RDPD: Rich Data Helps Poor Data via Imitation

Shenda Hong [1 2] **Cao Xiao** [3] **Trong Nghia Hoang** [4] **Tengfei Ma** [4] **Hongyan Li** [1] **Jimeng Sun** [2]

Abstract

In many situations, we need to build and deploy separate models in related environments with different data qualities. For example, an environment with strong observation equipments (e.g., intensive care units) often provides high-quality multi-modal time series data, which are acquired from multiple sensory devices and have rich-feature representations. On the other hand, an environment with poor observation equipments (e.g., at home) only provides low-quality, uni-modal time series data with poor-feature representations. To deploy a competitive model in *poor-data* environment without requiring direct access to multi-modal data acquired from *rich-data* environment, this paper develops and presents a knowledge distillation (KD) method (RDPD) to enhance a predictive model trained on poor data using knowledge distilled from a high-complexity model trained on *rich, private* data. We evaluated RDPD on three real-world time series datasets and shown that its distilled model consistently outperformed all baselines across all datasets, especially achieving the greatest performance improvement over a model trained only on low-quality data by 24.56% on PR-AUC and 12.21% on ROC-AUC, and over that of a state-of-the-art KD model by 5.91% on PR-AUC and 4.44% on ROC-AUC.

1. Introduction

Many *rich-data* environments encompass multiple data modalities. For example, multiple motion sensors in a lab can collect activity time series from various locations of a human body where signals generated from each location can be viewed as one modality. Multiple leads for Electrocardiogram (ECG) signals in hospital are used for diagnosing heart diseases, of which each lead is considered a modality. Multiple physiological signals are measured in intensive care units (ICU) where each type of measure is a modality. A series of recent studies have confirmed that finding patterns among rich multimodal data can increase the accuracy of diagnosis, prediction, and overall performance of the deep learning models (Xiao et al., 2018).

Despite the promises that rich multimodal data bring us, in practice we have more *poor-data* environments with data from fewer modalities of limited quality. For example, unlike in a *rich-data* environment such as hospitals where patients place multiple electrons to collect 12-lead ECG signals, in everyday home monitoring devices often only measure lead I ECG signal from arms. Although deep learning models often perform well in *rich-data* environment, their performance on *poor-data* environment is less impressive due to limited data modality and lower quality (Salehinejad et al., 2018).

We argue that given both rich- and poor-data from similar contexts, the models built on rich multi-modal data can help improve the other model built on poor data with fewer modalities or even a single modality. For example, a heart disease detection model trained on 12 ECG channels in a hospital can help improve a similar heart disease detection model trained on ECG signals from a single-channel at home.

The recent development of mimic learning or knowledge distillation (Hinton et al., 2015; Ba & Caruana, 2014; Lopez-Paz et al., 2015) has provided a way of transferring information from a complex model (teacher model) to a simpler model (student model). Knowledge distillation or mimic learning essentially compresses the knowledge learned from a complex model into a simpler model that is much easier to deploy. However they often require the same data for teacher and student models. Domain adaptation techniques address the problem of learning models on some source data distribution that generalize to a different target distribution. Deep learning based domain adaptation methods have focused mainly on learning domain-invariant representations (Glorot et al., 2011; Chen et al., 2012; Bousmalis et al., 2016). However they often need to be trained jointly on source and target domain data and are

[1]School of Electronics Engineering and Computer Science, Peking University, China [2]Georgia Institute of Technology, USA [3]Analytics Center of Excellence, IQVIA, USA [4]IBM Research, USA. Correspondence to: Jimeng Sun <jsun@cc.gatech.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

therefore unappealing to the settings when the target data source is unavailable during training.

In this paper, we propose RDPD (Rich Data to Poor Data) to build accurate and efficient models for poor data with the help of rich data. In particular, RDPD transfers knowledge from a teacher model trained on rich data to a student model operating on poor data by directly leveraging multi-modal data in the training process. Given a teacher model along with attention weights learned from multimodal data, RDPD is trained end-to-end for the student model operating on poor data to imitate the behavior (attention imitation) and performance (target imitation) of the teacher model.

In particular, RDPD jointly optimize the combined loss of attention imitation and target imitation. The loss of target imitation can utilize both hard labels from the data and soft labels provided by the teacher model. Here are the main contributions of this work: (1) We formally define the learning task from rich data to poor data, which has many real-world applications including healthcare. (2) We propose RDPD algorithm based on mimic learning, which takes a joint optimization approach to transfer knowledge learned by a teacher model using rich data to help improving a student model trained only on poor data. (3) We show that RDPD consistently outperformed all baselines across multiple time series datasets and achieve the greatest performance improvement over the Direct model and the standard distillation model.

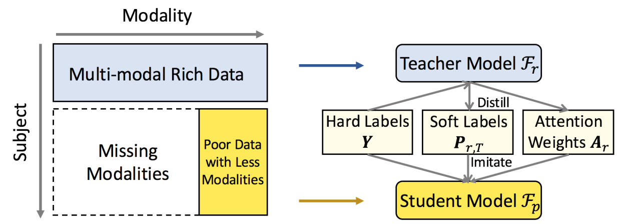

Figure 1. The framework of RDPD. Given teacher model along with attention weights learned from rich data, RDPD trains the student model on poor data while imitating the behavior and performance of teacher model. In particular, RDPD jointly optimize the combined loss of attention imitation (behavior) and target imitation (performance). The loss of target imitation also concerns both hard labels from data and soft labels provided by the teacher model.

2. Method

The design of RDPD is shown in Fig. 1. Mathematically, denote $\boldsymbol{X}_r$ as the multi-modal rich data with D_r modalities that is available in training phase, and $\boldsymbol{X}_p$ as the poor data with D_p modalities that is available in both training and testing phases. Here the modalities in $\boldsymbol{X}_p$ are a subset of $\boldsymbol{X}_r$, and $D_p < D_r$; $\boldsymbol{X}_p$ and $\boldsymbol{X}_r$ share the same labels $\boldsymbol{Y}$. Our task is to build a student model $\mathcal{F}_p$ which only takes $\boldsymbol{X}_p$ as input, and will benefit from knowledge transferred from $\boldsymbol{X}_r$.

2.1. Building Teacher Model

Although RDPD can be applied on time series in general, in this paper we only consider regularly sampled continuous time series $\boldsymbol{X}_r$ (e.g., sensor data). Assume a patient has time series from D_r modalities, for time series in each modality with length l, we split $\boldsymbol{X_r} \in \mathbb{R}^{l \times D_r}$ into M segments at length S, thus $l = M \times S$. We denote multi-modal segmented input time series as $\boldsymbol{S_r} \in \mathbb{R}^{M \times S \times D_r}$.

We applied stacked 1-D convolutional neural networks (CNN) on each segment and recurrent neural networks (RNN) across segments. Such a design has been demonstrated to be effective in many previous studies on multi-variate time series modeling (Ordóñez & Roggen, 2016; Choi et al., 2016). In detail, we apply 1-D CNN with mean pooling on each segment $\boldsymbol{s_r}^{(j)} \in \mathbb{R}^{S \times D_r}, j = 1, \cdots, M$ as given by $\boldsymbol{h_r}^{(j)} = Pooling(CNN_{1D}(\boldsymbol{s_r}^{(j)}))$. Parameters including number of filters, filter size and stride in CNN are shared among segments $\boldsymbol{s_r}^{(1)}, \cdots, \boldsymbol{s_r}^{(M)}$, and vary across different datasets. Then, we concatenate all convolved and pooled segments to get $\boldsymbol{H_r} = [\boldsymbol{h_r}^{(1)}, \cdots, \boldsymbol{h_r}^{(M)}]^T \in \mathbb{R}^{M \times K_r}$, where K_r is the number of filters in CNN_{1D}. Next we applied an RNN layer on $\boldsymbol{H_r}$ and denote the output as $\boldsymbol{Q_r}$ such that $\boldsymbol{Q_r} = RNN(\boldsymbol{H_r})$. And $\boldsymbol{Q_r} \in \mathbb{R}^{M \times U_r}$, where U_r is the number of hidden units in RNN layer. Here we use the widely-applied self-attention mechanism (Lin et al., 2017) as it is a natural choice to get better results by taking advantage of the correlations or importance of segments. It also generates attention weights $\boldsymbol{A_r}$ that could represent teacher's behaviors on each segment. The attention weights are calculated by Eq. 1.

$$\boldsymbol{A_r} = softmax(\boldsymbol{Q_r W}) \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{U_r \times 1}$, $\boldsymbol{A_r} \in \mathbb{R}^{M \times 1}$. We then multiplied the RNN output $\boldsymbol{Q_r}$ with corresponding attention weights $\boldsymbol{A_r}$. The weighted output $\boldsymbol{G_r}$ is given by $\boldsymbol{G_r} = \boldsymbol{A_r}^T \boldsymbol{Q_r}$. where $\boldsymbol{G_r} \in \mathbb{R}^{1 \times U_r}$. Finally, the weighted output $\boldsymbol{G_r}$ is further transformed by a dense layer with weights $\boldsymbol{W_d} \in \mathbb{R}^{U_r \times C}$ to output logits $\boldsymbol{O_r} \in \mathbb{R}^{1 \times C}$, $\boldsymbol{O_r} = \boldsymbol{G_r W_d}$ For simplicity, we can summarize the teacher model $\mathcal{F}_r$ as in Eq. 2: $\mathcal{F}_r$ takes $\boldsymbol{X_r}$ as inputs and outputs logits $\boldsymbol{O_r}$ and attention weights $\boldsymbol{A_r}$.

$$\mathcal{F}_r(\boldsymbol{X_r}) \quad = \quad \boldsymbol{A_r}, \boldsymbol{O_r} \tag{2}$$

The objective function of the teacher model measures prediction accuracy, and also provides knowledge to student model. Typically, $\boldsymbol{O_r}$ are transformed by softmax as final predicted probabilities, which can be used as distilled knowledge for student model to imitate. However, sharp distribution (e.g, hard labels) will be less informative. To alleviate this issue, we follow the idea in (Hinton et al., 2015) to produce more informative soft labels. Compared with hard label, the soft label imitation has much smoother probability distribution over classes, thus contains richer (larger entropy) informations. Concretely, we modify classic softmax to $\mathcal{S}(x, T)$ by dividing original logits $\boldsymbol{O_r}$ with a predefined hyper-parameter T (larger than 1). T is usually referred to as Temperature. The modified softmax (shows ith soft probability) is given by $\mathcal{S}(\boldsymbol{x}, T)_i = \frac{\exp(x_i/T)}{\sum_j \exp(x_j/T)}$ and the soft predictions are given by $\boldsymbol{P_{r,T}} = \mathcal{S}(\boldsymbol{O_r}, T)$. Finally, we use cross-entropy loss as prediction loss $\mathcal{L}_{teacher}$ (in Eq. 3) to measure the difference between soft predictions $\boldsymbol{P_{r,T}} \in \mathbb{R}^{1\times C}$ and ground truth $\boldsymbol{Y} \in \mathbb{R}^{1\times C}$. We optimize teacher model via minimizing $\mathcal{L}_{teacher}$.

$$\mathcal{L}_{teacher} = CrossEntropy(\boldsymbol{Y}, \boldsymbol{P_{r,T}}) \tag{3}$$

2.2. Imitating Attentions and Targets

After training teacher model on rich data, we now describe the imitation process for the student model. For attention imitation, we mean to mimic attention weights. For target imitation, the student model imitates the following components: 1) soft label that is more informative, 2) hard label that could improve performance (according to (Hinton et al., 2015)), and 3) a trainable combination of both soft label and hard label. Again, we start with constructing the student model $\mathcal{F}_p$ using a CNN + RNN architecture, but with fewer filters in CNN and fewer hidden units in RNN. Similar to Eq.2, $\mathcal{F}_p$ takes $\boldsymbol{X_p}$ as inputs and outputs logits $\boldsymbol{O_p}$ and attention weights $\boldsymbol{A_p}$ as in Eq. 4.

$$\mathcal{F}_p(\boldsymbol{X_p}) \quad = \quad \boldsymbol{A_p}, \boldsymbol{O_p} \tag{4}$$

Attention Imitation In Eq.1 we define attention weights to represent the influence of different time segments to the final predictions. We assume that the attention behavior of student model should resemble that of teacher model, and formulate the attention imitation as below. Given Eq.2 and Eq.4, to enforce $\boldsymbol{A_p}$ and $\boldsymbol{A_r}$ to have similar distributions, we minimize the Kullback-Leibler (KL) divergence $\mathcal{L}_{att}$ given by Eq. 5 to measure the information loss from distribution of attention in student model $\boldsymbol{A_p}$ to distribution of attention in teacher model $\boldsymbol{A_r}$.

$$\mathcal{L}_{att} \quad = \quad D_{KL}(\boldsymbol{A_p} || \boldsymbol{A_r}) \tag{5}$$

Imitating Hard Labels For hard label imitation, we optimize the student model by minimizing cross entropy loss $\mathcal{L}_{hard}$ (in Eq. 6) that measures the difference between predicted target values and ground truth values $\boldsymbol{Y} \in \mathbb{R}^{1\times C}$, where C is the number of target classes, $\boldsymbol{P_{p,1}} = \mathcal{S}(\boldsymbol{O_p}, 1)$.

$$\mathcal{L}_{hard} = CrossEntropy(\boldsymbol{Y}, \boldsymbol{P_{p,1}}), \tag{6}$$

Imitating Soft Labels Given soft labels from $\mathcal{F}_r$, we produce soft predictions $\boldsymbol{P_{p,T}}$ by the same temperature T on softmax in student model $\mathcal{F}_p$. Then, we optimize a cross entropy loss $\mathcal{L}_{soft}$ (in Eq. 7) that measures the differences between student and teacher.

$$\mathcal{L}_{soft} = T^2 CrossEntropy(\boldsymbol{P_{r,T}}, \boldsymbol{P_{p,T}}) \tag{7}$$

Here, $\boldsymbol{P_{r,T}}$ is defined as before. $\boldsymbol{P_{p,T}} = \mathcal{S}(\boldsymbol{O_p}, T)$. Since the magnitudes of gradients in Eq.7 is scaled by $1/T^2$ as we divided logits by T, we should multiply the soft imitation loss by T^2 to keep comparable gradient during implementation.

Imitating Combined Label While hard labels provide certain prediction outcomes and soft labels provide probabilistic predictions, the two labels may even be opposite. To resolve the gap between the two labels, a reasonable solution is to combine them to yield uncertain prediction (probabilities of each class). Besides, while hard label imitation helps student model learn more information from data, soft label imitation transfer more knowledge from the teacher model (smoother distribution), each will lead to either more bias (comes from data) or more variance (comes from model). To leverage their benefits and make them complement each other, we propose to minimize a linear combination of hard labels and soft labels as $\boldsymbol{P_{p,comb}} = \mathcal{S}(w_1 \boldsymbol{P_{p,1}} + w_2 \boldsymbol{P_{p,T}} + b, 1)$, where w_1, w_2, b are learnable parameters. For the combined imitation, we also use cross entropy loss $\mathcal{L}_{comb}$ (in Eq. 8) to define the loss between $\boldsymbol{P_{p,comb}}$ and ground truth $\boldsymbol{Y}$.

$$\mathcal{L}_{comb} = CrossEntropy(\boldsymbol{Y}, \boldsymbol{P_{p,comb}}) \tag{8}$$

2.3. Joint Optimization

Finally, for the student model to imitate attentions and targets simultaneously, we jointly optimize all loss functions above. Since they are computed using cross entropy loss, and we have to rectify them to get comparable loss values. Here, we simply summed them up to get the final objective function $\mathcal{L}_{student}$ given by Eq. 9.

$$\mathcal{L}_{student} = \mathcal{L}_{att} + \mathcal{L}_{hard} + \mathcal{L}_{soft} + \mathcal{L}_{comb} \tag{9}$$

3. Experiments

Datasets We used time series data including: (1) PAMAP2 Physical Activity Monitoring Data Set (PAMAP2) (Reiss & Stricker, 2012), (2) The PTB Diagnostic ECG Database (PTBDB) (Bousseljot et al., 1995) and (3) The Medical Information Mart for Intensive Care (MIMIC-III) (Johnson et al., 2016) in performance evaluation. Data statistics are summarized in Table 1.

Comparative Methods 1. Teacher: Teacher model is trained and tested on all channels. It serves as an

Table 1. Statistics of Datasets

	PAMAP2	PTBDB	MIMIC-III
# subjects	9	290	9,488
# classes	12	6	8
# attributes	52	15	6
Total time series length	2,872,533	59,619,455	455,424
Sample Frequency	100 Hz (IMU) 9 Hz (HR)	1,000 Hz	1 per hour

empirical upper bound of performance. **2. Direct**: Direct model is build on the partially observed data using RCNN, without attention imitation and soft label imitation. This model is equivalent to $\mathcal{L} = \mathcal{L}_{hard}$. **3. Knowledge Distillation (KD)**: KD (Hinton et al., 2015) model is constructed on the partially observed data, with soft label imitation and hard label imitation. This model is equivalent to $\mathcal{L} = \mathcal{L}_{hard} + \mathcal{L}_{soft}$. **4. $\mathbf{RDPD}_{r1}$**: The reduced version of RDPD without attention imitation. And the objective function would be $\mathcal{L} = \mathcal{L}_{comb} + \mathcal{L}_{hard} + \mathcal{L}_{soft}$. **5. $\mathbf{RDPD}_{r2}$**: The reduced version of RDPD without combined labels. This model is equivalent to KD model with attention imitation. And the objective function would be $\mathcal{L} = \mathcal{L}_{att} + \mathcal{L}_{hard} + \mathcal{L}_{soft}$. **6. RDPD**: Our whole model contains all proposed imitations. Using $\mathcal{L} = \mathcal{L}_{att} + \mathcal{L}_{hard} + \mathcal{L}_{soft} + \mathcal{L}_{comb}$ as objective function.

Results We compared the results of RDPD against other baselines and the reduced version of RDPD in Table 2 (PAMAP2 dataset), Table 3 (PTBDB dataset) and Table 4 (MIMIC-III dataset). RDPD outperformed other methods (except Teacher) in most cases and demonstrated the proposed attention imitation and target imitation successfully improved performance of student model. The teacher model performs best among all methods since it is trained using a full datasets with multiple modalities. It serves an empirical upper bound of the performance. In Table 3, RDPD works better than its reduced version in PR-AUC and F1-score but not ROC-AUC. The reason is that classes in PTBDB dataset is very imbalanced, some occasional samples in rare classes distort the final result.

4. Conclusion

In this paper we proposed to leverage the power of rich data to improve the learning from poor data with RDPD. RDPD learns end-to-end for the student model built on poor data to imitate the behavior (attention imitation) and performance (target imitation) of teacher model by jointly optimizing the combined loss of attention imitation and target imitation. We evaluated RDPD across multiple datasets and demonstrated its promising utility and efficacy.

Table 2. Performance comparison on PAMAP2 dataset. The task is multi-class classification (12 classes). All contains 52 channels, Wrist contains 17 channels signals of 1 IMU over the wrist on the dominant arm, Chest contains 17 channels signals of 1 IMU on the chest, Ankle contains 17 channels signals of 1 IMU on the dominant side's ankle.

Data	Method	ROC-AUC	PR-AUC	macro-F1
All	Teacher	0.928 ± 0.014	0.708 ± 0.039	0.608 ± 0.045
Wrist	Direct	0.800 ± 0.032	0.452 ± 0.051	0.376 ± 0.049
	Distill	0.825 ± 0.020	0.469 ± 0.052	0.380 ± 0.060
	RDPD_{r1}	0.837 ± 0.025	0.491 ± 0.037	0.406 ± 0.053
	RDPD_{r2}	0.836 ± 0.018	0.478 ± 0.038	0.401 ± 0.049
	RDPD	$\mathbf{0.838 \pm 0.012}$	$\mathbf{0.491 \pm 0.045}$	$\mathbf{0.425 \pm 0.057}$
Chest	Direct	0.836 ± 0.035	0.519 ± 0.065	0.449 ± 0.069
	Distill	0.868 ± 0.025	0.575 ± 0.043	0.486 ± 0.065
	RDPD_{r1}	0.872 ± 0.028	0.605 ± 0.030	0.518 ± 0.037
	RDPD_{r2}	0.879 ± 0.027	0.600 ± 0.051	0.478 ± 0.048
	RDPD	$\mathbf{0.883 \pm 0.016}$	$\mathbf{0.609 \pm 0.052}$	$\mathbf{0.529 \pm 0.051}$
Ankle	Direct	0.811 ± 0.035	0.513 ± 0.065	0.405 ± 0.080
	Distill	0.901 ± 0.015	0.621 ± 0.044	0.492 ± 0.070
	RDPD_{r1}	0.889 ± 0.021	0.581 ± 0.071	0.443 ± 0.095
	RDPD_{r2}	0.904 ± 0.019	0.629 ± 0.041	0.473 ± 0.069
	RDPD	$\mathbf{0.910 \pm 0.014}$	$\mathbf{0.639 \pm 0.030}$	$\mathbf{0.511 \pm 0.033}$

Table 3. Performance comparison on PTBDB dataset. The task is multi-class classification (6 classes). All contains 15 channels of ECG signals. Lead I contains single channel Lead I ECG signal, which is usually generated by mobile devices.

Data	Method	ROC-AUC	PR-AUC	macro-F1
All	Teacher	0.737 ± 0.035	0.293 ± 0.018	0.288 ± 0.028
Lead I	Direct	0.701 ± 0.023	0.279 ± 0.017	0.164 ± 0.020
	Distill	0.676 ± 0.045	0.282 ± 0.022	0.217 ± 0.016
	RDPD_{r1}	0.677 ± 0.036	0.255 ± 0.029	0.139 ± 0.027
	RDPD_{r2}	$\mathbf{0.707 \pm 0.073}$	0.282 ± 0.044	0.218 ± 0.024
	RDPD	0.706 ± 0.075	$\mathbf{0.293 \pm 0.025}$	$\mathbf{0.218 \pm 0.019}$

Table 4. Performance comparison on MIMIC-III dataset. The task is multi-class classification (8 classes). All contains 6 channels of patient vital signs. BP contains blood pressure systolic and blood pressure diastolic, which is usually monitors by house sphygmomanometer. HR is heart rate, RR is respiration rate.

Data	Method	ROC-AUC	PR-AUC	macro-F1
All	Teacher	0.696 ± 0.011	0.281 ± 0.009	0.256 ± 0.012
BP	Direct	0.610 ± 0.016	0.204 ± 0.011	0.149 ± 0.013
	Distill	0.611 ± 0.013	0.206 ± 0.007	0.150 ± 0.005
	RDPD_{r1}	0.607 ± 0.012	0.203 ± 0.003	0.148 ± 0.003
	RDPD_{r2}	0.613 ± 0.020	0.205 ± 0.009	0.147 ± 0.007
	RDPD	$\mathbf{0.614 \pm 0.018}$	$\mathbf{0.207 \pm 0.010}$	$\mathbf{0.150 \pm 0.006}$
HR	Direct	0.556 ± 0.019	0.176 ± 0.013	0.089 ± 0.042
	Distill	0.564 ± 0.021	0.175 ± 0.012	0.109 ± 0.030
	RDPD_{r1}	0.566 ± 0.010	0.178 ± 0.004	0.132 ± 0.005
	RDPD_{r2}	0.571 ± 0.011	0.176 ± 0.008	0.123 ± 0.016
	RDPD	$\mathbf{0.581 \pm 0.014}$	$\mathbf{0.182 \pm 0.004}$	$\mathbf{0.130 \pm 0.010}$
RR	Direct	0.570 ± 0.019	0.176 ± 0.012	0.109 ± 0.039
	Distill	0.614 ± 0.023	0.201 ± 0.009	0.162 ± 0.015
	RDPD_{r1}	0.611 ± 0.014	0.202 ± 0.007	0.160 ± 0.016
	RDPD_{r2}	0.614 ± 0.017	0.205 ± 0.006	0.169 ± 0.010
	RDPD	$\mathbf{0.619 \pm 0.022}$	$\mathbf{0.207 \pm 0.008}$	$\mathbf{0.169 \pm 0.007}$

References

Ba, J. and Caruana, R. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pp. 2654–2662, 2014.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. Domain separation networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 343–351, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9.

Bousseljot, R., Kreiseler, D., and Schnabel, A. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. *Biomedizinische Technik/Biomedical Engineering*, 40(s1):317–318, 1995.

Chen, M., Xu, Z., Weinberger, K. Q., and Sha, F. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1627–1634, USA, 2012. Omnipress. ISBN 978-1-4503-1285-1.

Choi, K., Fazekas, G., Sandler, M., and Cho, K. Convolutional recurrent neural networks for music classification. *arXiv preprint arXiv:1609.04243*, 2016.

Glorot, X., Bordes, A., and Bengio, Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 513–520, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. A structured self-attentive sentence embedding. *ICLR*, 2017.

Lopez-Paz, D., Bottou, L., Schölkopf, B., and Vapnik, V. Unifying distillation and privileged information. *ICLR*, 2015.

Ordóñez, F. J. and Roggen, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

Reiss, A. and Stricker, D. Creating and benchmarking a new dataset for physical activity monitoring. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 40. ACM, 2012.

Salehinejad, H., Barfett, J., Valaee, S., and Dowdell, T. Training neural networks with very little data - A draft. 2018.

Xiao, C., Choi, E., and Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 2018.