noisy samples using the average performance on the VO_2max prediction task. Results are shown in §6.3.

5.5. Metrics

The prediction performance of all models is evaluated based on standard regression evaluation metrics such as the Mean Squared Error (MSE) and Mean Absolute Error (MAE). The coefficient of determination (R^2) and the Pearson correlation coefficient (Corr) are also used to evaluate the model performance on the health-related outcome prediction task.

6. Results and Discussion

In this section, we present the results of applying UDAMA to the cardio-fitness prediction task and compare it with the baselines (§6.1). Additionally, we discuss the performance in addressing the domain shift problem (§6.2) and the impact of injected source domain knowledge (§6.3). Furthermore, we conduct an ablation study to verify the effectiveness of our framework's structure (§6.4). Finally, we discuss the model robustness in §6.5 with semi-synthetic data.

Table 1: **Evaluation of different methods on CRF prediction task.** Each result displays the mean value with standard deviation from three-fold cross-validation. In particular, **In-domain** means the model is trained on the same domain, namely trained on D_t and tested on D_t, while **Out-of-domain** corresponds to models trained on Fenland (D_s) and adapted/fine-tuned to BBVS (D_t). All **Out-of-domain** models are evaluated on the BBVS test set.

In-domain	Training method	$\mathbf{R}^2$	Corr	MSE	MAE
$D_t \rightarrow D_t$	*Supervised*	0.123 ± 0.111	0.622 ± 0.036	43.778 ± 6.012	5.263 ± 0.277
Out-of-domain	**Training method**	$\mathbf{R}^2$	**Corr**	**MSE**	**MAE**
$D_s \rightarrow D_t$	*Supervised*	-0.096 ± 0.100	0.007 ± 0.250	58.048 ± 10.061	6.336 ± 0.621
	WDGRL (Shen et al., 2017)	-0.100 ± 0.073	0.004 ± 0.161	55.611 ± 10.61	6.044 ± 0.615
	Autoencoder (Srivastava et al., 2015)	-0.067 ± 0.069	0.127 ± 0.222	53.254 ± 4.878	5.973 ± 0.194
	Deep-Coral (Sun and Saenko, 2016)	0.021 ± 0.073	0.360 ± 0.057	49.044 ± 6.553	5.638 ± 0.374
	Transfer learning (TF)	0.283 ± 0.037	0.621 ± 0.012	35.399 ± 5.910	4.744 ± 0.433
	DANN (Ganin et al., 2015)	0.288 ± 0.077	0.617 ± 0.037	35.458 ± 3.920	4.679 ± 0.382
	UDAMA (ours)	$\mathbf{0.459 \pm 0.063}$	$\mathbf{0.701 \pm 0.032}$	$\mathbf{27.469 \pm 6.456}$	$\mathbf{4.111 \pm 0.353}$

6.1. Fitness prediction

We took 60 participants from the BBVS dataset as test samples to predict their CRF by predicting VO_2max values. The comparison between the proposed domain adaptation framework and baseline approaches is shown in Table 1.

First, in the out-of-domain comparison, adversarial-based DA or transfer learning shows better performance than the discrepancy-based method under label distribution shift. In particular, the discrepancy-based method, Deep-coral, increases the Corr and MAE to 0.36 and 5.638, respectively. Meanwhile, WDGRL aims to minimize the feature difference by employing Wasserstein distance, yielding results comparable to the out-of-domain supervised method, which directly applies the model trained on Fenland for testing BBVS. In contrast, the adversarial-based methods here display better results. DANN learns a representation that is predictive of the regression task but uninformative to the input domain and improves the Corr and MAE to 0.617 and 4.679, respectively. According to Corr and MAE, methods such as transfer learning with fine-tuning techniques also improve performance, achieving 0.621 and 4.744, compared to the discrepancy-based method.

In general, high Corr and R2 values and low MSE and MAE demonstrate the model's ability to leverage noisy, large-scale labeled VO2max data for gold-standard VO2max prediction under label shift. However, the limited size of the test set might result in increased uncertainty during model evaluation. The results from both TF and DANN are similar as shown in Table 1. In contrast, UDAMA utilizing both fine-tuning and adversarial-based domain adaptation methods outperforms all the abovementioned baselines. We observe that the correlation (Corr) outperforms the basic transfer learning methods by 12.9%, the MSE increases by 22.4%, and R^2 improves by 62.2%. Moreover, compared with the in-domain supervised training on $\boldsymbol{D_t}$, UDAMA shows a significant increase, improving R^2 from 0.123 to 0.459. Our method also achieves good performance when generalizing the model from the in-domain to out-of-domain BBVS setting, compared with the dramatic performance drop-down, as shown in Table 1. Therefore, UDAMA can leverage the large-scale noisy datasets
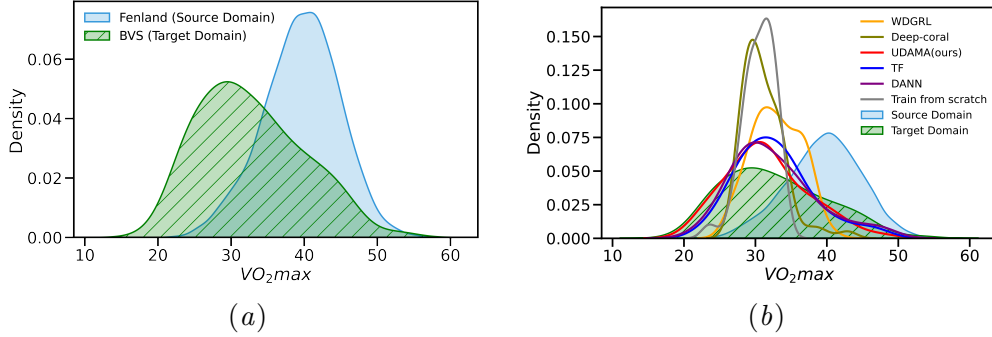
Figure 2: **The distribution of BBVS, Fenland, and distribution of prediction of BBVS test set using different methods**. **(a)** Left figure shows the distribution of BBVS and Fenland datasets. **(b)** Right figure shows the prediction distribution of the BBVS test set from different methods.

information and alleviate the model performance degeneration performance compared with directly validating models on small-scale sensing datasets.

6.2. Domain shift

To better compare whether UDAMA or baseline DA methods can solve the label distribution shift problem effectively. Figure 2 presents the predicted label distribution of the BBVS test set from different methods. First, it shows that the D_s dataset (i.e., Fenland) shares a different VO$_2$max underlying distribution compared to D_t BBVS. Our results demonstrate that UDAMA can learn the small dataset distribution during the adaptation phase and achieves promising results compared with other methods. Besides, we observe that both adversarial-based methods and transfer learning capture the mean and range of target domain distribution as shown in Figure 3(a), whereas the discrepancy-based methods fail to learn the general distribution. We attribute this performance degeneration to the fact that discrepancy-based approaches, mainly designed to minimize the divergence between feature spaces, cannot alleviate the impact of noisy labeling.

Moreover, we use the Hellinger Distance (HD), which calculates the similarity of distributions between prediction and ground truth to examine the distance between two label distributions. Specifically, our framework's prediction of fitness level lies in the same range as the ground truth, while methods like Deep-coral or WDGRL fail at learning within this range. Besides, the distribution of UDAMA ties close compared to baseline methods, where the normalized HD for UDAMA is 0.179, and HD for TF is 0.264, for Deep-coral is 0.305. These results indicate that our framework effectively alleviates the distribution shift problem of the VO$_2$max prediction task and UDAMA can leverage noisy silver-standard data to improve the performance on the gold-standard dataset.

6.3. Impact of Injected Knowledge from Source Domain

Instead of using a generator, we incorporated source domain knowledge to create the adversarial training environment. Figure 3(b) shows the average performance of adding the
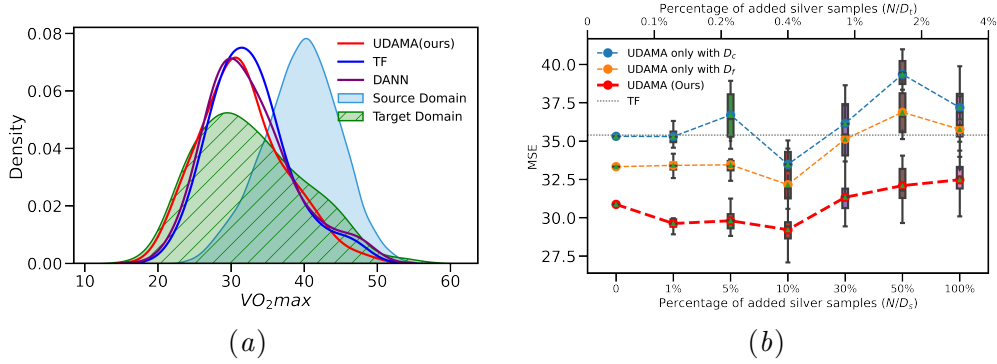
Figure 3: **(a)** Left figure shows UDAMA mitigates distribution shifts.**(b)** Right figure shows the impact of injected source domain samples.

different scales of injected source domain samples to the target domain. Each box plot shows the average MSE results of 15 runs with added random samples from the source domain. As shown in Figure $3(b)$, our method performs better than the baseline methods with added samples in all cases, even after adding 100% of noisy samples of D_t from D_s. Specifically, we observe that it achieves the best results and showcases the most positive transfer by only adding 0.4% of the source domain, which equals 10% of D_t. In contrast, if we continue to add more noisy information to the adaptation stage, the performance will gradually decrease as the source and target domain data reach the ratio of 1:1. After that, the adaptation tends to learn the noisy source domain representation instead of the target domain, yielding a negative transfer. Therefore, only injecting a few samples from the source domain to create the adversarial training environment might help to learn more domain-invariant features and achieve optimal results.

6.4. Ablation Study

Our framework comprises two joint discriminators, so we perform an ablation study to understand each discriminator's effect. Based on the observation that incorporating a 10% amount of BBVS from the D_t Fenland dataset achieves the best transfer, we train different discriminators under this setting. We observe that a single discriminator (coarse-grained or fine-grained) exhibits more competitive performance than the baseline TF or DANN methods, as shown in Table 2 and Figure $3(b)$. Specifically, the fine-grained discriminator alone significantly outperforms the baseline results: +15.8% TF (based on MSE). This indicates that utilizing a fine-grained distribution domain discriminator, which discriminates the domain label distribution, enhances the training framework. Therefore, with the combination of D_f and D_c, UDAMA can learn the cross-domain information without capturing the domain information of the input and alleviate the noisy labeling problem.

Additionally, our method is unique in that it uses wearable-based CRF prediction task, unlike other methods that solely rely on anthropometric data (Nes et al., 2011). Our experiments have demonstrated that combining sensor data with anthropometric information leads to improved prediction accuracy. Although the sensor data alone is unreliable and

Table 2: **Ablation study by removing one of the discriminators.**

Discriminator	R^2	Corr	MSE	MAE
Coarse-grained	0.339 ± 0.053	0.626 ± 0.032	33.485 ± 6.740	4.580 ± 0.337
Fine-grained	0.394 ± 0.030	0.666 ± 0.030	30.578 ± 5.300	4.498 ± 0.324
UDAMA (ours)	$\mathbf{0.459 \pm 0.063}$	$\mathbf{0.701 \pm 0.032}$	$\mathbf{27.469 \pm 6.456}$	$\mathbf{4.111 \pm 0.353}$

Table 3: **Simulated Label Distribution Shift**. Kullback–Leibler divergence (KL divergence) calculates the distribution difference between the shifted D_s and fixed D_t.

Source dataset	KL divergence	UDAMA Corr	DANN Corr
Fenland	0.461	0.701 ± 0.032	0.617 ± 0.037
Left shifted Fenland	1.607	0.656 ± 0.065	0.589 ± 0.027
Right shifted Fenland	3.188	0.646 ± 0.035	0.608 ± 0.037

insufficient for accurate VO2max prediction, when combined with anthropometric data, the accuracy of UDAMA increases from 0.679 (using only anthropometric data) to 0.701 (using anthropometric data and wearable sensor data such as acceleration, heart rate, and heart rate variability). As a result, using anthropometric and low-cost wearable devices together enables more accurate VO$_2$max prediction through UDAMA.

6.5. Robustness assessment with semi-synthetic data shift

Although the distribution shift among gold- and silver-standard labels are prevalent in healthcare applications, there are very few open available datasets with such a challenge where we can apply our method. Therefore, to further evaluate the efficacy of UDAMA, we generate semi-synthetic datasets with various label distribution shifts for the cardio-fitness prediction.

Motivated by the label distribution shift simulation for classification, as seen in (Lipton et al., 2018), we shift the labels in the $\boldsymbol{D_s}$ by a fixed offset and Gaussian noise. By conducting experiments with varying degrees of label shifts, we gain a comprehensive understanding of the effect of label shifts on CRF prediction tasks. As the shift becomes greater from the target domain, the performance decreases. In particular, we show two extreme cases by pushing the source domain shift to left and right to stress-test UDAMA, and the results are shown in Table 3. Despite the increasing KL divergence, which indicates a greater deviation from the ground truth data, our method still displays robust performance compared to the baseline DA method, especially in the case of low fitness on the left side. These stress tests highlight the versatility and robustness of our model in dealing with different input distributions.

7. Discussion

While deep learning has strong potential in numerous healthcare applications, the quality and size of labeled datasets remain a significant problem due to the high cost and labor-intensive nature of data collection. Consequently, the majority of high-quality datasets

with gold-standard labels are small-scale, making it challenging to train efficient and generalizable models. Using large-scale wearable sensing datasets with inaccurate labeling could enhance the performance of such tiny datasets' models. However, distribution shifts, such as cohort and task shifts, may also contribute to performance degradation during model generalization and validation. In this work, we proposed the unlabeled domain adaptation via multi-discriminator adversarial training framework (UDAMA) to address the problem by leveraging a large-scale noisy silver-standard dataset. Our proposed method alleviates the domain shift problem and improves the performance of the challenging CRF prediction task. Moreover, our method has substantial potential to be applied to various other healthcare scenarios that demonstrate similar characteristics. While we have not specifically explored these applications due to limited benchmark datasets, the robust performance of our method under various distribution shifts using diverse datasets, containing both real and semi-synthetic data between gold-standard and noisy labels, shows the proposed method has the ability to generalize. We believe that the proposed method is applicable to a multitude of healthcare domains such as sleep monitoring, continuous monitoring of chronic diseases, and mental health research

Acknowledgments

References

Uk biobank cardio assessment manual version 1.0. https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/Cardio.pdf, 2011.

Fatih Abut and Mehmet Akay. Machine learning and statistical methods for prediction of maximal oxygen uptake: Recent advances. *Medical Devices: Evidence and Research*, 08 2015. doi: 10.2147/MDER.S57281.

Fatih Abut, Mehmet Fatih Akay, and James George. Developing new vo2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection. *Computers in biology and medicine*, 79:182–192, 2016.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks, 2016. URL https://arxiv.org/abs/1608.06019.

Soren Brage, N Brage, P.W. Franks, Ulf Ekelund, and N.J. Wareham. Brage s, brage n, franks pw, ekelund u, wareham nj. reliability and validity of the combined heart rate and movement sensor actiheart. eur j clin nutr 59, 561-570. *European journal of clinical nutrition*, 59:561–70, 04 2005. doi: 10.1038/sj.ejcn.1602118.

Antoine de Mathelin, Guillaume Richard, Francois Deheeger, Mathilde Mougeot, and Nicolas Vayatis. Adversarial weighting for domain adaptation in regression, 2020. URL https://arxiv.org/abs/2006.08251.

Yuntao Du, Zhiwen Tan, Qian Chen, Xiaowen Zhang, Yirong Yao, and Chongjun Wang. Dual adversarial domain adaptation, 2020. URL https://arxiv.org/abs/2001.00153.

Michael Esco, Emmanual Mugu, H. Williford, Aindrea Mchugh, and Barbara Bloomquist. Cross-validation of the polar fitness testtm via the polar f11 heart rate monitor in predicting vo2max. *Journal of Exercise Physiology*, 14:43–52, 10 2011.

Maria Faurholt-Jepsen, Søren Brage, Lars Vedel Kessing, and Klaus Munkholm. State-related differences in heart rate variability in bipolar disorder. *Journal of psychiatric research*, 84:169–173, Jan 2017. ISSN 1879-1379. doi: 10.1016/j.jpsychires.2016.10.005. URL https://pubmed.ncbi.nlm.nih.gov/27743529. 27743529[pmid].

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. 2015. doi: 10.48550/ARXIV.1505.07818. URL https://arxiv.org/abs/1505.07818.

Tomas Gonzales, Justin Jeon, Timothy Lindsay, Kate Westgate, Ignacio Perez-Pozuelo, Stefanie Hollidge, Katrien Wijndaele, Kirsten Rennie, Nita Forouhi, Simon Griffin, Nick Wareham, and Soren Brage. Resting heart rate as a biomarker for tracking change in cardiorespiratory fitness of uk adults: The fenland study, 07 2020a.

Tomas Gonzales, Kate Westgate, Stefanie Hollidge, Tim Lindsay, Justin Jeon, and Soren Brage. Estimating maximal oxygen consumption from heart rate response to submaximal ramped treadmill test, 02 2020b.

Tomas I. Gonzales, Kate Westgate, Tessa Strain, Stefanie Hollidge, Justin Jeon, Dirk L. Christensen, Jorgen Jensen, Nicholas J. Wareham, and Søren Brage. Cardiorespiratory fitness assessment using risk-stratified exercise testing and dose-response relationships with disease outcomes. *Scientific Reports*, 11(1):15315, Jul 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-94768-3. URL https://doi.org/10.1038/s41598-021-94768-3.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. 13(null):723–773, mar 2012. ISSN 1532-4435.

André Henriksen, Martin Mikalsen, Ashenafi Woldaregay, Miroslav Muzny, Gunnar Hartvigsen, Laila Hopstock, and Sameline Grimsgaard. Using fitness trackers and smartwatches to measure physical activity in research: Analysis of consumer wrist-worn wearables. *Journal of Medical Internet Research*, 20:e110, 03 2018. doi: 10.2196/jmir.9157.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation, 2017. URL https://arxiv.org/abs/1711.03213.

Natasha Jaques, Ognjen (Oggi) Rudovic, Sara Taylor, Akane Sano, and Rosalind Picard. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In Neil Lawrence and Mark Reid, editors, *Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, volume 66 of

Proceedings of Machine Learning Research, pages 17–33. PMLR, 20 Aug 2017. URL https://proceedings.mlr.press/v66/jaques17a.html.

Kurt Jensen, Morten Frydkjær, Niels MB Jensen, Lucas M Bannerholt, and Søren Gam. A maximal rowing ergometer protocol to predict maximal oxygen uptake. *International Journal of Sports Physiology and Performance*, 16(3):382–386, 2021.

Stephen Kaptoge, Lisa Pennells, Dirk De Bacquer, Marie Therese Cooney, Maryam Kavousi, Gretchen Stevens, Leanne Margaret Riley, Stefan Savin, Taskeen Khan, Servet Altay, Philippe Amouyel, Gerd Assmann, Steven Bell, Yoav Ben-Shlomo, Lisa Berkman, Joline W. Beulens, Cecilia Björkelund, Michael Blaha, Dan G. Blazer, Thomas Bolton, Ruth Bonita Beaglehole, Hermann Brenner, Eric J. Brunner, Edoardo Casiglia, Parinya Chamnan, Yeun-Hyang Choi, Rajiv Chowdry, Sean Coady, Carlos J. Crespo, Mary Cushman, Gilles R. Dagenais, Ralph B. D'Agostino Sr, Makoto Daimon, Karina W. Davidson, Gunnar Engström, Ian Ford, John Gallacher, Ron T. Gansevoort, Thomas Andrew Gaziano, Simona Giampaoli, Greg Grandits, Sameline Grimsgaard, Diederick E. Grobbee, Vilmundur Gudnason, Qi Guo, Hanna Tolonen, Steve Humphries, Hiroyasu Iso, J. Wouter Jukema, Jussi Kauhanen, Andre Pascal Kengne, Davood Khalili, Wolfgang Koenig, Daan Kromhout, Harlan Krumholz, T. H. Lam, Gail Laughlin, Alejandro Marín Ibañez, Tom W. Meade, Karel G. M. Moons, Paul J. Nietert, Toshiharu Ninomiya, Børge G. Nordestgaard, Christopher O'Donnell, Luigi Palmieri, Anushka Patel, Pablo Perel, Jackie F. Price, Rui Providencia, Paul M. Ridker, Beatriz Rodriguez, Annika Rosengren, Ronan Roussel, Masaru Sakurai, Veikko Salomaa, Shinichi Sato, Ben Schöttker, Nawar Shara, Jonathan E. Shaw, Hee-Choon Shin, Leon A. Simons, Eleni Sofianopoulou, Johan Sundström, Henry Völzke, Robert B. Wallace, Nicholas J. Wareham, Peter Willeit, David Wood, Angela Wood, Dong Zhao, Mark Woodward, Goodarz Danaei, Gregory Roth, Shanthi Mendis, Oyere Onuma, Cherian Varghese, Majid Ezzati, Ian Graham, Rod Jackson, John Danesh, and Emanuele Di Angelantonio. World health organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *The Lancet Global Health*, 7(10):e1332–e1345, Oct 2019. ISSN 2214-109X. doi: 10.1016/S2214-109X(19)30318-3. URL https://doi.org/10.1016/S2214-109X(19)30318-3.

Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis, 2020.

Wouter M. Kouw and Marco Loog. An introduction to domain adaptation and transfer learning, 2018. URL https://arxiv.org/abs/1812.11806.

Jari A. Laukkanen, Timo A. Lakka, Rainer Rauramaa, Raimo Kuhanen, Juha M. Venäläinen, Riitta Salonen, and Jukka T. Salonen. Cardiovascular Fitness as a Predictor of Mortality in Men. *Archives of Internal Medicine*, 161(6):825–831, 03 2001. URL https://doi.org/10.1001/archinte.161.6.825.

Tim Lindsay, Kate Westgate, Katrien Wijndaele, Stefanie Hollidge, Nicola Kerrison, Nita Forouhi, Simon Griffin, Nick Wareham, and Soren Brage. Descriptive epidemiology of physical activity energy expenditure in uk adults (the fenland study). *In-*

ternational Journal of Behavioral Nutrition and Physical Activity, 16, 12 2019. doi: 10.1186/s12966-019-0882-6.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. *CoRR*, abs/1802.03916, 2018. URL http://arxiv.org/abs/1802.03916.

Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *CoRR*, abs/2103.03571, 2021. URL https://arxiv.org/abs/2103.03571.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks, 2015. URL https://arxiv.org/abs/1502.02791.

Ke Mei, Chuang Zhu, Lei Jiang, Jun Liu, and Yuanyuan Qiao. Cross-stained segmentation from renal biopsy images using multi-level adversarial learning, 2020. URL https://arxiv.org/abs/2002.08587.

Bjarne Nes, Imre Janszky, Lars Vatten, Tom Nilsen, Stian Aspenes, and Ulrik Wisloff. Estimating v̇o2peak from a nonexercise prediction model: The hunt study, norway. *Medicine and science in sports and exercise*, 43:2024–30, 04 2011. doi: 10.1249/MSS.0b013e31821d3f6f.

Novi Patricia and Barbara Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1442–1449, 2014. doi: 10.1109/CVPR.2014.187.

Ignacio Perez-Pozuelo, Dimitris Spathis, Emma Clifton, and Cecilia Mascolo. *Wearables, smartphones, and artificial intelligence for digital phenotyping and health*, pages 33–54. 01 2021. ISBN 9780128200773. doi: 10.1016/B978-0-12-820077-3.00003-1.

Christian S. Perone, Pedro Ballester, Rodrigo C. Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling, 2018. URL https://arxiv.org/abs/1811.06042.

Guy Plasqui and Klaas Westerterp. Accelerometry and heart rate as a measure of physical fitness. *Medicine and science in sports and exercise*, 38:1510–4, 08 2006. doi: 10.1249/01.mss.0000228942.55152.84.

Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *CoRR*, abs/1811.12808, 2018. URL http://arxiv.org/abs/1811.12808.

G. Rietjens, H Kuipers, A. Kester, and HA Keizer. Validation of a computerized metabolic measurement system (oxycon-pro®) during low and high intensity exercise1. *International journal of sports medicine*, 22:291–4, 06 2001. doi: 10.1055/s-2001-14342.

Sherif Sakr, Radwa El Shawi, Amjad Ahmed, Waqas Qureshi, Clinton Brawner, Steven Keteyian, Michael Blaha, and Mouaz Al-Mallah. Using machine learning on cardiorespiratory fitness data for predicting hypertension: The henry ford exercise testing (fit) project. *PloS one*, 13:e0195344, 04 2018. doi: 10.1371/journal.pone.0195344.

Anna Shcherbina, C. Mattsson, Daryl Waggott, Heidi Salisbury, Jeff Christle, Trevor Hastie, Matthew Wheeler, and Euan Ashley. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of Personalized Medicine*, 7:3, 05 2017. doi: 10.3390/jpm7020003.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation, 2017. URL https://arxiv.org/abs/1707.01217.

Afshin Shoeibi, Navid Ghassemi, Marjane Khodatars, Mahboobeh Jafari, Sadiq Hussain, Roohallah Alizadehsani, Parisa Moridian, Abbas Khosravi, Hossein Hosseini-Nejad, Modjtaba Rouhani, A. Zare, Saeid Nahavandi, Dipti Srinivasan, Amir Atiya, and U Rajendra Acharya. Epileptic seizure detection using deep learning techniques: A review, 07 2020.

Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J. Wareham, and Cecilia Mascolo. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*. ACM, apr 2021. doi: 10.1145/3450439.3451863. URL https://doi.org/10.1145%2F3450439.3451863.

Dimitris Spathis, Ignacio Perez-Pozuelo, Tomas I. Gonzales, Yu Wu, Soren Brage, Nicholas Wareham, and Cecilia Mascolo. Longitudinal cardio-respiratory fitness prediction through wearables in free-living environments, 2022. URL https://arxiv.org/abs/2205.03116.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms, 2015. URL https://arxiv.org/abs/1502.04681.

Oliver Stegle, Sebastian V. Fallert, David J. C. MacKay, and Soren Brage. Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering*, 55(9):2143–2151, 2008. doi: 10.1109/TBME.2008.923118.

Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. *CoRR*, abs/1607.01719, 2016. URL http://arxiv.org/abs/1607.01719.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance, 2014. URL https://arxiv.org/abs/1412.3474.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation, 2017. URL https://arxiv.org/abs/1702.05464.

Rahul Venkataramani, Hariharan Ravishankar, and Saihareesh Anamandra. Towards continuous domain adaptation for healthcare, 2018. URL https://arxiv.org/abs/1812.01281.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, oct 2018a. doi: 10.1016/j.neucom.2018.05.083. URL https://doi.org/10.1016%2Fj.neucom.2018.05.083.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *CoRR*, abs/1802.03601, 2018b. URL http://arxiv.org/abs/1802.03601.

Thomas White, Kate Westgate, Nicholas Wareham, and Soren Brage. Estimation of physical activity energy expenditure during free-living from wrist accelerometry in uk adults. *PLOS ONE*, 11:e0167472, 12 2016. doi: 10.1371/journal.pone.0167472.

Özal Yildirim, Ulas Baloglu, and U Rajendra Acharya. A deep learning model for automated sleep stages classification using psg signals. *International journal of environmental research and public health*, 16:599, 02 2019. doi: 10.3390/ijerph16040599.

Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/717d8b3d60d9eea997b35b02b6a4e867-Paper.pdf.

Mo Zhou, Yoshimi Fukuoka, Ken Goldberg, Eric Vittinghoff, and Anil Aswani. Applying machine learning to predict future adherence to physical activity programs. *BMC Medical Informatics and Decision Making*, 19, 08 2019. doi: 10.1186/s12911-019-0890-0.
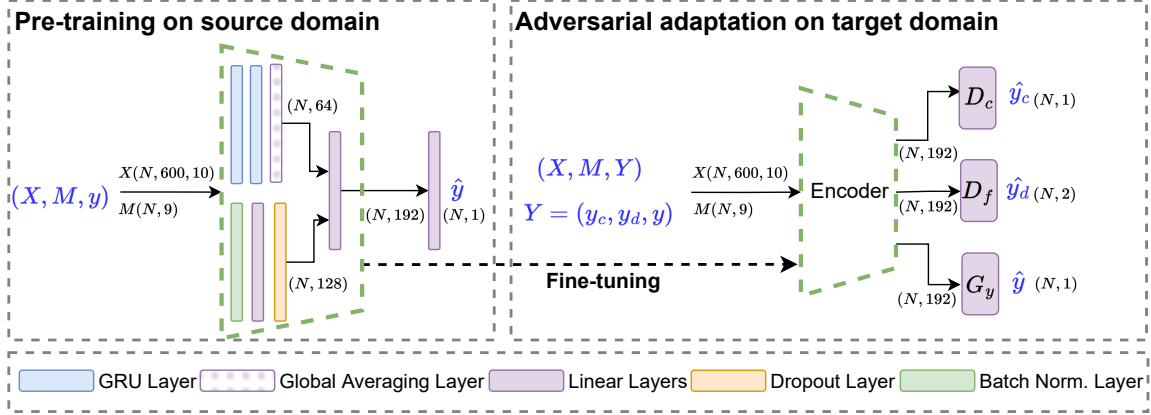
Figure 4: **UDAMA detailed model architecture.** Visualization of the detailed network structure and the input and output dimensionality for each module.

Appendix A. Notation Table and Training Structure

We provide the details of the notation table and training structure here.

Table 4: **Notation.**

Notation	Description
D_s	source domain with noisy (silver-standard) labels
D_t	target domain with high-quality (gold-standard) labels
$X \in \mathbb{R}^{N \times T \times F}$	input time-series sequences
$M \in \mathbb{R}^{N \times F}$	input user metadata
N	number of samples
T	length of input sequences
F	number of features
$y_c \in \mathbb{R}^N$	categorical value representing the coarse-grained binary domain label
$y_d \in \mathbb{R}^N$	numerical value that denotes the fine-grained domain distribution label
$y \in \mathbb{R}^N$	numerical scalar for task target value
D_c	coarse-grained domain discriminator
D_f	fine-grained domain discriminator
G_y	regression predictor
E	feature encoder

Appendix B. Evaluation

First, we describe the cohort selection of the VO$_2$max prediction task (§B.1). Then, we discuss the data extraction and different measurement VO$_2$max (§B.2) and feature extraction (§B.3) stage that occurs before the construction of the deep neural network.

B.1. Cohort Selection

Two datasets, Fenland and BVS, are utilized for experiments and evaluations in our work. Table 5 indicates the descriptive characteristics of the two analysis samples. In addition, the mean and standard deviations for each characteristic are presented in this table.

Table 5: **Characteristics for study analytical samples: Fenland and BBVS study.**

	Fenland Study		BBVS	
	Men (n= 5229)	Women (n= 5830)	Men (n= 98)	Women (n= 83)
	mean std	*mean std*	*mean std*	*mean std*
Demographics				
Age (years)	47.70 7.57	47.66 7.36	53.59 7.31	54.39 6.63
Anthropometrics				
Height (m)	1.78 0.07	1.64 0.06	1.79 0.07	1.64 0.06
Body mass (kg)	85.85 13.83	70.54 13.92	84.63 10.15	69.31 10.95
BMI (kg/m2)	27.16 3.97	26.17 4.97	26.54 2.94	25.72 3.44
Physical activity				
MVPA (min/day)	35.87 22.35	34.40 22.59	40.97 25.23	41.73 22.26
VPA (min/day)	3.27 8.57	3.31 15.67	5.94 12.61	4.21 8.76
Resting Heart Rate				
RHR (bpm)	61.48 8.68	64.46 8.28	59.60 8.06	61.91 6.93
Cardiorespiratory fitness				
VO2max (ml O2/min/kg)	41.95 4.61	37.44 4.73	35.69 6.99	29.60 5.80

B.1.1. FENLAND STUDY

The Fenland dataset with large-scale weakly-labeled VO$_2$max is used as the source domain (D_s) in this work. In particular, Fenland study (Lindsay et al., 2019) is a prospective population-based cohort study of individuals aged 35-65 investigating the interaction between environmental and genetic factors in determining obesity, type 2 diabetes, and related metabolic disorders. The study has collected from 12,435 participants around Cambridgeshire in the East of England between 2005 and 2015. After a baseline clinic visit, participants were instructed to wear the chest sensor Actiheart for six consecutive days to collect the heart rate and movement data. Data from 11,059 participants were included in this study following the exclusion of those with insufficient or corrupt data or missing variables based on §B.3.

B.1.2. BIOBANK VALIDATION STUDY (BBVS)

BBVS dataset is the target domain (D_t) with clean gold-standard VO$_2$max labels (Gonzales et al., 2021). The BBVS dataset is a subset of 191 participants from the Fenland study and aims to validate CRF measurement methods for the UKB-CRF test. During the study, similar electrocardiography (Cardiosoft) and an Actiwave CARDIO device (CamNtech, Papworth, UK) were utilized to collect signals. In particular, participants need to wear face masks with a computerized metabolic system to measure the VO$_2$max during the study. Similarly, data from 181 participants were included after excluding participants with insufficient data. All participants from two datasets provided written informed consent, and the University of Cambridge Ethics Committee approved the study.

B.2. Data extraction and Cardiorespiratory fitness assessment

For both Fenland and BBVS study, participants wore the two standard Electrocardiography (ECG) electrodes Actiheart devices attached to the chest, which measured heart rate and movement recording at 60-second intervals (Brage et al., 2005) for the wearable free-living sensing. The monitoring phase lasted for a total of six days.

In the Fenland study, VO$_2$max was measured using a previously validated submaximal treadmill test(Gonzales et al., 2020b) and considered as weak (silver-standard) labeling [1]. During a lab visit, all participants conducted a treadmill test to determine their heart rate response to a submaximal test, informing their VO$_2$max by using a linear regression method. HR was recorded for 15 minutes during the exercise, and RHR was calculated as the mean heart rate measured during the last 3 minutes. The RHR is a mix of the RHR mentioned above and the Sleeping HR as recorded by the ECG during the free-living period. Although using the submaximal test and linear regression capture the fitness level, measurement bias from -3.0 to -1.6 ml O2/min/kg and a Pearson's r ranging from 0.57 to 0.79.

In comparison, VO$_2$max of participants in the BBVS dataset were directly measured during a maximal exercise test [2]which was completed to exhaustion (Gonzales et al., 2021), . Following standardized techniques used for the UKB-CRF test (CRF, 2011), participants did 5 maximal exercise tests to elicit VO$_2$max, including a UKB flat test, two UKB ramped tests with varying ramp rates and a steady-state test (exclusive to the validation study), and another ramped test (for validation only). After that, the gold-standard VO$_2$max data was collected. For BBVS participants, similar sensor and anthropometric data were collected in the same manner as in Fenland.

B.3. Feature Extraction

All heart rate data collected during lab-visit underwent pre-processing for noise filtering (Stegle et al., 2008). If participants had fewer than 72 hours of concurrent wear data (three full days of recording) or inadequate individual calibration data, they were eliminated from the study (treadmill test-based data). Non-wear periods were excluded from the analyses through non-wear detection procedures. This pre-processing algorithm discovered lengthy durations of non-physiological heart rate and extended periods of no movement

---

1. https://www.mrc-epid.cam.ac.uk/research/studies/fenland/fenland-study-phase-1/

2. https://www.mrc-epid.cam.ac.uk/research/research-areas/physical-activity/

reported by the device's accelerometer (>90 minutes). We used the calculation 1 MET = 71 J/min/kg (3.5 ml O2 min1 kg1) to convert movement intensities into standard metabolic equivalent units (METs). These conversions were then used to classify intensity levels, with behaviors less than 1.5METs classed as sedentary, those between 3 and 6 METs as moderate to vigorous physical activity (MVPA), and those greater than 6METs as vigorous physical activity (VPA). Since time can greatly impact physical activities, we encoded the sensor timestamps using cyclical temporal features (Spathis et al., 2021). Additionally, given the sensors' high sampling rate (1 sample/minute) after matching the HR and Acceleration modalities, learning patterns from such a lengthy sequence (a week's worth of sensor data contains more than 10,000 time steps) is unfeasible, even with the most powerful sequence models. Therefore, instead of segmenting the data into time windows for temporal data processing, we downsampled the wearable signals (Accelerometer and ECG) by a ratio of 15 to decrease the sequence length to 600 time steps. After that, we normalized the data by performing min-max scaling on all input features (sequence-wise for sensor data and column-wise for metadata). Then, each feature vector with 19 features combining time series and metadata were put into various deep neural networks. A detailed view of these features is provided Table 6 in Appendix C.

Appendix C. Hyper-parameters

In this section, we provide the details of the feature choices of the input feature vectors X and M. A detailed view of these features is provided in Table 6. Further, we provide the details of the hyperparameters in Table 7.

Table 7: **Hyper-parameter tuning**.

Parameter	Search space	Selected value
Dropout	{0.2,0.3}	0.3
Optimizer	Adam	Adam
Learning rate	{1e-2, 1e-3}	1e-3
Epochs	[0,100]	Early stopping
Batch size	{8,26,32}	8
α	{0.01,0.02,0.03}	0.01
λ	{(0.9, 0.1), (0.8, 0.2), (0.7, 0.3), (0.6, 0.4), (0.5, 0.5)}	(0.9,0.1)

Table 6: **Description of the features/variables used in our analysis as inputs to the models.** The features with asterisks(*****) are time-series and remaining are metadata. The final set of features is 19.

Features/Variables	Description
Sensors	
Acceleration*	Acceleration measured in mg
Heart rate (HR)*	Mean HR resampled in 15sec intervals, measured in BPM
Heart Rate Variability (HRV)*	HRV calculated by differencing the second-shortest and the second-longest inter-beat interval (as seen in Faurholt-Jepsen et al. (2017)), measured in ms
Acceleration-derived Euclidean Norm Minus One (ENMO)*	ENMO-like variable (Acceleration/0.0060321) + 0.057 (as seen in White et al. (2016))
Acceleration-derived Metabolic Equivalents of Task (METs)*	
Sedentary*	If Accelerometer <1, take daily count and average
Moderate to Vigorous*	If Accelerometer >= 1, take daily count and average
Vigorous*	If Accelerometer >= 4.15, take daily count and average
Anthropometrics	
Age	Age, measured in years
Sex	Sex is binary (female/male)
Weight	Weight, measured in kilograms
Height	Height, measured in meters.centimeters
Body Mass Index (BMI)	BMI is calculated by Weight/(Height^2), measured in kg/m^2
Resting Heart Rate	
Wearable-derived RHR	RHR is calculated by averaging the 4th, 5th, and 6th minute of the baseline visit and adding to that the Sleeping Heart Rate that has been inferred by the wearable device. Gonzales et al. (2020a)
Seasonality	
Month of year	The month number is used along with a coordinate encoding that allows the models to make sense of their cyclical sequence.