EEG-Based Driver Drowsiness Estimation Using Feature Weighted Episodic Training

Yuqi Cui, Yifan Xu, and Dongrui Wu, *Senior Member, IEEE*

***Abstract*— Drowsy driving is pervasive, and also a major cause of traffic accidents. Estimating a driver's drowsiness level by monitoring the electroencephalogram (EEG) signal and taking preventative actions accordingly may improve driving safety. However, individual differences among different drivers make this task very challenging. A calibration session is usually required to collect some subject-specific data and tune the model parameters before applying it to a new subject, which is very inconvenient and not user-friendly. Many approaches have been proposed to reduce the calibration effort, but few can completely eliminate it. This paper proposes a novel approach, feature weighted episodic training (FWET), to completely eliminate the calibration requirement. It integrates two techniques: feature weighting to learn the importance of different features, and episodic training for domain generalization. Experiments on EEG-based driver drowsiness estimation demonstrated that both feature weighting and episodic training are effective, and their integration can further improve the generalization performance. FWET does not need any labelled or unlabelled calibration data from the new subject, and hence could be very useful in plug-and-play brain-computer interfaces.**

***Index Terms*— Drowsy driving, domain generalization, EEG, episodic training, feature weighting.**

I. Introduction

Driving safety is very important to our everyday life. However, according to the World Health Organization[1] "Global Status Report on Road Safety 2018", "*the number of road traffic deaths continues to rise steadily, reaching 1.35 million in 2016. . . . Road traffic injuries are the eighth leading cause of death for all age groups. More people now die as a result of road traffic injuries than from HIV/AIDS, tuberculosis or diarrhoeal diseases. Road traffic injuries are currently the leading cause of death for children and young adults aged 5–29 years.*"

Manuscript received June 10, 2019; revised August 30, 2019 and September 25, 2019; accepted October 2, 2019. Date of publication October 7, 2019; date of current version November 6, 2019. This work was supported by the National Natural Science Foundation of China under Grant 61873321. *(Yuqi Cui and Yifan Xu contributed equally to this work.) (Corresponding author: Dongrui Wu.)*

The authors are with the Key Laboratory of the Ministry of Education for Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: yqcui@hust.edu.cn; yfxu@hust.edu.cn; drwu@hust.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2019.2945794

[1]https://www.who.int/violence_injury_prevention/road_safety_status/2018/English-Summary-GSRRS2018.pdf

In addition to the reliability of the vehicle and the driver's experience, driving safety is also strongly related to the driver's alertness (or, drowsiness). Drowsy driving is the fourth major contributor to road crashes, following only to alcohol, speeding, and inattention [1]. Drowsiness impacts the driver's ability to quickly and appropriately respond to road emergencies, and hence may lead to accidents [2]. Therefore, accurate estimation of the driver's drowsiness level is very important in preventing road accidents.

Many approaches have been reported [3]–[7], which can be roughly categorized into two directions: contactless detections and wearable sensor based detections. The former use cameras and/or other sensors, which are not attached to the driver's body, to monitor the driver's facial activities and/or driving patterns to estimate the drowsiness level [6], [8], [9]. The latter use wearable sensors to measure the driver's physiological signals, e.g., electroencephalogram (EEG) [10], electrocardiography (ECG) [10], [11], electromyography (EMG) [12], [13], etc, and then perform drowsiness estimation. The heart rate and heart rate variability can be easily obtained from ECG signals. They both vary significantly between alertness and drowsiness, and hence can be indicators of drowsiness [14], [15]. EMG signal is usually combined with other signals to determine the drowsiness level. For instance, Lee *et al.* [16] proposed a driver fatigue detection approach using EMG and galvanic skin responses. Fu *et al.* [17] proposed to use EEG, EMG and respiration signals to dynamically detect driver fatigue. In this paper, we focus on using only EEG signals for driver drowsiness estimation.

Since EEG directly measures the brain states, it is very suitable for human psychophysiological state evaluation [18]. The power spectrum of EEG has been used to estimate driver drowsiness level [19]–[22], especially the theta (4-7Hz) and alpha (8-12Hz) bands [18], [23], [24]. Additionally, different brain regions have different abilities in assessing the driver's drowsiness level. Previous studies have shown that theta and alpha band activities in the central and occipital regions are more correlated to fatigue [25]–[27]. These results indicate that it may be beneficial to give different brain regions different weights in drowsiness estimation.

A major challenge in EEG-based driver drowsiness estimation is that, due to individual differences, it is very difficult to develop a generic estimator, whose parameters are fixed and optimal for all subjects. Hence, a subject-specific calibration session is usually required to tune the estimator, which is

1534-4320 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.

time-consuming and not user-friendly. Lots of efforts have been made to reduce or eliminate this calibration. One of the most frequently used approach is transfer learning [28], [29], which uses data from other subjects/sessions (called source domains) to facilitate the learning for a new subject (called target domain). For instance, Lin and Jung [30] proposed a conditional transfer learning framework to promote positive transfer for each individual. It first assesses an individual's transferability for positive transfer, and then selectively leverages the data from others with comparable feature spaces. This approach has demonstrated promising performance in EEG-based emotion classification. Zanini *et al.* [31] proposed a Riemannian space transfer learning framework, which uses a reference covariance matrix at the resting state to align data from different domains, before applying a Riemannian space classifier. He and Wu [32] proposed a similar EEG data alignment approach in the Euclidean space, which is more efficient than the Riemannian space data alignment approach, and can be used as a pre-processing step before any Euclidean space classifier. However, all these approaches considered only classification problems, and all required some labeled or unlabeled data from the target subject for calibration. So, they cannot be used in true plug-and-play brain-computer interfaces.

This paper considers a much more realistic, also more challenging, scenario: there are no calibration data (either labeled or unlabeled) from the target subject at all; we want to build a model from the auxiliary subjects and apply it directly to the target subject. Each auxiliary subject can be viewed as an independent source domain, and this problem setting is called *domain generalization* in computer vision.

Many neural network based approaches have been proposed in recent years for domain generalization [33]–[38], which can be summarized into two categories:

1) Train a robust cross-domain model using a specially designed neural network architecture to reduce the domain shift. For instance, Ghifary *et al.* [33] proposed multi-task auto-encoder, which learns to transform the image in one domain into analogs in multiple related domains. These features, which are robust to variations across domains, are then fed into a classifier. Li *et al.* [35] proposed a low-rank parameterized convolutional neural network to compensate the domain shift. Li *et al.* [34] used adversarial auto-encoders to align the distributions among different domains by minimizing the maximum mean discrepancy (MMD), and matched the aligned distribution to an arbitrary prior distribution via adversarial feature learning. The first step ensures the learned feature representation is universal to the known source domains, and the second step ensures the features can generalize well to the unseen target domain.
2) Train models with regularization or meta-learning scheme regardless of the model structure. Balaji *et al.* [38] proposed a meta-regularization approach for domain generalization, which encodes domain generalization using a novel regularization function that makes the model trained in one domain to perform well in another domain. The regularization function was found in a learning-to-learn (or meta-learning) framework. Li *et al.* [37] proposed a model agnostic training procedure for domain generalization. Their algorithm simulated the shift between source and target domains during training by synthesizing virtual target domains within each mini-batch. The meta-optimization objective ensures performance improvements in both domains. Li *et al.* [36] further proposed an episodic training (ET) procedure that trains a single deep network while exposing it to the domain shift that characterises a novel domain at runtime. Specifically, it decomposes a deep network into two components: feature extractor and classifier, and then trains each component by simulating it interacting with a partner which may not be well tuned for the current domain.

This paper extends ET from classification to regression, and applies it to EEG-based driver drowsiness estimation. Our main contributions are:

1) We propose a feature weighting (FW) scheme that automatically assigns each feature a weight, by taking different importance of different brain regions into consideration.
2) We extend ET in [36] from classification to regression, and simplify it so that the computational cost is reduced without sacrificing the generalization performance.
3) We integrate FW and ET into a single learning framework, feature weighted episodic training (FWET), to achieve better generalization performance than each individual module.

The remainder of this paper is organized as follows: Section II introduces our dataset, feature extraction method, and the proposed FWET approach. Section III evaluates the performance of FWET in EEG-based driver drowsiness estimation. Section IV draws conclusion.

II. Feature Weighted Episodic Training (FWET)

This section introduces the dataset for EEG-based driver drowsiness estimation, and our proposed FWET approach, whose overall flowchart is shown in Fig. 1, along with several other variants.

A. Dataset

The data were collected in a simulated driving experiment, which was identical to that used in [19], [21], [39], [40]. Sixteen healthy subjects (age 24.2 $\pm$ 3.7, ten males, six females) with normal or corrected to normal vision were recruited to participate in a sustained-attention driving experiment [39], [40], which consisted of a real vehicle mounted on a motion platform with six degrees of freedom immersed in a 360-degree virtual reality scene. The experiment simulated driving on an empty highway at 100km/h. Every 5-10 seconds, a random lane-departure event was activated, which caused the car to drift from the center of the lane. The participants were asked to steer the car back to the center of the

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.

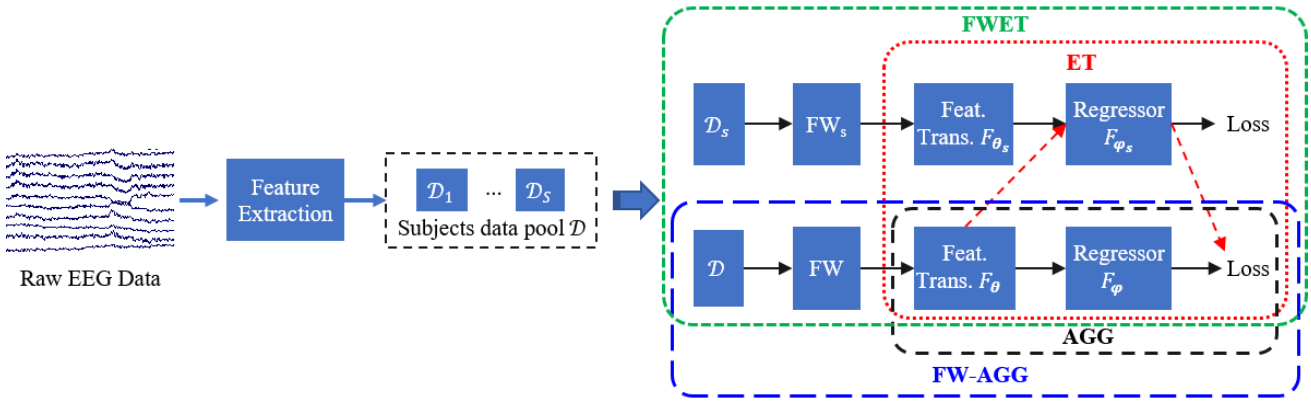

Fig. 1. AGG, FW-AGG, ET and FWET for EEG-based driver drowsiness estimation.

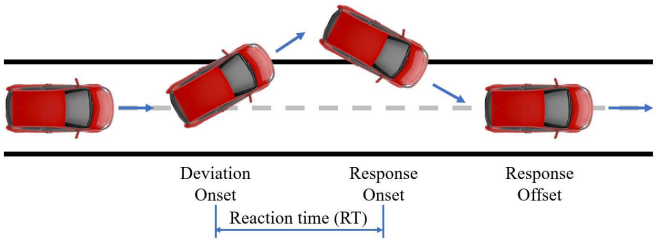

Fig. 2. Illustration of the way the reaction time was computed.

lane immediately. The reaction time was calculated as the time difference between the drift and the moment the subject started to act, as shown in Fig. 2. If the participant did not respond to the lane-departure event, such as falling asleep, the vehicle would hit the boundary of the road and continue moving forward along the boundary. The next lane-departure event happened after the response offset. Each participant performed the experiment for 60-90 minutes in the afternoon when the circadian rhythm of sleepiness reached its peak [41].

The Institutional Review Board of the Taipei Veterans General Hospital approved the experimental protocol. Each participant read and signed an informed consent form before the experiment began.

The reaction time τ was later converted into a drowsiness index (DI) [21], [22], [42]–[44],

$$DI = \max\left(0, \frac{1 - e^{-(\tau - \tau_0)}}{1 + e^{-(\tau - \tau_0)}}\right), \tag{1}$$

where τ_0 was set to 1 in our work. The DIs were then smoothed by a 90s moving-average window. (1) maps the reaction time to [0, 1] and overcomes its long-tail effect (very large reaction time was rare, but it did exist; such extreme values would significantly deteriorate the overall estimation). The fatigue level has been demonstrated to have a strong correlation with the reaction time [45]. Since the DI is positively correlated with the reaction time, DI is also an indicator of the fatigue level.

Note that the value of τ_0 could also be set individually for each subject. For instance, in [42], τ_0 was set to the 5 percentile value of the reaction time in each session. However, in a real-world online plug-and-play brain-computer-interface system, we do not have training data from the target subject, thus setting τ_0 individually is not possible. Nevertheless, to demonstrate the robustness of our proposed approach, we also compare the performances using $\tau_0 = 1$ and individualized τ_0 in Section III-F, which is possible in offline driver drowsiness estimation.

During the experiment, EEG signals were recorded using a 500Hz 32-channel Neuroscan system (30-channel EEGs plus 2-channel earlobes). Since data from one subject were not recorded correctly, we only used 15 subjects in our paper. To ensure a fair comparison, we used the first 3,600 seconds data from each subject.

B. Preprocessing and Feature Extraction

We used EEGLAB [46] for data preprocessing. We first performed 1-50Hz band-pass filtering to remove artifacts and noise, and then down-sampled the data from 500Hz to 250Hz and re-referenced them to the averaged earlobes.

We tried to predict the DI for each subject every 3 seconds, using 30-second EEG signal before each sample point. We computed the average power spectral density (PSD; their absolute values, instead of relative values, were used) in theta and alpha bands using Welch's method [47], with Hamming window, 1024 points fast Fourier transform, and 50% overlapping. The PSDs were then converted into dBs and used as our features. Each feature vector had $30 \times 2 = 60$ dimensions. All algorithms in our experiments used the same PSD features described above.

Each 30-second EEG signal may include brain activities, e.g., visual stimulus of the lane departure event and the wheel steering intention, and interferences from the wheel steering motor execution and other body movements. These brain activities and interferences are inevitably happening in real-world driving scenarios, and a good drowsiness estimation algorithm should be able to cope with them. Moreover, there are some other activities that are normal in realistic driving situations but were not considered in our experiments, e.g., the motor executions of acceleration and braking, talking, etc. These should be considered in the future improved experiment design.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.

C. Problem Setting

Assume Subject s has N_s labeled EEG trials $\mathcal{D}_s = \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{N_s}$, where $\mathbf{x}_i^s \in \mathbb{R}^{d\times 1}$ is a d-dimension feature vector extracted from the i-th EEG trial of Subject s, and y_i^s is the corresponding DI. Assume also that we have S subjects in our training set, and we want to predict the DI for $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$ from an unseen target subject t.

Our model contains two components: $F_{\boldsymbol{\theta}}$, the feature transformation network, and $F_{\boldsymbol{\psi}}$, the regression network. Hence the prediction for $\mathbf{x}_i^t$ is $\hat{y}_i^t = F_{\boldsymbol{\psi}}(F_{\boldsymbol{\theta}}(\mathbf{x}_i^t))$.

D. Aggregation Training (AGG)

The simplest domain generalization approach is to combine all source subjects' data to train one single model, which is usually a very strong baseline. This method is called aggregation training (AGG) in [36].

In this paper, we perform AGG using a multi-layer perceptron (MLP) neural network with one hidden layer and ReLU activation function. The loss function is:

$$\mathcal{L}_{AGG} = \sum_{s=1}^{S}\sum_{i=1}^{N_s} \ell(y_i^s,\ F_{\boldsymbol{\psi}}(F_{\boldsymbol{\theta}}(\mathbf{x}_i^s))), \tag{2}$$

where ℓ is the squared error in regression. The parameters $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ are learned through gradient descent optimization.

E. Feature Weighting (FW)

Previous studies [25]–[27] have shown that EEG features (channel-wise PSD features in this paper) in different brain regions have different correlations to the drowsiness. Thus, we use the following FW scheme to assign different weights to different EEG channels:

$$\hat{w}_l = e^{w_l}/\sum_{j=1}^{d} e^{w_j},\quad l = 1,\ldots,d \tag{3}$$

$$\hat{\mathbf{x}}_i^s = \hat{\mathbf{w}} \circ \mathbf{x}_i^s,\quad s = 1,\ldots,S \tag{4}$$

where $\mathbf{w} = [w_1,\ldots,w_d]^T \in \mathbb{R}^{d\times 1}$ and $\hat{\mathbf{w}} = [\hat{w}_1,\ldots,\hat{w}_d]^T \in \mathbb{R}^{d\times 1}$ are the original and transformed weight vectors, respectively, and $\circ$ denotes element-wise product. We do not use the weight $\mathbf{w}$ directly in (4); instead, we use its $softmax$ version $\hat{\mathbf{w}}$, to make sure the weights are non-negative and sum up to 1.

F. Episodic Training (ET)

ET for domain generalization was recently proposed by Li *et al.* [36] for image recognition. We simplify their algorithm and integrate it with FW. The original ET algorithm in [36] contains three regularization terms. In our work, we only adopt the first loss term (described as *epif* in Section III-D) for simplicity and speed. As it will be shown in Section III-G, our simplification greatly reduces the computational cost of the original ET, without sacrificing its generalization performance.

A common approach in transfer learning to learn domain-invariant features is to train a feature extractor $F_{\boldsymbol{\theta}}$ that makes the marginal distribution $P(F_{\boldsymbol{\theta}}(\mathbf{x}^s))$ consistent for different source domains, $s = 1,\ldots,S$. However, since the DIs of different subjects vary due to individualized differences, i.e., the conditional distributions $P(y^s|F_{\boldsymbol{\theta}}(\mathbf{x}^s))$ are different for different subjects, aligning the marginal distributions only may not lead to satisfactory generalization performance. ET considers the conditional distributions $P(y^s|F_{\boldsymbol{\theta}}(\mathbf{x}^s))$ directly, and trains an $F_{\boldsymbol{\theta}}$ that aligns $P(y^s|F_{\boldsymbol{\theta}}(\mathbf{x}^s))$ in all source domains, which usually generalizes well to the unseen target domain $\mathcal{D}_t$.

We first establish a subject-specific feature transformation (FT) model $F_{\boldsymbol{\theta}_s}$ and a subject-specific regression model $F_{\boldsymbol{\psi}_s}$ for each source subject to learn the domain-specific information. We also want to train an FT model $F_{\boldsymbol{\theta}}$ that makes the transformed features from Subject s still perform well when applied to a regressor $F_{\boldsymbol{\psi}_j}$ trained on Subject j ($j \neq s$). Hence, the following loss function is used:

$$\mathcal{L}_{FT}^{s,j} = \sum_{i=1}^{N_s} \ell(y_i^s,\ \overline{F}_{\boldsymbol{\psi}_j}(F_{\boldsymbol{\theta}}(\mathbf{x}_i^s))), \tag{5}$$

where $\overline{F}_{\boldsymbol{\psi}_j}$ means that $F_{\boldsymbol{\psi}_j}$ is not updated during back propagation.

The overall loss function of ET, when Subject s's data are fed into Subject j's regressor, is:

$$\mathcal{L}_{ET}^{s,j} = \mathcal{L}_{AGG}^{s} + \lambda \cdot \mathcal{L}_{FT}^{s,j}. \tag{6}$$

where

$$\mathcal{L}_{AGG}^{s} = \sum_{i=1}^{N_s} \ell(y_i^s,\ F_{\boldsymbol{\psi}}(F_{\boldsymbol{\theta}}(\mathbf{x}_i^s))). \tag{7}$$

$\lambda = 0.1$ was used in our experiments.

Note that since there is a (purposeful) mismatch between $F_{\boldsymbol{\psi}}$ and $F_{\boldsymbol{\theta}}$ in $\mathcal{L}_{FT}^{s,j}$, the gradient $\partial\mathcal{L}_{FT}^{s,j}/\partial F_{\boldsymbol{\theta}}$ may be unstable and sometimes have gradient explosion. Therefore, we clipped the gradient $\partial\mathcal{L}_{FT}^{s,j}/\partial F_{\boldsymbol{\theta}}$ to $[-10, 10]$.

G. FWET

Our proposed algorithm, FWET, which integrates FW and ET, is shown in Algorithm 1. It learns $\mathbf{w}$ in FW and $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ in ET simultaneously through gradient descent optimization.

All $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, $\boldsymbol{\theta}_s$ and $\boldsymbol{\psi}_s$, $s = 1,\ldots,S$, are uniformly initialized. Take $\boldsymbol{\theta}$ as an example. Let M be the number of features in each layer. Then, each element of $\boldsymbol{\theta}$ is initialized as a uniformly distributed random variable in $[-\sqrt{1/M}, \sqrt{1/M}]$.

III. Experimental Results

This section studies the performance of FWET in EEG-based driver drowsiness estimation.

A. Evaluation Method and Performance Measures

We used leave-one-subject-out cross-validation to validate the performance of our model. Since this was a regression problem, we used two metrics to evaluate the prediction results: root mean squared error (RMSE) and Pearson correlation coefficient (CC), which respectively measure the error and the correlation between the predicted DIs and the groundtruth DIs.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Pseudocode of FWET

Input: Training subject data $\{\mathbf{x}_i^s, y_i^s\}_{i=1}^{N_s}$, $s = 1, \ldots, S$;
ET weight λ;
Batch size N;
Learning rate α.
Output: FWET model parameters $\mathbf{w}$, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$.
for $s = 1 : S$ **do**
  Initialize domain-specific FW vector $\mathbf{w}_s = \mathbf{1}$;
  Randomly initialize domain-specific model parameters $\boldsymbol{\theta}_s$ and $\boldsymbol{\psi}_s$;
end
Initialize $\mathbf{w} = \mathbf{1}$ in FWET;
Randomly initialize $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ in FWET;
// Warm up
for $s = 1 : S$ **do**
  Train the domain-specific model parameters $\mathbf{w}_s$, $\boldsymbol{\theta}_s$ and $\boldsymbol{\psi}_s$ for one epoch, using only data from Subject s;
end
while *training* **do**
  for $s = 1 : S$ **do**
    Sample a batch $\{\mathbf{x}_i^s, y_i^s\}_{i=1}^N$ from Subject s;
    Compute $\hat{\mathbf{x}}_i^s$ using (4), $i = 1, \ldots, N$;
    Compute the sum of squared loss for the domain-specific model
    $\mathcal{L}_{DS} = \sum_{i=1}^N \ell(y_i, F_{\boldsymbol{\psi}_s}(F_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}_i^s)))$;
    $\mathbf{w}_s = \mathbf{w}_s - \alpha \nabla_{\mathbf{w}_s} \mathcal{L}_{DS}$;
    $\boldsymbol{\theta}_s = \boldsymbol{\theta}_s - \alpha \nabla_{\boldsymbol{\theta}_s} \mathcal{L}_{DS}$;
    $\boldsymbol{\psi}_s = \boldsymbol{\psi}_s - \alpha \nabla_{\boldsymbol{\psi}_s} \mathcal{L}_{DS}$;
  end
  for $s = 1 : S$ **do**
    for $j = 1 : S$ *and* $j \neq s$ **do**
      Sample a batch $\{\mathbf{x}_i^s, y_i^s\}_{i=1}^N$ from Subject s;
      Compute the loss $\mathcal{L}_{ET}^{s,j}$ in (6) on the batch;
      $\mathbf{w} = \mathbf{w} - \alpha \nabla_{\mathbf{w}} \mathcal{L}_{ET}^{s,j}$;
      $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}_{ET}^{s,j}$;
      $\boldsymbol{\psi} = \boldsymbol{\psi} - \alpha \nabla_{\boldsymbol{\psi}} \mathcal{L}_{ET}^{s,j}$;
    end
  end
end

We compared six different algorithms:

- *kNN*, which was a k-nearest neighbors regressor with $k = 5$. The prediction was the average of the five nearest neighbors.
- *RR*, which was ridge regression with L2 regularization coefficient $\alpha = 0.1$.
- *AGG*, which was an MLP neural network with only one hidden layer, trained using the loss function in (2). The number of hidden layer units was 40.
- *FW-AGG*, which performed FW before AGG.
- *ET*, which trained an AGG model and S domain-specific models together using ET. Each such model has the same structure as AGG above, i.e., a 3-layer MLP. The first layer was treated as $F_{\boldsymbol{\theta}}$. The other two layers were treated as $F_{\boldsymbol{\psi}}$.
- *FWET*, which performed FW before ET.

The first two algorithms are commonly used baselines for regression problems. The last four are AGG based. We compare them to analyze the individual contributions of FW and ET in FWET. The last four models were trained using mini-batch gradient descent with momentum, with batch size 32, learning rate 0.001, momentum 0.9, and weight decay 0.00005. We sampled 10% data from each training subject as the validation set in early-stopping to reduce overfitting. The maximum number of training epochs was set to 500, and early-stopping patience was 10 epochs. One epoch means one *training* iteration in Algorithm 1. We repeated all algorithms five times and report the average performance.

B. Experimental Results

The regression performance for each subject, averaged over five repeats, is shown in Fig. 3. FW-AGG, ET and FWET outperformed kNN, RR and AGG for most subjects. One exception is Subject 10, on which FW-AGG and FWET gave negative CCs.

To explore why FW-AGG and FWET gave weird CCs on Subject 10, we plot the feature distributions of Subject 10, along with those from the other 14 subjects, in Fig. 4. We first plot the 10 and 90 percentile of PSD features from each subject in Fig. 4(a) and a t-SNE visualization in Fig. 4(b) to see if there are differences on feature distribution between subjects. Clearly, the distributions of the 51st and 52nd features of Subject 10 are dramatically different from those of other subjects, which may be due to outliers. We can also see that there are some data points from Subject 10 that are not consistent with the data from other subjects. The unsatisfactory performance of FW-AGG and FWET on Subject 10 suggests that maybe FW is sensitive to outliers.

In offline applications, we know the features from the target subject. So, preprocessing may be used to remove the outlier features. For example, when the 51st and 52nd outlier features of Subject 10 were removed, the corresponding boxplots of the RMSEs and CCs of FW-AGG and FWET are shown in Fig. 5. They were considerably improved over the RMSEs and CCs in Fig. 3.

The last group in each subfigure of Fig. 3 also shows the average performance across the 15 subjects, whose values are given in Table I. AGG is a nonlinear model with more parameters than kNN and RR. Theoretically, it should outperform kNN and RR if well-trained. However, Table I shows that this was not the case. AGG had slightly worse average RMSE than kNN, and slightly worse average CC than RR. There may be two reasons: 1) there were not enough training data to tune AGG well; and, 2) AGG was over-fitted on the training data, so it did not generalize well to a new subject. After introducing FW and ET, both training performance and generalization ability were improved, and both FW-AGG and FWET outperformed the three baselines (kNN, RR, and AGG). More specifically, ET outperformed AGG, improving 4.9% and 0.2% on the RMSE and the CC, respectively. After adding FW to AGG, FW-AGG further outperformed ET by 4.5% on the RMSE and 4.3% on the CC. FWET achieved the best

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.

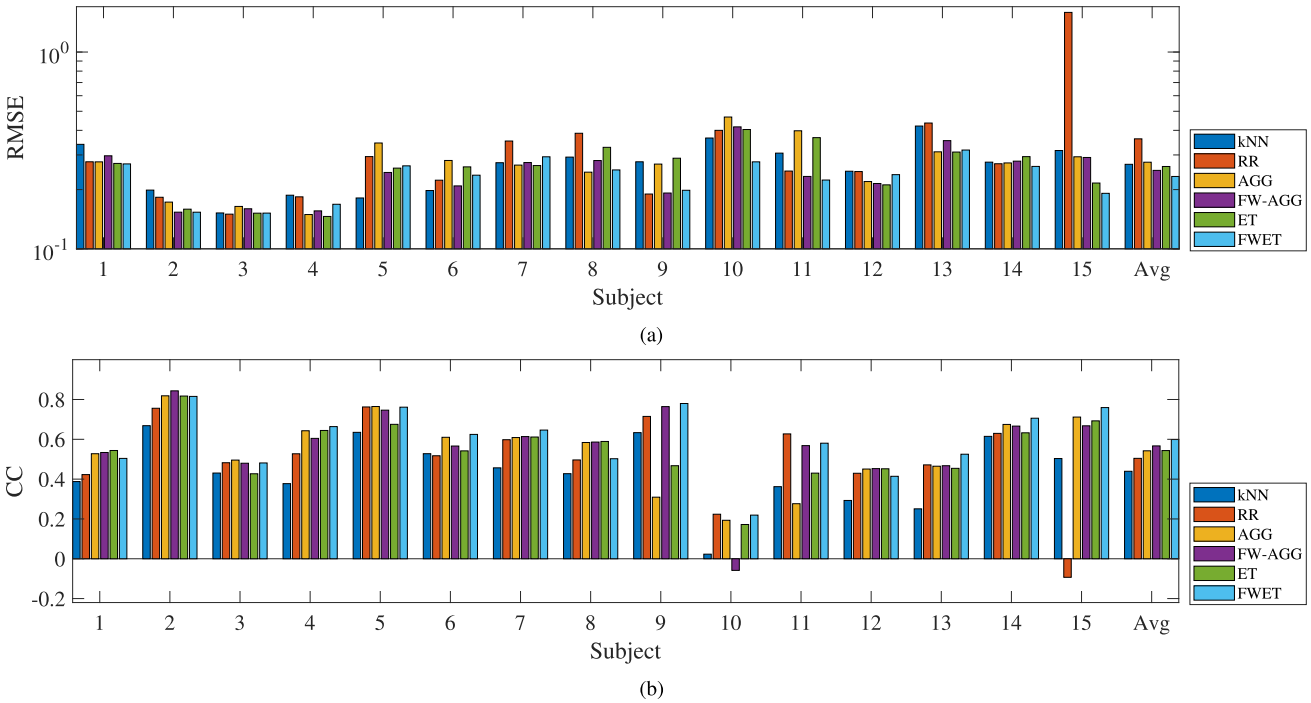

Fig. 3. (a) RMSEs and (b) CCs in leave-one-subject-out cross-validation. The experiments were repeated five times, and the averages are shown.

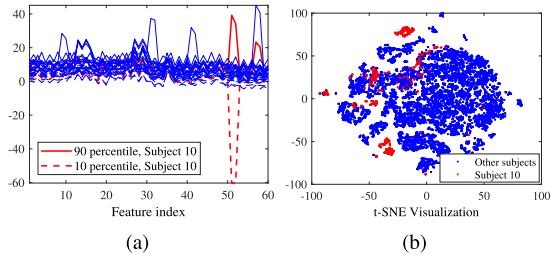

Fig. 4. (a) 90 and 10 percentiles of features from Subject 10, w.r.t. the corresponding feature percentiles (90: solid curves; 10: dashed curves) from the other 14 subjects; (b) *t*-SNE visualization of the features from different subjects.

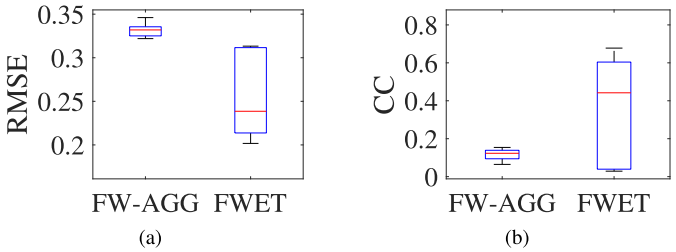

Fig. 5. Boxplots of (a) RMSEs and (b) CCs of FW-AGG and FWET, when Subject 10 was the target (test) subject, and the 51st and 52nd outlier features were removed.

performance, and further improved the RMSE and the CC by 6.9% and 5.7%, respectively, over FW-AGG.

To determine if the differences between different algorithms were statistically significant, we also performed non-parametric multiple comparison tests on the RMSEs and CCs using Dunn's procedure [48], with a p-value correction using the False Discovery Rate method [49]. The results are shown in Table II, where the statistically significant ones are marked in bold. FWET statistically significantly outperformed kNN, RR and AGG on the RMSE, and also kNN and RR on the CC. Though the performance improvements of FWET over FW-AGG and ET were not statistically significant, we have seen from Fig. 3 and Table I that on average FWET still slightly outperformed them.

In summary, we have shown that it is always preferable to use FWET over the other five algorithms.

TABLE I
AVERAGE RMSES AND CCS ACROSS THE 15 SUBJECTS AND FIVE RUNS

	kNN	RR	AGG	FW-AGG	ET	FWET
RMSE	0.2688	0.3622	0.2756	0.2504	0.2621	**0.2332**
CC	0.4394	0.5044	0.5422	0.5668	0.5434	**0.5989**

TABLE II
p-VALUES OF NON-PARAMETRIC MULTIPLE COMPARISONS ON THE RMSES AND CCS

		kNN	RR	AGG	FW-AGG	ET
RMSE	RR	.3697				
	AGG	.4558	.3814			
	FW-AGG	.1478	.0936	.1376		
	ET	.1982	.1390	.2146	.3783	
	FWET	**.0182**	**.0098**	**.0170**	.1745	.1335
CC	RR	**.0063**				
	AGG	**.0000**	.0875			
	FW-AGG	**.0000**	.0347	.3043		
	ET	**.0001**	.1657	.3440	.1908	
	FWET	**.0000**	**.0021**	.0873	.1832	.0403

C. Effects of FW

The four AGG based algorithms have randomness involved, e.g., initialization, batch selection, etc. It's interesting to study their stability. Recall that we had 15 subjects, and each

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.

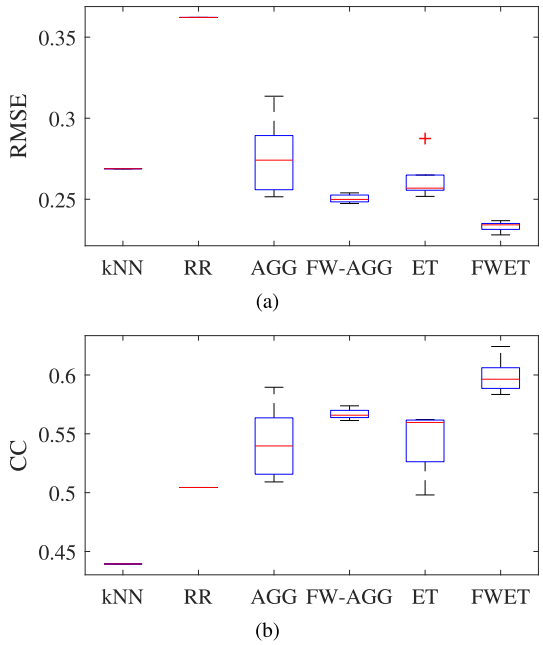

Fig. 6. Boxplots of the average (a) RMSEs and (b) CCs of the six algorithms.

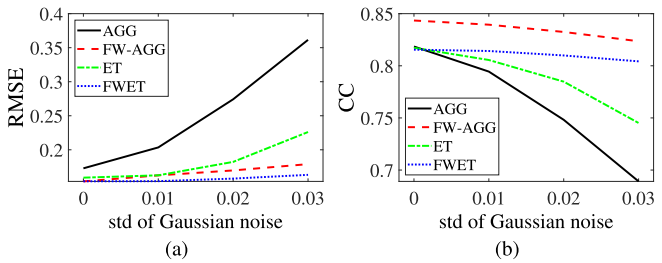

Fig. 7. (a) RMSE and (b) CC when Gaussian noise was added to the learned parameters of different algorithms. The models were trained on Subjects 1, 3-15 and tested on Subject 2.

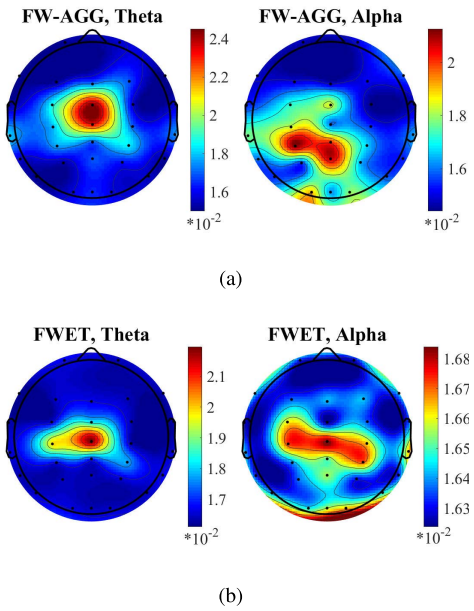

Fig. 8. EEG channel importance in theta and alpha bands, converted from *softmax*(w) in (a) FW-AGG and (b) FWET.

algorithm was run five times when each subject was used as the target subject. The final results were assembled into a 15×5 RMSE matrix and a 15×5 CC matrix. We could plot a boxplot for each subject to show the stability of different algorithms, but that would take too much space, and is difficult to see the forest for the trees. So, we first computed the average performance of each algorithm over the 15 subjects, i.e., we took the average of the RMSE (CC) matrix along the columns to obtain a 1×5 row vector, and then plotted the box-plots of the five average RMSEs (CCs) in Fig. 6. The RMSEs and CCs of kNN and RR did not have uncertainty, because there was no randomness in these algorithms. Among the four AGG based algorithms, AGG and ET had large variations, and FW-AGG and FWET had very small variations, suggesting one more advantage of introducing FW to AGG, beyond better RMSE and CC.

Fig. 6 shows that generally FW helped reduce the variation from different runs. It's interesting to study why. Several studies had analyzed the relationship between the generalization performance and sharp minima [50], [51]. It is believed that sharp minima may lead to bad generalization performance. ET tends to have flatter minima, which had already been demonstrated in [36]. We want to investigate if FW has a similar effect. We added random Gaussian noise to the learned parameters and checked how quickly the performance degraded. A rapid decrease indicates that the model is at a sharp minimum, which is bad for generalization.

As shown in Fig. 7, FW-AGG was more robust to the perturbations than AGG, and FWET was also more robust than ET. These observations demonstrated that FW led the model to flatter minima in the parameter space, which helped improve its generalization ability.

We also visualize the importance of different regions in each power band, determined by $\mathbf{w}$ in FW. Fig. 8(a) shows the topoplots of $\mathbf{w}$ in theta and alpha bands after the $softmax$ function in FW-AGG, when the last 14 subjects were used in training. For the theta band, the central brain region had the maximum weights, i.e., it contributed the most to drowsiness estimation. For the alpha band in FW-AGG, both the central and the occipital brain regions contributed more to drowsiness estimation than other regions. These were partially consistent with the findings in [26], where Zhao *et al.* studied mental fatigue in 90-minute continuous simulated driving, and found that the frontal, *central* and occipital regions in the theta band, and the *central*, parietal, *occipital* and temporal regions in the alpha band, all had significant difference at the beginning and the end of the driving.

Fig. 8(b) shows the topoplots of $\mathbf{w}$ in theta and alpha bands after the $softmax$ function in FWET, when the last 14 subjects were used in training. We can observe roughly the same patterns as in Fig. 8(a) for FW-AGG. However, note that the magnitude ranges in Fig. 8(b) were much smaller than those in Fig. 8(a), i.e., $\mathbf{w}$ in FWET had smaller variance than that in FW-AGG.

We also computed the average PSD values for alert and drowsy states over the 15 subjects. We considered the subject

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.

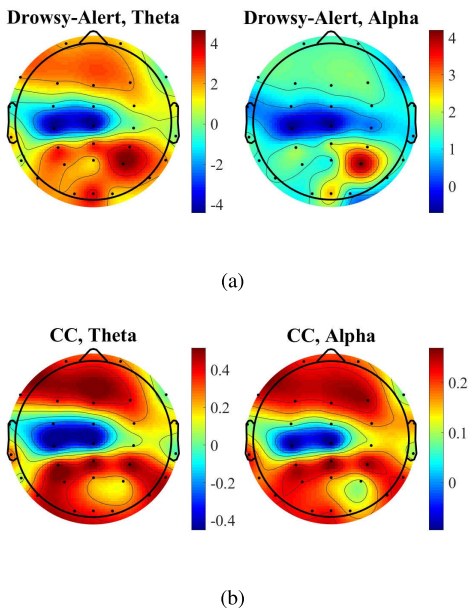

Fig. 9. (a) the differences between the topoplots of the drowsy and alert states. (b) the Pearson correlation coefficient between each PSD feature and the DI.

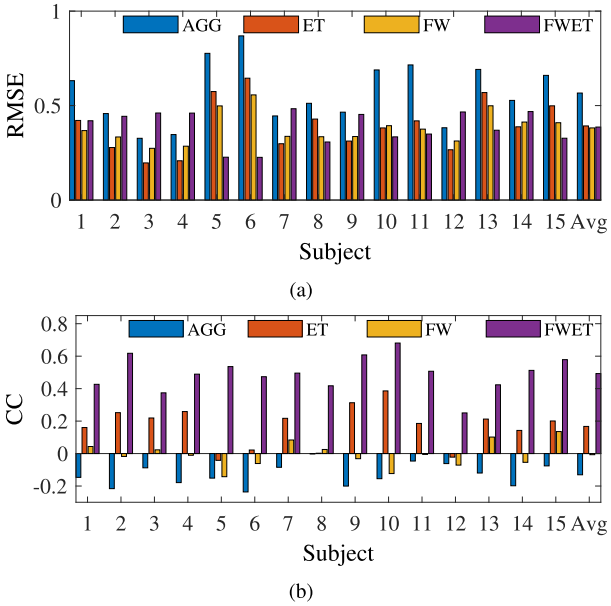

Fig. 10. Average RMSEs (a) and CCs (b) when Subject j's regression network F_{ψ_j} was applied to data from Subject s ($s \neq j$). The feature transformation was $F_{\theta}(\boldsymbol{x})$ for both AGG and ET. The feature transformation was $F_{\theta}(\hat{\mathrm{w}} \circ \boldsymbol{x})$ for both FW-AGG and FWET.

be alert (drowsy) when his/her DI was lower (higher) than the 15 (85) percentile of the DIs over the entire session. Fig. 9(a) shows the differences between the topoplots of the drowsy and alert states, and Fig. 9(b) the Pearson correlation coefficient between each PSD feature and the DI. Interestingly, Figs. 8(b) and 9(b) are not similar, i.e., though the channel weights $\mathbf{w}$ helped improve the drowsiness estimation performance, they were different from the correlation coefficients between the corresponding features and the DI.

Finally, although FW looks similar to the attention mechanism [52], which is being widely used in computer vision and natural language processing, they are different. The attention mechanism assigns dynamic weights to the neighboring locations, which change as the input varies. FW uses a fixed weight for each EEG channel, as the contributions of different brain regions usually do not change much in the same mental task.

D. Effects of ET

This subsection first presents two experiments to understand how ET helped extract more generalizable features from different subjects, and then studies the effect of adding more regularization terms in ET and FWET.

We used data from all 15 subjects to train AGG, ET, FW and FWET, which had different feature extractor F_{θ}. To compare these two F_{θ}, we input Subject s's data to each F_{θ}, and used Subject j's ($j \neq s$) regressor F_{ψ_j} (which was trained on data from Subject j only) for regression. For AGG and ET, the final regression model was $F_{\psi_j}(F_{\theta}(\mathbf{x}^s))$. For FW-AGG and FWET, the final regression model was $F_{\psi_j}(F_{\theta}(\hat{\mathbf{w}} \circ \mathbf{x}^s))$. We tried all $j \neq s$ for each s, i.e., 14 different F_{ψ_j}, and computed the average performance for each s. The smaller (larger) the RMSE (CC) is, the better the generalization performance is. The results are shown in Fig. 10, where the subject index means that subject's data were used as input (Subject s in the above description). ET always achieved a smaller RMSE and a larger CC, suggesting that ET extracted more generalizable features. FWET had comparable RMSEs as FW-AGG, but generally larger CCs than FW-AGG, suggesting again that ET extracted more generalizable features.

Three different regularizations were used in ET in [36] for classification problems:

1) *epif* (short for *episodic feature*), which requires the trained *feature extractor* to work well with all domain-specific *classifiers*.
2) *epic* (short for *episodic classifier*), which requires the trained *classifier* to work well with all domain-specific *feature extractors*.
3) *epir* (short for *episodic random*), which requires the *feature extractor* to work well with a randomly initialized *classifier* (representing a completely new domain).

We only adopted *epif* in our ET, because it was much easier and faster to optimize. The average training time per iteration, when different regularization terms were used in ET and FWET, are shown in Table III. Intuitively, the computational cost increased when more regularization terms were used.

Table III also shows the RMSEs and CCs when more regularizations were used. The weights for the three regularization terms were all set to 0.1. For both ET and FWET, using *epif* only achieved comparable performance with models using more regularization terms, and sometimes even slightly better. For the same type of regularization, FWET always outperformed ET, suggesting again the benefit of FW.

In summary, we have shown that our proposed ET and FWET are efficient and effective, and their extracted features

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.

TABLE III
AVERAGE RMSES, CCS AND TRAINING TIME (S) WHEN DIFFERENT REGULARIZATION TERMS WERE USED IN ET AND FWET

	RMSE	CC	Time (s)
ET (ET-epif)	0.2621	0.5434	**0.5302**
ET-epif-epic	**0.2577**	**0.5486**	0.7050
ET-epif-epic-epir	0.2682	0.5230	0.9235
FWET (FWET-epif)	**0.2332**	**0.5989**	**0.9651**
FWET-epif-epic	0.2398	0.5771	1.0530
FWET-epif-epic-epir	0.2384	0.5795	1.2812

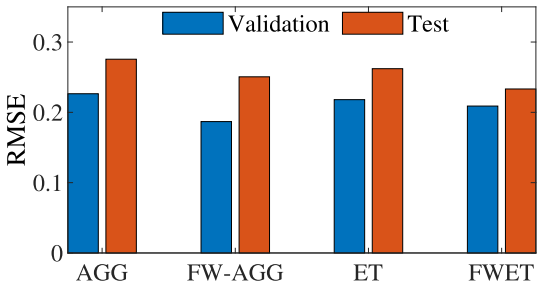

Fig. 11. Validation and test RMSEs of the four AGG-based algorithms. A smaller gap between the validation RMSE and the test RMSE indicates better generalizability.

have comparable (or even slightly better) generalization performance with those with more regularization terms.

E. Performance Gap between Validation and Test

Early stopping on a validation set is frequently used in machine learning to reduce overfitting, and was also the case in this paper. However, the validation performance is usually more optimistic than the test performance. A model with stronger generalization ability should have a smaller performance gap between the validation performance and the test performance.

Fig. 11 shows the validation and test RMSEs of the four AGG-based algorithms. Although AGG had the smallest validation RMSE, its test RMSE was the largest, i.e., the performance gap between the validation and test RMSEs were the largest, suggesting poor generalization ability. The validation-test RMSE gaps of FW-AGG, ET and FWET were considerably reduced. Particulary, FWET had the smallest gap, and the best test RMSE, suggesting its strong generalizability.

F. Individualized τ_0

$\tau_0 = 1$ in (1) was used in all above experiments. This is because we considered the most challenging case in brain-computer interfaces, i.e., we do not have any labeled data from the new subject. However, if we have some labeled data from the new subject, or some prior knowledge about the reaction time of the new subject, then it is possible to set τ_0 individually. This subsection demonstrates the performance of FWET in this case.

Following the practice in [42], we set τ_0 in (1) to be 5 percentile value of the reaction time of the corresponding subject, and repeated the experiments. The performances of FWET for constant and individualized τ_0 are shown in Fig. 12. Using individualized τ_0 reduced the RMSE for almost every subject (except Subject 4), although the CCs were roughly the same. This demonstrates that more information about the new subject can generally improve the drowsiness estimation performance.

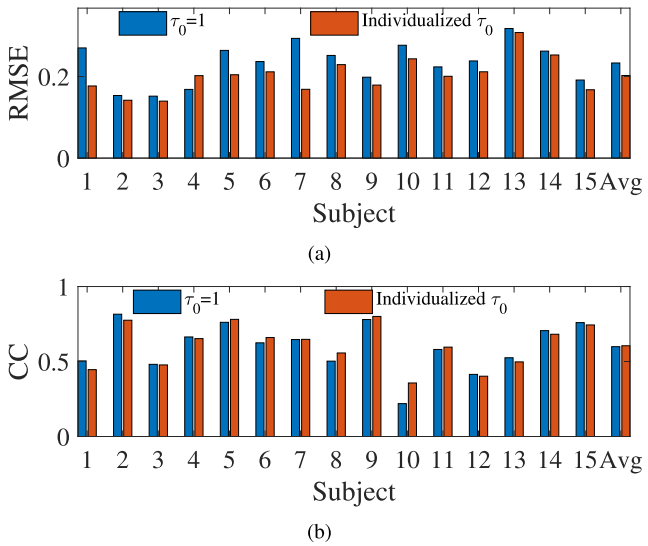

Fig. 12. (a) RMSEs and (b) CCs in leave-one-subject-out cross-validation of FWET, using $\tau_0 = 1$ and individualized τ_0.

G. Discussion

This paper extends domain generalization, a concept mostly used in computer vision, to brain-computer interfaces. There are some important differences between these two application areas, which should be paid attention to in future research:

1) *The number of source domains*. In computer vision applications, the number of domains is usually small, e.g., PACS[2] has four domains, IXMAS[3] has five domains, and MNIST[4] usually has seven rotated domains. Scalability is usually ignored in such applications. However, in brain-computer interfaces, more and more datasets with a large number of subjects are collected, and the scalability with respect to the number of domains can no longer be ignored.
2) *The variation of the label distribution in different domains*. Most existing domain generalization approaches only focus on learning a feature transformation T that makes all source domains to have roughly the same marginal distribution $P(T(X))$, without considering the label distribution $P(Y)$. In EEG-based driver drowsiness estimation, the distribution of DIs varies significantly among different subjects. This makes generalization across different subjects difficult in brain-computer interfaces.

[2]https://domaingeneralization.github.io/
[3]http://4drepository.inrialpes.fr/pages/home
[4]http://yann.lecun.com/exdb/mnist/

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.

IV. Conclusion

EEG-based driver drowsiness estimation could be very important in improving driving safety. Unfortunately, individual differences among different drivers make it very difficult to design a generic estimation algorithm, whose parameters are fixed and optimal for all subjects. Usually some subject-specific calibration data are needed to tune the model parameters before applying it to a new subject, which is very inconvenient and not user-friendly. Many approaches have been proposed to reduce this calibration effort, but few can completely eliminate it. This paper has proposed an FWET approach to completely eliminate the calibration requirement. It integrates two techniques: FW to learn the importance of different features, and ET for domain generalization. Experiments demonstrated that both FW and ET are effective, and their integration can further improve the generalization performance. FWET does not need any labelled or unlabelled calibration data from the new subject at all, and hence could be very useful in plug-and-play brain-computer interfaces. Our future research will apply FWET to more such applications.

References

[1] F. Sagberg, P. Jackson, H.-P. Krüger, A. Muzet, and A. J. Williams, *Fatigue, sleepiness and reduced alertness as risk factors in driving*. Oslo, Norway: Institute of Transport Economics, 2004.

[2] K. Kozak *et al.*, "Evaluation of lane departure warnings for drowsy drivers," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, Los Angeles, CA, USA, Oct. 2006, pp. 2400–2404.

[3] H. Abbood, W. Al-Nuaimy, A. Al-Ataby, S. A. Salem, and H. S. AlZubi, "Prediction of driver fatigue: Approaches and open challenges," in *Proc. 14th U.K. Workshop Comput. Intell. (UKCI)*, West Yorkshire, U.K., Sep. 2014, pp. 1–6.

[4] M. I. Chacon-Murguia and C. Prieto-Resendiz, "Detecting driver drowsiness: A survey of system designs and technology," *IEEE Consum. Electron. Mag.*, vol. 4, no. 4, pp. 107–119, Oct. 2015.

[5] H.-B. Kang, "Various approaches for driver and driving behavior monitoring: A review," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Sydney, NSW, Australia, Dec. 2013, pp. 616–623.

[6] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: A review," *Sensors*, vol. 12, no. 12, pp. 16937–16953, 2012.

[7] Y.-T. Wang *et al.*, "Developing an EEG-based on-line closed-loop lapse detection and mitigation system," *Frontiers Neurosci.*, vol. 8, p. 321, Oct. 2014.

[8] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 63–77, Mar. 2006.

[9] T. D'Orazio, M. Leo, C. Guaragnella, and A. Distante, "A visual approach for driver inattention detection," *Pattern Recognit.*, vol. 40, no. 8, pp. 2341–2355, 2007.

[10] E. Michail, A. Kokonozi, I. Chouvarda, and N. Maglaveras, "EEG and HRV markers of sleepiness and loss of control during car driving," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Vancouver, BC, Canada, Aug. 2008, pp. 2566–2569.

[11] G. Jahn, A. Oehme, J. F. Krems, and C. Gelau, "Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use in a driving study," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 8, no. 3, pp. 255–275, 2005.

[12] M. Akin, M. B. Kurt, N. Sezgin, and M. Bayram, "Estimating vigilance level by using EEG and EMG signals," *Neural Comput. Appl.*, vol. 17, no. 3, pp. 227–236, 2008.

[13] S. Hu and G. Zheng, "Driver drowsiness detection with eyelid related parameters by support vector machine," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7651–7658, 2009.

[14] K. Fujiwara *et al.*, "Heart rate variability-based driver drowsiness detection and its validation with EEG," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1769–1778, Jun. 2019.

[15] M. Miyaji, H. Kawanaka, and K. Oguri, "Driver's cognitive distraction detection using physiological features by the AdaBoost," in *Proc. 12th Int. IEEE Conf. Intell. Transp. Syst.*, Oct. 2009, pp. 1–6.

[16] B.-L. Lee, D.-S. Lee, and B.-G. Lee, "Mobile-based wearable-type of driver fatigue detection by GSR and EMG," in *Proc. IEEE Region Conf. (TENCON)*, Macao, China, Nov. 2015, pp. 1–4.

[17] R. Fu, H. Wang, and W. Zhao, "Dynamic driver fatigue detection using hidden Markov model in real driving condition," *Expert Syst. Appl.*, vol. 63, pp. 397–411, Nov. 2016.

[18] P. Aricò, G. Borghini, G. Di Flumeri, N. Sciaraffa, A. Colosimo, and F. Babiloni, "Passive BCI in operational environments: Insights, recent advances, and future trends," *IEEE Trans. Biomed. Eng*, vol. 64, no. 7, pp. 1431–1436, Jul. 2017.

[19] Y. Cui and D. Wu, "EEG-based driver drowsiness estimation using convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process.*, Guangzhou, China, Oct. 2017, pp. 822–832.

[20] C.-H. Chuang, Z. Cao, P.-T. Chen, C.-S. Huang, N. R. Pal, and C.-T. Lin, "Dynamically weighted ensemble-based prediction system for adaptively modeling driver reaction time," Sep. 2018, *arXiv:1809.06675*. [Online]. Available: https://arxiv.org/abs/1809.06675

[21] D. Wu, C.-H. Chuang, and C.-T. Lin, "Online driver's drowsiness estimation using domain adaptation with model fusion," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Xi'an, China, Sep. 2015, pp. 904–910.

[22] D. Wu, V. J. Lawhern, S. Gordon, B. J. Lance, and C.-T. Lin, "Driver drowsiness estimation from EEG signals using online weighted adaptation regularization for regression (OwARR)," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1522–1535, Dec. 2017.

[23] C.-H. Chuang, Z. Cao, J.-T. King, B.-S. Wu, Y.-K. Wang, and C.-T. Lin, "Brain electrodynamic and hemodynamic signatures against fatigue during driving," *Frontiers Neurosci.*, vol. 12, p. 181, Mar. 2018.

[24] W. Klimesch, P. Sauseng, and S. Hanslmayr, "EEG alpha oscillations: The inhibition–timing hypothesis," *Brain Res. Rev.*, vol. 53, no. 1, pp. 63–88, 2007.

[25] J. Perrier, S. Jongen, E. Vuurman, M. L. Bocca, J. G. Ramaekers, and A. Vermeeren, "Driving performance and EEG fluctuations during on-the-road driving following sleep deprivation," *Biol. Psychol.*, vol. 121, pp. 1–11, Dec. 2016.

[26] C. Zhao, M. Zhao, J. Liu, and C. Zheng, "Electroencephalogram and electrocardiograph assessment of mental fatigue in a driving simulator," *Accident Anal. Prevention*, vol. 45, pp. 83–90, Mar. 2012.

[27] C.-T. Lin, C.-H. Chuang, Y.-K. Wang, S.-F. Tsai, T.-C. Chiu, and L.-W. Ko, "Neurocognitive characteristics of the driver: A review on drowsiness, distraction, navigation, and motion sickness," *J. Neurosci. Neuroeng.*, vol. 1, no. 1, pp. 61–81, 2012.

[28] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[29] A. M. Azab, J. Toth, L. S. Mihaylova, and M. Arvaneh, *Signal Processing and Machine Learning for Brain-Machine Interfaces*. Edison, NJ, USA: Institution of Engineering and Technology, 2018, pp. 81–101, ch. 5.

[30] Y.-P. Lin and T.-P. Jung, "Improving EEG-based emotion classification using conditional transfer learning," *Frontiers Hum. Neurosci.*, vol. 11, p. 334, Jun. 2017.

[31] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 5, pp. 1107–1116, May 2018.

[32] H. He and D. Wu, "Transfer learning for brain-computer interfaces: A Euclidean space data alignment approach," *IEEE Trans. Biomed. Eng.*, to be published.

[33] M. Ghifary, W. B. Klein, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proc. ICCV*, Santiago, Chile, Jun. 2015, pp. 2551–2559.

[34] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5400–5409.

[35] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5543–5551.

[36] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," Jan. 2019, *arXiv:1902.00113*. [Online]. Available: https://arxiv.org/abs/1902.00113

[37] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. 32th AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Apr. 2018, pp. 3490–3497.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.

[38] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "MetaReg: Towards domain generalization using meta-regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2018, pp. 998–1008.

[39] C.-H. Chuang, L.-W. Ko, T.-P. Jung, and C.-T. Lin, "Kinesthesia in a sustained-attention driving task," *NeuroImage*, vol. 91, no. 1, pp. 187–202, May 2014.

[40] S.-W. Chuang, L.-W. Ko, Y.-P. Lin, R.-S. Huang, T.-P. Jung, and C.-T. Lin, "Co-modulatory spectral changes in independent brain processes are correlated with task performance," *NeuroImage*, vol. 62, no. 3, pp. 1469–1477, Sep. 2012.

[41] J. Horne and L. Reyner, "Vehicle accidents related to sleep: A review," *Occupational Environ. Med.*, vol. 56, no. 5, pp. 289–294, 1999.

[42] C.-S. Wei, Y.-P. Lin, Y.-T. Wang, T.-P. Jung, N. Bigdely-Shamlo, and C.-T. Lin, "Selective transfer learning for EEG-based drowsiness detection," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Hong Kong, Oct. 2015, pp. 3229–3232.

[43] C.-S. Wei, Y.-P. Lin, and T.-P. Jung, "Exploring the EEG correlates of neurocognitive lapse with robust principal component analysis," in *Proc. Int. Conf. Augmented Cognition*, Toronto, ON, Canada, Jul. 2016, pp. 113–120.

[44] C.-S. Wei, Y.-P. Lin, Y.-T. Wang, C.-T. Lin, and T.-P. Jung, "A subject-transfer framework for obviating inter- and intra-subject variability in EEG-based drowsiness detection," *NeuroImage*, vol. 174, pp. 407–419, Jul. 2018.

[45] Q. Ji, Z. Zhu, and P. Lan, "Real-time nonintrusive monitoring and prediction of driver fatigue," *IEEE Trans. Veh. Technol.*, vol. 53, no. 4, pp. 1052–1068, Jul. 2004.

[46] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[47] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, no. 2, pp. 70–73, Jun. 1967.

[48] O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, no. 3, pp. 214–252, 1964.

[49] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc., Ser. B*, vol. 57, pp. 289–300, Jan. 1995.

[50] P. Chaudhari *et al.*, "Entropy-SGD: Biasing gradient descent into wide valleys," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 1–19.

[51] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 1–16.

[52] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on July 06,2020 at 15:22:48 UTC from IEEE Xplore. Restrictions apply.