PiRL: Participant-Invariant Representation Learning for Health Care using Wearable Data

Zhaoyang Cao
Department of Computational Applied Mathematics and Operations Research
Rice University
Houston, TX 77030
zc48@rice.edu

Han Yu, Huiyuan Yang, Akane Sano
Department of Electrical and Computer Engineering
Rice University
Houston, TX 77030
{hy29, hy49, akane.sano}@rice.edu

Abstract

Due to the individual heterogeneities among human subjects, researchers observed performance gaps between generic (one-size-fits-all) models and person-specific models in data-driven health applications. However, in real-world applications, generic models are usually more favored due to the factors such as the new-user-adaptation issue, system complexities, etc. To improve the performance of the generic model, we propose a representation learning framework that learns participant-invariant representations, named PiRL. The proposed framework constrains the latent space using maximum mean discrepancy (MMD) to close the distribution gap among subjects. Further, a triplet loss is utilized to optimize the learned representations for downstream health applications. We evaluate our frameworks on two public datasets for human physical and mental health problems detecting sleep apnea and stress, respectively. As preliminary results, we found the proposed approach shows around a 5% increase in accuracy with statistical differences compared to baseline.

1 Introduction

With the development of wearable devices, researchers have been developing deep learning models using time-series wearable sensor data for solving health-related problems. For example, for heart diseases, Oh *et al.* proposed an automated system that combines a convolutional neural network (CNN) and long short-term memory network (LSTM) for the diagnosis of arrhythmia [14]. Erdenebayar *et al.* designed a deep neural network, recurrent neural networks, and gated-recurrent unit to distinguish apnea and hypopnea events using an electrocardiogram (ECG) signal [3]. Additionally, for mental health, Yu and Sano applied semi-supervised learning on leveraging unlabeled data to estimate the wearable-based momentary stress [22]. Radhika *et al.* proposed the frameworks that investigate the effectiveness of transfer learning and deep multimodal fusion on CNN stress detection models [16] [17].

However, due to the heterogeneity among human data, reflected in both the collected data and labels, generic models' performances are usually unsatisfactory because learned representations used for prediction are associated with subject-specific information. As shown in related works [5, 11, 13, 18, 23] that generic machine learning models perform poorly with low prediction accuracy

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

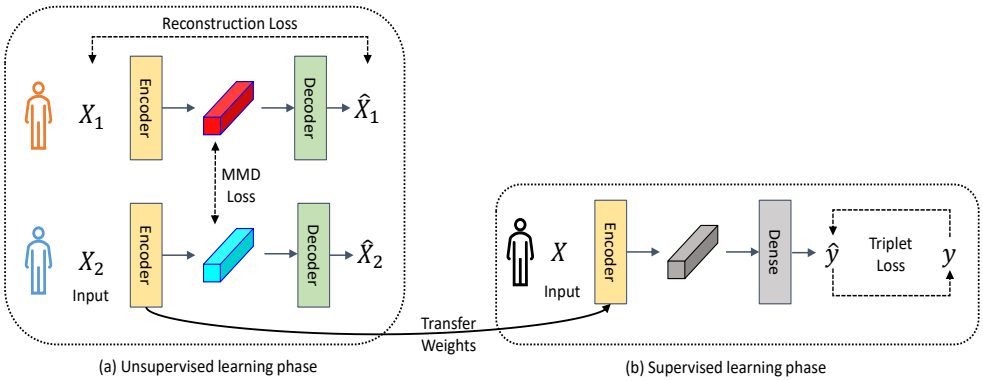

Figure 1: PiRL Network Architecture

[8]. For instance, Bsoul *et al.* showed that the accuracy of the subject-dependent classifier model on recognizing apnea episodes is 6% lower than that of the subject-independent model [2]. Moreover, Nath *et al.* [12] used LSTM network to show that subject-independent model obtains over 90% accuracy in emotion recognition, while subject-dependent model only approach 70% accuracy.

Although person-specific models have been utilized to achieve promising results in health applications [7, 9, 20], we cannot neglect its drawback. Obviously, person-specific models cannot be easily extended to new datasets from the populations the models have not seen yet [13]. Also, it is challenging to develop person-specific models because it is expensive to collect enormous datasets from individuals.

Compared to previous models used on health detection tasks (CNN and LSTM by Oh *et al.*, deep multimodal fusions by Radhika *et al.*), we proposed a novel representation learning method with penalized MMD loss and triplet loss. Similar studies had been done by some researchers, like Bethge *et al.*, who utilized MMD loss only to impose domain-invariance representations. Although the classifier is shared for all sample data, each participant has its private encoder in Bethge *et al.*'s research. However, to avoid the personalized information and extract more generic representations in our study, all participants used identical encoders.

In this work, we aim to improve the performances of the generic models by proposing a representation learning method to learn participant-invariant features. The proposed framework alleviates the heterogeneous issues by integrating the maximum mean discrepancy loss in representation learning to minimize the distribution shifts among features from different subjects. Further, a triplet loss, which aims to learn the label-distinguishable embedding, is considered in part of the framework as another constraint to avoid trivial solutions in learned representations. We evaluate the proposed method using two public datasets, including downstream tasks: sleep apnea detection and stress detection for physical health and mental health problems, respectively. Regarding the generic model without any penalized loss function as the baseline, the results suggest that the proposed method can help improve the model performance significantly compared to the baseline results.

2 Methodology

We propose a representation learning framework that aims to extract participant-invariant representations in this study, named PiRL. The main PiRL framework is visualized in Figure 1 and detailed architecture is shown in Appendix A. We employ a 1D CNN-based auto-encoder structure as a deep feature extractor from the input wearable data. On top of the auto-encoder, an MMD loss is utilized to constrain the representations from distribution shifts to encourage the learning embedding to be participant-invariant. During training for downstream tasks, we optimize the model with a triplet loss for label-distinguishable representations. The following subsections will introduce the aforementioned components in detail.

2.1 Representation learning

We utilize a 1D CNN-based autoencoder to extract learning representations from raw time-series sequences X. The encoder extracts latent representations from input sequences as e with a series of 1D CNN layers; whereas the decoder aims to output the reconstructed signal as $\hat{X}$ from e using up-sampling layers. The objective of the autoencoder is:

$$\mathcal{L}_{ae} = \|X - \hat{X}\|_2^2 \tag{1}$$

2.1.1 MMD loss

To encourage the model learns the participant-invariant representations, we constrain the model with an MMD loss function, which is widely used in eliminating distribution shifts among different groups of data [6].

Given training samples from two different subjects as X_i and X_j with a total subject number of N, the MMD loss can be considered as follows:

$$\mathcal{L}_{mmd}(p, q, \mathcal{H}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sup_{f \in \mathcal{H}, ||f||_{\mathcal{H}} \leq 1} (\mathop{\mathbb{E}}_{p(X_i)} [f(X_i)] - \mathop{\mathbb{E}}_{q(X_j)} [f(X_j)])$$

where p and q are the distribution of variable X_i and X_j, and $\mathcal{H}$ is the Hilbert space. During the training process, the model is optimized to minimize the distribution distances between the pairs of subjects. Thus, the overall objective of unsupervised learning is:

$$\mathcal{L} = \mathcal{L}_{ae} + \lambda \cdot \mathcal{L}_{mmd} \tag{2}$$

where λ is the coefficient of MMD loss and set to 0.2.

2.2 Fine-tuning with Triplet Loss

The supervised learning models for downstream tasks are built on top of the learned representations. We design a fully connected dense layer as a classifier that follows the aforementioned CNN-based encoder that is pre-trained with the auto-encoder. Then the supervised learning structure is fine-tuned according to different downstream tasks. Further, we also apply the triplet loss to optimize the representation of the training labels and avoid the trivial solutions learned in the pre-training procedure [19]. The objective of the triplet loss is given below:

$$\mathcal{L}_{triplet} = \max(d(a, p) - d(a, n) + margin, 0) \tag{3}$$

where *a* represents anchor sample data, *p* represents 'positive' sample data with the same label from the anchor, *n* represents 'negative' sample data with the opposite label against the anchor and the margin is a positive scalar. From the equation 3, we can learn that the triplet objective is to decrease the sample distance from the same labels and separate the sample distance with different labels. During the supervised learning phase, the coefficient of triplet loss is set to 0.2.

3 Results and Discussion

We tested the proposed PiRL frameworks on two datasets, including CLAS and Apnea-ECG datasets for the applications in mental health and physical health, respectively. Appendix B included detailed information on two datasets and experimental settings such as hyper-parameters. For each supervised prediction model, we pre-trained and fine-tuned the encoder at the beginning of each epoch. We evaluated the performance of participant-invariant representations by measuring prediction accuracy.

3.1 Stress Detection using Electrodermal Activity (EDA) with CLAS Dataset

Table 1 shows the prediction accuracy in four types of supervised learning models. The MMD loss-based model obtains higher accuracy of 66.5% compared to the baseline results (64.3 %). The prediction accuracy of the triplet loss only and the MMD + triplet loss models both exceeded 70%. The results illustrated that the model with triplet loss works better in stress prediction than the one with MMD loss-based models. To examine the statistical significance of the accuracy, we conducted an ANOVA (post-hoc: Tukey) test, and the corresponding results are also shown in Table 1. The

prediction accuracy of the baseline framework is treated as the reference group. We found that both MMD and triplet loss based models have statistically higher accuracy, but triplet loss improved the model performance most obviously in stress detection.

Table 1: Performance in stress detection on CLAS dataset. P-values are calculated by ANOVA (post-hoc: Tukey)

	Baseline	MMD	Triplet	MMD+Triplet
Accuracy	64.3%	66.5%	70.1%	70.6%
SD	0.012	0.014	0.011	0.010
P-value < 0.01	-	✓	✓	✓

3.2 Sleep Apnea Detection using ECG with Apnea-ECG Dataset

Table 2 shows the prediction accuracy of apnea detection using four types of supervised learning models. The baseline model obtained a prediction accuracy of 75.2% with a standard deviation of 0.014. The accuracy of the remaining three frameworks reached over 79% and they were all statistically significantly higher than the baseline results with less standard deviation. The best framework for detecting apnea is the combination of MMD and triplet loss since it achieved the highest prediction accuracy.

Table 2: Performance in sleep apnea detection on Apena-ECG dataset. P-values are calculated by ANOVA (post-hoc: Tukey)

	Baseline	MMD	Triplet	MMD+Triplet
Accuracy	75.2%	79.5%	79.1%	79.9%
SD	0.014	0.010	0.013	0.009
P-value < 0.01	-	✓	✓	✓

4 Ablation Study

4.1 Person-Specific Models' Performance

Person-specific models' performance was calculated and used to make the comparison between generic models and proposed PiRL frameworks. We separated each participant into a training set and testing set with a ratio of 70%:30% for all participants in the training set from the EDA dataset and Apnea-ECG dataset, separately. The final prediction accuracy of the person-specific models was recorded as the average accuracy from each participant and shown below in Table 3. The accuracy of the person-specific model on the Apnea-ECG dataset is worth noticing. Although the accuracy is higher than 95%, limitations still exist that restricted us to apply to the more general cases because the preliminary results are based on patients and the extracted representations may not be applicable to yet unseen population. The results illustrated that our proposed frameworks partially bridge the performance gap between generic and person-specific models to some extent.

Table 3: Accuracy of person-specific models and generic models

	Generic	MMD	Triplet	Person-Specific
Accuracy on CLAS dataset	64.3%	66.5%	70.1%	86.8%
Accuracy on Apnea-ECG dataset	75.2%	79.5%	79.1%	95.7%

4.2 Domain-Adversarial Training of Neural Networks

Domain-Adversarial Training of Neural Networks (DANN) architecture is a potent tool to constrain participant-invariant features by accomplishing the following adversarial tasks: minimizing the loss of label prediction and maximizing the loss of domain classification. Subsequently, with the addition of gradient reversal layer before the domain classifier, the overall objective function is the sum of two

minimizing problems [4]. In our case, feature extractor refers to auto-encoder, label predictor refers to the supervised learning phase, and we implemented the domain classifier in the training process together with reconstruction loss. In order to generate the domain classification, we treated each participant as an independent domain, with the number of domains equal to the number of participants. Since it is complicated to separate representations into over 40 domains (participants) accurately, the effect of the domain classifier will be weakened and thus, the DANN framework doesn't enhance the prediction performance compared to our baseline model based on the experimental results.

4.3 Latent Space Visualization

Visualization of the latent space is often helpful in demonstrating how the representations perform in a reduced-dimensional space. Hence, we use t-distributed stochastic neighbor embedding (t-SNE) to realize the latent space visualization in this study as it is commonly used for dimension reduction and visualization of high-dimensional datasets [21]. We plot the distributions of learned representations between baseline and MMD loss-based framework in 2-dimensional coordinates to evaluate the effectiveness of the MMD loss function.

Figure 2 is the latent space visualization of invariant representations that aims to further explore the influence of MMD loss on CLAS dataset.By comparing the distribution of baseline and MMD-based representations based on our visualization, the second component's dimension shows a decrease from baseline to MMD between participants, whereas the first component shows no obvious variation in range. In general, as the representations of each participant spread out in distinct clusters, the baseline PiRL framework successfully revealed the heterogeneity across populations. However, for the participant-invariant representation learning, the data from different participants should be entangled and show no collapse in the feature space, but currently, representations are still separable during the visualization [1]. Therefore, we still need to try some other approaches to better implement the latent space visualization.

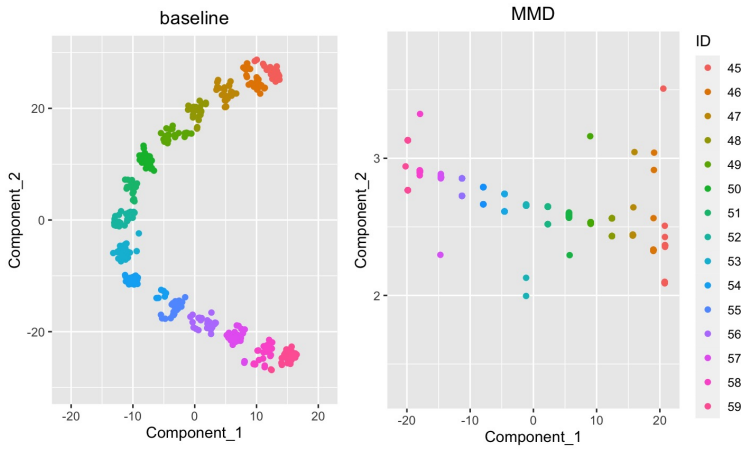

Figure 2: Latent Space Visualization Under MMD

5 Conclusions and Future Work

In this work, we proposed PiRL - which utilizes MMD and triplet loss for learning participant-invariant representations - to improve the performance of generic health detection models. We evaluated the performance and effectiveness of our framework using two public datasets for mental health and physical health. As preliminary results, we demonstrated that our proposed PiRL outperformed the baseline models and helped generic models achieve better performances. The limitation of MMD loss is that it can only shorten the distribution distance in certain dimensions. In future work, we will investigate other approaches to further optimize the representations in data-driven health detection applications.

References

[1] Aditya Kumar Akash, Vishnu Suresh Lokhande, Sathya N Ravi, and Vikas Singh. Learning invariant representations using inverse contrastive loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6582–6591, 2021.

[2] Majdi Bsoul, Hlaing Minn, and Lakshman Tamil. Apnea medassist: real-time sleep apnea monitor using single-lead ecg. *IEEE transactions on information technology in biomedicine*, 15(3):416–427, 2010.

[3] Urtnasan Erdenebayar, Yoon Ji Kim, Jong-Uk Park, Eun Yeon Joo, and Kyoung-Joung Lee. Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram. *Computer methods and programs in biomedicine*, 180:105001, 2019.

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

[5] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. Continuous stress detection using a wrist device: in laboratory and real life. In *proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*, pages 1185–1193, 2016.

[6] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[7] Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005.

[8] Lydia Kogler, Veronika I Müller, Amy Chang, Simon B Eickhoff, Peter T Fox, Ruben C Gur, and Birgit Derntl. Psychosocial versus physiological stress—meta-analyses on deactivations and activations of the neural correlates of stress reactions. *Neuroimage*, 119:235–251, 2015.

[9] Saskia Koldijk, Mark A Neerincx, and Wessel Kraaij. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Transactions on affective computing*, 9(2):227–239, 2016.

[10] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. Clas: A database for cognitive load, affect and stress recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*, pages 1–4. IEEE, 2019.

[11] Yoshiki Nakashima, Jonghwa Kim, Simon Flutura, Andreas Seiderer, and Elisabeth André. Stress recognition in daily work. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pages 23–33. Springer, 2015.

[12] Debarshi Nath, Mrigank Singh, Divyashikha Sethia, Diksha Kalra, and S Indu. A comparative study of subject-dependent and subject-independent strategies for eeg-based emotion recognition using lstm network. In *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis*, pages 142–147, 2020.

[13] Kizito Nkurikiyeyezu, Anna Yokokubo, and Guillaume Lopez. The effect of person-specific biometrics in improving generic stress predictive models. *arXiv preprint arXiv:1910.01770*, 2019.

[14] Shu Lih Oh, Eddie YK Ng, Ru San Tan, and U Rajendra Acharya. Automated diagnosis of arrhythmia using combination of cnn and lstm techniques with variable length heart beats. *Computers in biology and medicine*, 102:278–287, 2018.

[15] Thomas Penzel, George B Moody, Roger G Mark, Ary L Goldberger, and J Hermann Peter. The apnea-ecg database. In *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*, pages 255–258. IEEE, 2000.

[16] K Radhika and V Ramana Murthy Oruganti. Transfer learning for subject-independent stress detection using physiological signals. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–6. IEEE, 2020.

[17] K Radhika and V Ramana Murthy Oruganti. Deep multimodal fusion for subject-independent stress detection. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 105–109. IEEE, 2021.

[18] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[19] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[20] Gaetano Valenza, Luca Citi, Antonio Lanatá, Enzo Pasquale Scilingo, and Riccardo Barbieri. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Scientific reports*, 4(1):1–13, 2014.

[21] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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

[23] Alexandros Zenonos, Aftab Khan, Georgios Kalogridis, Stefanos Vatsikas, Tim Lewis, and Mahesh Sooriyabandara. Healthyoffice: Mood recognition at work using smartphones and wearable sensors. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 1–6. IEEE, 2016.

A Model Description

A.1 Unsupervised Learning Phase

For the CLAS dataset, the encoder consists of 9 CNN layers, whereas the decoder consists of 8 up-sampling layers. For the Apnea-ECG dataset, the encoder consists of 12 CNN layers, whereas the decoder consists of 11 up-sampling layers.

A.2 Supervised Learning Phase

On both datasets, the encoder structure is same with the structure from unsupervised learning shape. In addition, the dense layer on two datasets both consist of 2 fully connected layers.

B Experimental Setting

B.1 Dataset

For this study, we investigated two datasets, described as follows:

B.1.1 CLAS Dataset

The CLAS dataset consists of recordings of physiological signals such as ECG, EDA, and Plethysmography (PPG) [10]. EDA time series data were chosen from 62 participants which 58 participants are included in this study. Labels for arousal, valence, and stress were included in the data and assigned based on the stimuli tags for interactive tasks. We were interested in stress detection using EDA data.

Table 3 summarizes the EDA data used for the experiments. Training set contains 746 time series sample data with label 0 ('non-stressed') and 247 time series sample data with label 1 ('stressed') from overall 45 participants. Testing set contains 269 sample data in non-stressed label and 90 sample data in stressed label form overall 13 participants.

Table 4: EDA data separation.

	#Participants	#Non-stressed samples	#Stressed samples
Training Set	45	746	247
Testing Set	13	269	90
Total	58	1015	337

B.1.2 Apnea-ECG Dataset

The recordings of Apnea-ECG dataset includes continuous ECG signals and sets of annotations for apnea (respiratory signals). The length of the recordings ranges from slightly less than 7 hours to about 10 hours each. The Apnea-ECG dataset consisted of 70 participants was divided into a training set of 35 records (a01 through a20, b01 through b05, and c01 through c10), and a testing set of 35 records [15].

Table 4 summarizes the ECG data separation of the study. The labels of the sample are binary, which 0 represents normal breathing and 1 represents disordered breathing.

Table 5: Apnea-ECG data separation.

	#Participants	#Normal breathing samples	#Disordered breathing samples
Training Set	35	10496	6514
Testing Set	35	10685	6548
Total	70	21181	13062

B.1.3 Data Normalization

We used min-max normalization on the time series data for the CLAS and Apnea-ECG datasets to ensure that the normalized data had a similar scale. The formula of min-max normalization is given below:

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)}$$

B.2 Training Process

Our tasks are divided into two ways of evaluating representation learning. As for the representation learning, the objection function is mean squared error (MSE) loss between the reconstruction time series data and the original data. We need to obtain robust representations from the autoencoder that is used in the following step. As for the supervised health condition detection, CNN-based encoder was pre-trained with the auto-encoder. Then the supervised learning structure is fine-tuned according to different downstream task

In the convolutional autoencoder, we implemented 100 epochs per trail of training for both the baseline model and regularized model with MMD loss. We shuffled the data, set the batch size and learning rate to 32 and 0.001 at the start of each epoch in CLAS dataset, respectively. The batch size is 256 in Apnea-ECG dataset, which is the only difference between two datasets in training process. We specifically adjusted the weighted coefficient of the MMD loss to be 0.2 in the regularized model to prevent the dominance of the MMD loss. Additionally, the representation was set to be an 8×1 vector in order to guarantee that the representations we extracted are robust and contain sufficient personalized information.

In the supervised evaluation model, the parameters such as batch size and learning rate were maintained the same. These four supervised model types—baseline model, MMD loss only, triplet loss only, MMD and triplet loss were being taken into consideration. Noteworthy, the coefficient of the triplet loss was set to be 0.2 as well. Then we pre-trained the autoencoder, initialized and saved all the representations and parameters in the encoder part. After that, we loaded the saved parameters to the supervised models and fine-tuned them. To get the predicted binary results of labels, we feeded the representations obtained from the encoder to the classifier, which consists of two fully connected layers. The final prediction accuracy was recorded as the mean of 10 trials of training.