Online Mental Fatigue Monitoring via Indirect Brain Dynamics Evaluation

Article *in* Neural Computation · March 2021

DOI: 10.1162/neco_a_01382

5 authors, including:

Yuangang Pan
Agency for Science, Technology and Research (A*STAR)
25 PUBLICATIONS **51** CITATIONS
SEE PROFILE

Ivor W Tsang
University of Technology Sydney
288 PUBLICATIONS **14,373** CITATIONS
SEE PROFILE

Yueming Lyu
University of Technology Sydney
10 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Avinash Singh
University of Technology Sydney
53 PUBLICATIONS **212** CITATIONS
SEE PROFILE

Online Mental Fatigue Monitoring via Indirect Brain Dynamics Evaluation

Yuangang Pan[1] **Ivor W. Tsang**[1] **Yueming Lyu**[1] **Avinash K Singh**[1] **Chin-Teng Lin**[1]

[1]Australian Artificial Intelligence Institute, University of Technology Sydney.

Keywords: Online Mental Fatigue Monitoring, Brain Dynamics, Preference, Multi-channel Fusion, electroencephalogram (EEG)

Abstract

Driver mental fatigue leads to thousands of traffic accidents. The increasing quality and availability of low-cost electroencephalogram (EEG) systems offer possibilities for practical fatigue monitoring. However, non-data-driven methods, designed for practical complex situations, usually rely on handcrafted data statistics of EEG signals. To reduce human involvement, we introduce a data-driven methodology for online mental fatigue detection, called Self-Weight Ordinal REgression (SWORE). Reaction Time (RT), referring to the duration taken by humans in reaction to an emergency, is widely considered as an objective behavioral measure for mental fatigue state. Since regression methods adopted in previous literature are sensitive to extreme RTs, we propose an indirect RT estimation based on preferences to explore the relationship between EEG and RT, which generalizes to any scenario when an objective fatigue indicator is available. In particular, SWORE evaluates the noisy EEG signals from multiple channels in terms of two states: shaking-state and steady-state. Modeling the shaking-state can discriminate the reliable channels from the non-informative ones; while modeling the steady-state can suppress the task non-relevant fluctuation within each channel. In addition, an online generalized Bayesian moment matching (Online GBMM) algorithm is proposed to online calibrate SWORE efficiently per participant. Experimental results on 40 participants show that SWORE can maximally achieve consistent with RT, demonstrating the feasibility and adaptability of our proposed framework in practical mental fatigue estimation.

1 Introduction

Mental fatigue is a common physiological phenomenon (Borghini et al., 2014), which induces sub-optimal functioning, and may even lead to accidents with severe consequences (Van Cutsem et al., 2017). The National Highway Traffic Safety Administration estimates that about 100,000 official reports of crashes are the direct result of driver mental fatigue each year which results in an estimated 1,550 deaths, 71,000 injuries, and 12.5 billion in monetary losses. In response to these critical issues, several algorithms have been developed to detect mental fatigue using electro-cardio signal (ECG) (Fallahi et al., 2016), functional Near Infrared Spectroscopy (fNIRS), electrooculogram (EOG) (Laurent et al., 2013), electroencephalographic (EEG) (Lin et al., 2013; Jagannath and Balasubramanian, 2014; Sauvet et al., 2014; Wang et al., 2015), etc. Among these signals, EEG signals are assumed to be most accurate and valid to fetch the information related to driver's mental fatigue, owing to their high temporal resolution and the availability of a vast variety of preprocessing methods (Graimann et al., 2009; Sahayadhas et al., 2012; Palanivel Rajan and Dinesh, 2015).

Previous methods for developing automatic systems, to detect driver drowsiness from EEG signals, can be broadly classified into two categories: non-data driven or data driven. Non-data driven approaches, such as power spectrum based analysis (Jap et al., 2009; Wang et al., 2018), entropy-based analysis (Kar et al., 2010), brain networks based analysis (Li et al., 2017), usually resort to handcrafted estimators, like changes in power or statistically related features, to evaluate mental fatigue using EEG signals from multiple channels (Gurudath and Riley, 2014; Gharagozlou et al., 2015). However, these evaluation metrics require extra expert interpretation and complex calculation processes. Meanwhile, EEG signals are known to be highly specific and vary in great detail among individuals. Thus, non-data driven approaches relying on predefined criteria are not robust enough to account for individual variability and dwindle their applications in practical implementations.

In terms of data-driven mental fatigue evaluation, the reaction time (RT) to a certain assigned task is widely adopted as supervision, to indicate the fatigue level. Some linear (Lin et al., 2010; Resalat and Saba, 2015) and non-linear (Liu et al., 2016; Cui and Wu, 2017; Pan et al., 2020) methods show that it is possible to detect mental fatigue with high accuracy. It is impressive but rather blind to the wealth of the dynamics and behavioral variability (Müller et al., 2008; Ratcliff et al., 2009; Yarkoni et al., 2009; Xu et al., 2018). Although some recent work (Wei et al., 2018; Cui et al., 2019) suggested addressing the concerns of inter- and intra-subject variability through transfer learning, those techniques are available only to offline analysis methods with sufficient training samples.

Previous offline analysis methods often result in poor fatigue detection performance due to limited training data in practical implementation (See Fig.1). For example, either deep learning (Goodfellow et al., 2016) methods, requiring massive training data, or Riemannian methods (Barachant et al., 2012; Congedo et al., 2017), incurring high computation cost, fail to meet the harsh requirement in practical scenario. In addition, the repercussion of mental fatigue, the decline of mental alertness, and bad driving performances, are the reflection of the whole brain among different areas. Recent works demonstrate its efficacy by discriminating functional interactions between dif-

ferent brain regions either based on heuristic metrics (Wang et al., 2018; Richer et al., 2018) or complex analysis (Li et al., 2017). However, it cannot fully reveal functional interactions between multiple channels in terms of mental fatigue since the analysis is independent of the later mental fatigue evaluation.

To address the above concerns, we introduce a data-driven methodology, called Self-Weight Ordinal REgression (SWORE), for online driver mental fatigue detection which models the functional interactions between different brain regions. Instead of formulating our SWORE as a regression task with RT being the direct supervision, we consider a more general problem setting, i.e., learning to rank. SWORE learns from brain dynamics preferences and aims to achieve consistency with RT indirectly in the sense of ranking. The brain dynamics preferences can be constructed via some objective fatigue indicator, such as RT if available, or some power spectral features (Wang et al., 2018; Bose et al., 2019). Preferences-based indirectly mental fatigue evaluation is proved to alleviate the overfitting issue of directly predicting RT in a regression task (Pan et al., 2020). In particular, SWORE models the brain dynamic preferences in terms of two states: shaking state and steady state. SWORE automatically discriminates the reliable channels from the non-informative ones by modeling the shaking state and suppresses the mental fatigue non-relevant fluctuation within each channel by modeling the steady state. Moreover, an online generalized Bayesian moment matching (Online GBMM) algorithm is proposed for Bayesian posterior update. Once a new sample (the reaction time corresponding to the newly recorded EEG signals) is available, Online GBMM would efficiently calibrate the SWORE model with the simple update rules. In summary, the main contributions of this paper are as follows:

- We propose an online mental fatigue monitoring system that can evaluate mental fatigue momentarily with a high prediction performance.
- We propose a Self-Weight Ordinal REgression (SWORE) model to reliably aggregates brain dynamics related preferences from multiple noisy channels in terms of two states: shaking state and steady state.
- We propose an online generalized Bayesian moment matching (Online GBMM) algorithm for online calibrating the SWORE model with the analytic update rules.
- We conduct comprehensive experimental results on 40 participants to verify the reliability of our system in online mental fatigue monitoring scenarios. Further, we explore the parameter sensitivity and model uncertainty of SWORE w.r.t. the Online GBMM algorithm.

The sequel of this paper is organized as follows. Section 2 introduces the background of mental fatigue monitoring and motivates the practice of online mental fatigue monitoring. In Section 3, an indirect mental fatigue monitoring model, Self-Weighted Ordinal REgression (SWORE), is introduced to model the heterogeneous brain dynamic preferences in terms of two states. Section 4 describes an analytic update strategy for online calibrating the SWORE model. Section 5 discusses the details of mental fatigue evaluation in the online scenario. Section 6 demonstrates the reliability of the proposed SWORE model with EEG signals collected from forty participants. Section 7 concludes the paper and envisions future work.

2 Background and Problem Statement

In the following, we introduce the background of mental fatigue monitoring and then discuss the deficiency of previous approaches in the online scenario. Further, several sub-goals are listed which are necessary for achieving a robust online mental fatigue evaluation model.

2.1 Mental Fatigue Monitoring

First of all, the reaction time (RT) to an emergence is generally accepted as the most intuitive and resourceful metric to evaluate mental fatigue. Meanwhile, the EEG (Lin et al., 2013; Wang et al., 2015) signals as the feature vector are adopted which is well known to be most accurate and valid to fetch the information related to driver's mental fatigue (Graimann et al., 2009; Sahayadhas et al., 2012; Palanivel Rajan and Dinesh, 2015), compared to ECG, fNIRS) (Nguyen et al., 2017), etc.

Therefore, a common practice for mental fatigue monitoring is to build a learning model that can predict humans' reaction time to an emergency using the EEG signals recorded beforehand (Lal et al., 2003; Kohlmorgen et al., 2007; Soon et al., 2008; Jap et al., 2009).

2.2 Impaired Performance on Non-stationary Brain Dynamics in Online Applications

Some of previous works derived from linear (Resalat and Saba, 2015; Lin et al., 2010) and non-linear (Liu et al., 2016; Cui and Wu, 2017; Pan et al., 2020) methods show that it is possible to detect mental fatigue with high accuracy. It is impressive but it would rather go blind to the wealth of the dynamics and behavioral variability (Müller et al., 2008; Ratcliff et al., 2009; Yarkoni et al., 2009; Wei et al., 2018; Cui et al., 2019). Particularly, brain dynamics are non-stationary, which is characterized by significant trial-by-trial and subject-by-subject variability (Ratcliff et al., 2009; Yarkoni et al., 2009). However, the above methods, designed for offline analysis with sufficient training samples, would result in poor generalization performance in real scenes without efficient online calibration.

For better illustration, we trained three support vector machine regression (SVR)[1] using 20, 40, 60 sequential trials and visualized their prediction performances on the rest trials, respectively. From Fig.1, we can find that: (1) apart from few local mispredictions, SVR can exactly predict the RT on the training trials and is insensitive to the extreme values. It proves that SVR has sufficient fitting capability for mental fatigue monitoring and has superior robustness to extreme values compared to deep regression models (Pan et al., 2020). (2) All SVR models show bad prediction performance[2] on the remaining trials, respectively. This is consistent with our conjecture that previous

[1] We adopted SVR for showcase due to its nonlinear properties and superior generalization performance on small training dataset (Schlkopf et al., 2018). SVR is implemented using the Libsvm with the parameter option "-s 3 -t 2".

[2] The Y-axis is in log scale. And the prediction discrepancy would be more significant in normal scale.

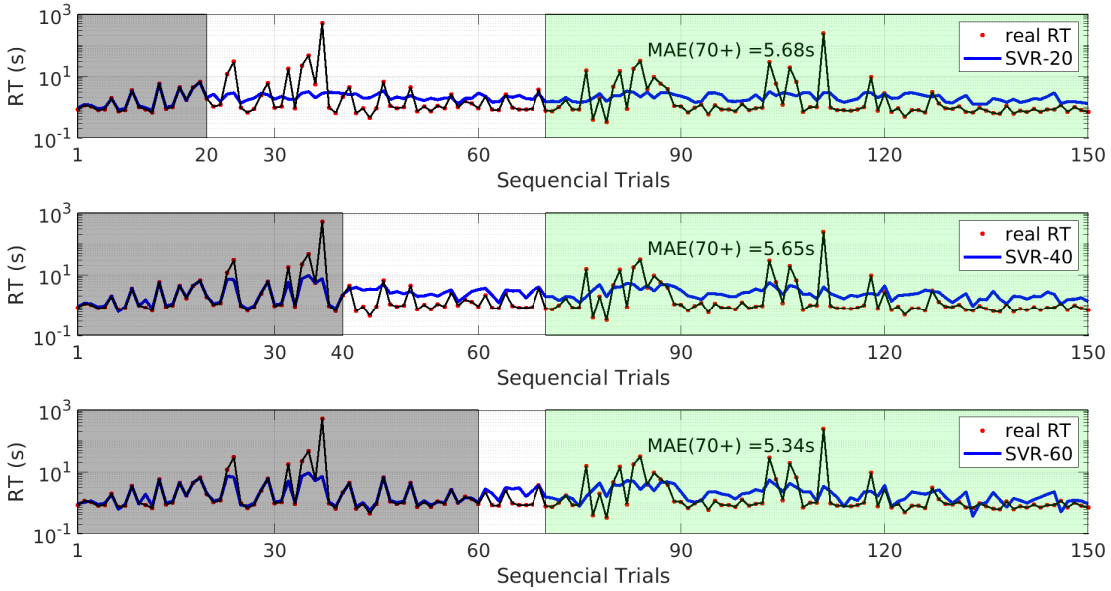

Figure 1: Bad prediction performance of SVR using 20, 40, 60 training trials, respectively. EEG signals from multiple channels are simply concatenated into a long feature vector and SVR is trained using this feature vector. For the sake of fair comparison, we collect the mean absolute error (MAE) of three models on the remaining trials starting from the 71st trial. We only present the result of the first participant for a showcase.

offline analysis will suffer from severe generalization issues in online scenarios. (3) Increasing the training trials marginally improves the prediction accuracy. Particularly, when the number of training trials increases from 20 to 60, the mean absolute error only decreases by 0.34s.

2.3 Online Mental Fatigue Evaluation

In pursuit of online mental fatigue evaluation, a major concern behind previous methods is the computational efficiency in terms of time and memory. Deep learning (Goodfellow et al., 2016) methods, although achieve superior performance (Pan et al., 2020), require massive training data. Meanwhile, Riemannian methods (Barachant et al., 2012; Congedo et al., 2017) achieve good performance with a small amount of training trials, but incur an overhead computational cost. Another important factor among existing proposed methods is the lacking of efficient aggregation mechanism to distill reliable predictions from multiple noisy channels. In particular, majority voting and concatenation are all verified to suffer from overfitting and poor generalization performance (Pan et al., 2020).

Based on the above analysis, we summarize three sub-problems, that will be addressed in this paper, to deduce a robust online mental fatigue evaluation model:

- How to reliably detect metal fatigue using the EEG signals as well as the corresponding RTs?

- How to automatically eliminate the non-informative channels during the learning process?
- How to effectively calibrate the learning model with an EEG signal when its truth RT is available?

We propose our SWORE mode in Sect. 3 to address the first two problems and introduce efficient online calibration strategies in Sect. 4 to answer the third problem.

3 Self-weighted Ordinal Regression for Brain Dynamics

In this section, we evaluate mental fatigue indirectly with brain dynamics related preferences to avoid overfitting to the extreme RTs, instead of modeling it as a regular regression task (Resalat and Saba, 2015; Lin et al., 2010).

3.1 Brain Dynamic Preferences

As shown in Fig. 1, it is usually difficult for a learning model to get the exact estimation of RT since the RT values are not smoothly changing, and the relationship between RT values and fatigue levels are not exact but relative due to time and subject variations. The performance would further worsen in the online setting as only a few training trials are available. Meanwhile, a rough but reliable estimation is acceptable in real-world situations of mental fatigue monitoring (Colosio et al., 2017). Therefore, in this paper, we model the brain dynamics related preferences instead of the exact values of RT.

Remark 1 (From regression to ordinal regression)**.** *Let's revisit the prediction of RT in the perspective of ordinal regression. RT is actually defined in the complete ordered field* $\mathbb{R}$ *which owns its structure meanings. The relative structure information is entirely preserved among the pairwise comparisons of RTs. Therefore, if there exists a learning model which can maximally preserve the whole structure information, a new trial can find its own position (a rough estimation of RT) by its comparisons with previous recorded EEG signals. Please refer to Sect. 5.3 for more details.*

Particularly, instead of modeling the global pattern of the brain dynamics within a regression model, we consider the local discrepancy between the current and next brain dynamic states. First, the difference between RTs is leveraged as the indicator for the local discrepancy,

$$y = \begin{cases} \text{up:} & 1 & RT_t < RT_{t+1} \\ \text{down:} & -1 & RT_t > RT_{t+1} \end{cases}, \tag{1}$$

where RT_t and RT_{t+1} denote the current and next reaction time, respectively. Further, the brain dynamic preference[3] (x_t^n, x_{t+1}^n) (typically a pair of d-dimensional feature vectors) can be constructed with the corresponding pairwise EEG signals recorded from each channel ($\forall n = 1, 2, \cdots, N$). Every EEG sensor used for recording is assumed to

[3] We used the term "preference" intentionally to show that brain dynamics keep changing w.r.t. human behaviour and it happens because the human brain prefers one decision over others.

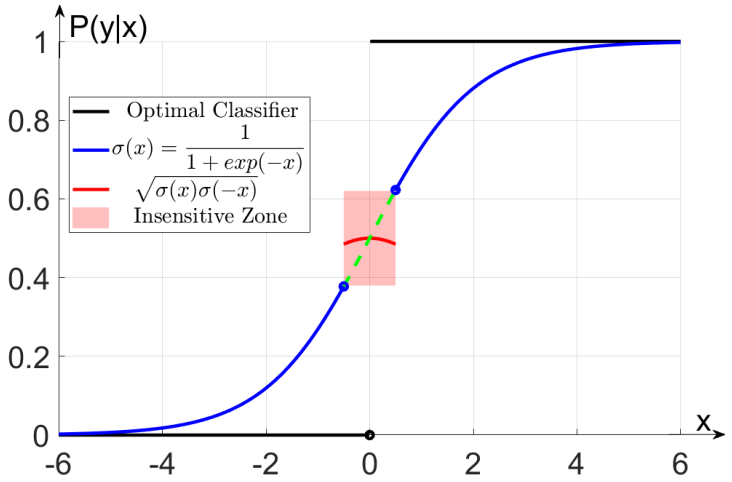

Figure 2: Gradient flattening w.r.t. sigmoid function. The dash line represents the original sigmoid function.

record independently from scalp without influencing other sensors (Homan et al., 1987; Teplan et al., 2002), so the brain dynamic preferences are constructed for each channel independently. Therefore, there are totally N brain dynamic preferences constructed for each comparison.

Remark 2 (Indirect Mental Fatigue Monitoring)**.** *The word "indirect" is adopted for comparing with the usage of RT as the direct supervision in the regression task. Meanwhile, the objective fatigue indicator used for constructing the brain dynamics preferences is not limited to RT. Other well-studied and easily accessible power spectral features (Borghini et al., 2014; Chai et al., 2016), such as dynamic time warping, entropy, and functional connectivity, can also be adopted as fatigue supervision (Wang et al., 2018; Bose et al., 2019) for constructing the brain dynamics preferences.*

Meanwhile, due to individual variability, the mental fatigue criteria defined by a specific RT value vary from person to person. Ranking-based criteria can avoid this since it can capture the normal level by modeling the ordering connection of several EEG signals.

3.2 Heterogeneous Brain Dynamic Preferences

The prediction of brain dynamic preferences can be formulated as a learning-to-rank problem, in which our goal is to estimate the optimal classifier in Fig. 2. Many binary classification models can be adopted to model this problem, e.g., the logistic ordinal regression:

$$P(y|w,x_t^n,x_{t+1}^n) = \sigma(yw^T\Delta x_n), \quad \text{where } \Delta x_n = x_{t+1}^n - x_t^n \text{ and } \sigma(z) = 1/(1+e^{-z}). \quad (2)$$

However, the vanilla logistic classification model suffers from the reliability issue when applied to brain dynamics, since a subtle discrepancy around classification boundary $P(y|x) = 0.5$ leads to the steepest gradient (see Fig. 2). Note the subtle difference

between the RTs may not be caused by the intrinsic difference between two brain dynamics but the unknown noise.

To improve the model stability, we introduce the "insensitive zone", which flattens the steepest gradient around the boundary and therefore enables the classification model less sensitive to the subtle difference between the response times.

Accordingly, we categorize the brain dynamic preferences into two states, according to the discrepancy between the RTs, namely, shaking state ($\mathscr{Y}_1$), where the discrepancy between the brain dynamics is significant, and steady state ($\mathscr{Y}_2$), where the brain dynamics remain stable.

$$y \in \begin{cases} \text{Shaking State } \mathscr{Y}_1 : \begin{cases} \text{up} & RT_t < RT_{t+1} \\ \text{down} & RT_t > RT_{t+1} \end{cases} & |RT_t - RT_{t+1}| > \tau \\ \text{Steady State } \mathscr{Y}_2 & & |RT_t - RT_{t+1}| \leq \tau \end{cases}, \tag{3}$$

where τ is the predefined parameter controlling the model sensitivity.

Shaking State: In terms of the shaking state $y \in \mathscr{Y}_1$, it has two cases: an up ($RT_{t+1} > RT_t$) and a down ($RT_{t+1} < RT_t$), which can be formulated as the learning to rank problem.

However, considering the functions of different regions in the human brain, the relative contributions of different channels to human reaction time may vary a lot. If we simply aggregate the N brain dynamic preferences recorded in different channels without making any distinctions about the channel reliability, the performance of the learning model would inevitably degrade (Pan et al., 2020). Inspired by Raykar et al. (2010), which aggregates the noisy annotations from multiple crowd workers while considering the worker ability, we propose to estimate the reliability of each channel explicitly during the aggregation process. In particular, we formulate a robust pairwise learning to rank model as follows,

$$P(y|w,\pi_{1:N},x_0^{1:N},x_1^{1:N}) = \prod_{n=1}^{N} \left[\pi_n \sigma(yw^T \Delta x_n) + (1-\pi_n)\sigma(-yw^T \Delta x_n)\right] \quad y \in \mathscr{Y}_1. \tag{4}$$

Note that Eq.(4) actually models each brain dynamic preference as the weighted arithmetic mean of two cases. The weight $\pi_n \in [0,1]$, estimated during the training process, denotes the relative contribution of the n-th channel w.r.t. to the learning task, $\forall n = 1,2,\cdots,N$.

Remark 3 (Superiority over the regular weighted average)**.** *From the perspective of EEG channel analysis, Eq.*(4) *provides us a new aggregation mechanism to combine the information from different channels. Different from majority voting which simply categorizes the channels into reliable ones and noisy ones, Eq.*(4) *performs a fine-grained analysis and further categorizes the noisy channels into non-relevant ones and negative reliable ones. Therefore, three types of channels can be recognized with the channel reliability $\pi_n \in [0,1]$, namely positive reliable ones ($\pi_n \to 1^-$)*[4]*, non-relevant ones ($\pi_n \approx 0.5$) and negative reliable ones ($\pi_n \to 0^+$), $\forall n = 1,2,\cdots,N$.* ■

[4] $\pi_n \to 1^-$ denotes π_n is up to approximate 1, while $\pi_n \to 0^+$ denotes π_n is down to approximate 0.

Steady State: On the other hand, the steady state $y \in \mathscr{Y}_2$ denotes the brain dynamic preferences with comparable RTs.

To improve the robustness of the learning model w.r.t. the easily corrupted brain dynamics, a technique called gradient flattening is introduced to model the insensitive zone in Fig. 2. Namely, it flattens the steepest gradient at the classification boundary $P(y|x) = 0.5$, enabling the learning model to be less sensitive to subtle noise. In particular, we model the brain dynamic preferences at steady state as follows:

$$P(y|w, x_0^n, x_1^n) = \sqrt{\sigma(w^T \Delta x_n)\sigma(-w^T \Delta x_n)} \quad y \in \mathscr{Y}_2, \tag{5}$$

which is the geometric mean of an up ($RT_t < RT_{t+1}$) and a down ($RT_t > RT_{t+1}$). Please refer to learning-to-rank literature for other options (Zhou et al., 2008). Furthermore, Eq.(5) can also be integrated to robust aggregation model (Eq.(4)) which considers the channel reliability π_n. Due to the symmetric of Eq.(5), we define

$$P(y|w, \pi_n, x_1^n, x_2^n) \stackrel{\Delta}{=} P(y|w, x_1^n, x_2^n) \quad \forall y \in \mathscr{Y}_2.$$

Remark 4 (Gradient flattening enhances model robustness)**.** *The gradient flattening used in Eq.*(5) *can be understood as a regularization. It enables our model to be robust to the fluctuation between brain dynamics which is not relevant to RTs.* ■

3.3 Self-Weighted Ordinal Regression Model

In summary, our Self-Weighted Ordinal REgression (SWORE) for heterogeneous brain dynamic preferences can be formulated as follows,

$$P(y|w, \pi_{1:N}, x_0^{1:N}, x_1^{1:N}) = \begin{cases} \prod_{n=1}^N \left[\pi_n \sigma(y w^T \Delta x_n) + (1-\pi_n)\sigma(-y w^T \Delta x_n)\right] & y \in \mathscr{Y}_1, \\ \prod_{n=1}^N \sqrt{\sigma(w^T \Delta x_n)\sigma(-w^T \Delta x_n)} & y \in \mathscr{Y}_2. \end{cases} \tag{6}$$

Remark 5 (Reliability of the SWORE model)**. (1) Inter-channel Reliability:** *SWORE only trusts the brain dynamic preferences from (positive and negative) reliable channels. Since SWORE trains a mixture of two complementary classifiers with shared parameter w, it categorizes the channels into positive channels ($\pi_n \to 1^-$), negative channels ($\pi_n \to 0^+$) and non-relevant channels ($\pi_n \approx 0.5$). Based on channel reliability π, SWORE can automatically choose the suitable classifier to extract the correct information from the positive and negative channels and update the shared parameter w accordingly. Further, SWORE ignores the information from the non-relevant channels by assigning a constant likelihood (i.e.* 0.5*) to each brain dynamic preference from the non-relevant channels.*

(2) Intra-channel Reliability: *SWORE only extracts task-related information from each brain dynamic preference. Since the probability of the steady state $y \in \mathscr{Y}_2$ (Eq.*(5)*) does not depend on the channel reliability π_n, gradient flattening actually performs as a regularization on the regression weight w and enables SWORE being robust to the random fluctuations which are widely-existing in brain dynamics.* ■

4 Efficient Online Updating Strategy

As we discussed in Sect. 2.2, brain dynamics are non-stationary. If the SWORE model cannot be updated to date, it would suffer from low generalization performance. Therefore, in this section, we introduce an efficient online updating strategy for the SWORE model. It can update SWORE in good approximation while introducing marginal computation cost.

4.1 Bayesian Moment Matching

Bayesian moment matching (BMM) is a Bayesian approach used to estimate the model parameters. Specifically, it estimates the parameters of the approximated posterior by matching a set of sufficient moments of the exact complex posterior. Moreover, BMM can be naturally extended to the sequential update paradigm for large-scale or streaming datasets, e.g. OnlineBMM (Jaini et al., 2017). That is, the approximated posterior is updated with each sample each time instead of the whole dataset.

First, SWORE is extended to its Bayesian version. Specifically, a Gaussian prior is introduced for weight vector w, i.e. $w \sim N(\mu, \Sigma)$, while a Beta prior is introduced for each channel reliability π_n, namely, $p_0(\pi) = \prod_{n=1}^{N} \text{Beta}(\pi_n|\alpha_n, \beta_n)$. Given a brain dynamic preference (x_t^n, x_{t+1}^n) recorded in the n-th channel with its ordinal supervision y, the posterior of the model parameters can be represented as:

$$P(w, \pi|y, \Delta x_n) = \frac{P(y|w, \pi, \Delta x_n) p_0(w) p_0(\pi)}{P(y|\Delta x_n)}. \tag{7}$$

Note that we only consider the posterior distribution w.r.t. brain dynamic preferences from one channel. Since different channels are modeled independently, the following equations can be easily extended to the posterior distribution w.r.t. the brain dynamic preferences from all channels.

The main issue with Eq.(7) is that the joint posterior distribution $P(w, \pi|y, \Delta x_n)$ is complicated or even intractable. To keep the computation tractable, we adopt the mean-filed assumption and project the posterior into the same form with the prior (product of a Normal with Betas, i.e. $P(w, \pi|y, \Delta x_n) \approx q(w)q(\pi) = N(w|\mu, \Sigma) \prod_{n=1}^{N} \text{Beta}(\pi_n|\alpha_n, \beta_n)$). Then the posterior parameters are estimated by matching a set of sufficient moments of the approximate posterior with the exact posterior:

- Match the moments between $q(w)$ and $P(w|y, \Delta x_n)$: $\int w q(w) dw = \int w P(w|y, \Delta x_n) dw$ and $\int w w^T q(w) dw = \int w w^T P(w|y, \Delta x_n) dw$. Due to the non-conjugation between the marginalized likelihood $P(y|w, \Delta x_n)$[5] and the normal prior $N(w|\mu, \Sigma)$, the posterior $P(w|y, \Delta x_n)$ is complex. Therefore, the posterior parameters $(\mu^{new}, \Sigma^{new})$ cannot be computed analytically because of the intractability of the integrals in the moment constraints.

- Match the moments between $q(\pi)$ and $P(\pi|y, \Delta x_n)$: $\int \pi_n q(\pi) d\pi = \int \pi_n P(\pi|y, \Delta x_n) d\pi$ and $\int \pi_n^2 q(\pi) d\pi = \int \pi_n^2 P(\pi|y, \Delta x_n) d\pi$, $n = 1, 2, \cdots, N$. Fortunately, we can solve the moment constraints with closed-form integrals, and get the posterior parameters $(\alpha_n^{new}, \beta_n^{new}), \forall n = 1, 2, \cdots, N$ accordingly.

[5] $P(y|w, \Delta x_n) = E_{\text{Beta}(\pi|\alpha,\beta)}[P(y|w, \pi, \Delta x_n)]$.

4.2 Generalized Bayesian Moment Matching

Inspired by the Bayesian approximation method proposed by Weng and Lin (2011), which extended the Stein's Lemma (Woodroofe, 1989), we propose to estimate the posterior parameters $(\mu^{new}, \Sigma^{new})$ of the approximate posterior $q(w)$ by differential operations instead of integral operations. Therefore, the BMM algorithm is extended to a general situation where the likelihood function is twice differentiable.

Theorem 1. *Assume $f(w)$ is the marginalized likelihood of one brain dynamic preference and almost twice differentiable. Upon updating this preference, the posterior parameters $(\mu^{new}, \Sigma^{new})$ of weight w can be estimated as:*

$$\mu^{new} \approx \mu + \Sigma \times \frac{d \log f(w)}{dw}\Big|_{w=\mu}, \tag{8a}$$

$$\Sigma^{new} \approx \Sigma + \Sigma \times \frac{d^2 \log f(w)}{dw dw^T}\Big|_{w=\mu} \times \Sigma. \tag{8b}$$

We set $w = \mu$ as we expect that the posterior density of w to be concentrated on μ (Weng and Lin, 2011). Refer to the appendix for the detailed proof of Theorem 1. ■

In the following, we resort to Generalized Bayesian Moment Matching (GBMM) method to estimate the posterior parameters. We take the brain dynamic preference (x_t^n, x_{t+1}^n) at the shaking state $y \in \mathscr{Y}_1$ as an example. The equations can be easily extended to the brain dynamic preference at the steady state.

In the following, we first update the hyperparameters (μ, Σ) of w, then update the hyperparameter (α_n, β_n) of π_n $\forall n = 1, 2, \cdots, N$. To update w, we integrate out π_n to obtain the marginalized likelihood function $f(w)$:

$$f(w) = \int P(y|w, \pi, \Delta x_n) \text{Beta}(\pi|\alpha, \beta) d\pi = \frac{\alpha_n}{\alpha_n + \beta_n} \sigma(y w^T \Delta x_n) + \frac{\beta_n}{\alpha_n + \beta_n} \sigma(-y w^T \Delta x_n).$$

According to Eq.(8a) in Theorem 1, we can update μ as follows:

$$\mu^{new} \approx \mu + \Sigma \times \frac{d \log f(w)}{dw}\Big|_{w=\mu} = \mu + (A - a) \times \Sigma \times \Delta x_n. \tag{9}$$

where $A = \frac{\alpha_n}{\alpha_n + \beta_n e^{-\mu^T \Delta x_n}}$ and $a = \frac{1}{1 + e^{-\mu^T \Delta x_n}}$. According to Eq.(8b) in Theorem 1, we can update Σ as follows:

$$\Sigma^{new} \approx \Sigma + \Sigma \times \frac{d^2 \log f(w)}{dw dw^T}\Big|_{w=\mu} \times \Sigma \approx \Sigma + \kappa \mathbf{I} + [A(1-A) - a(1-a)] \times \Sigma \times \Delta x_n \Delta x_n^T \times \Sigma. \tag{10}$$

where κ is a small positive value to ensure a positive definite variance matrix. $\mathbf{I}$ is an identity matrix.

Similarly, to update π, we first integrate out w to obtain the marginalized likelihood $f(\pi_n)$, $\forall n = 1, 2, \ldots, N$ for each preference:

$$\begin{aligned} f(\pi_n) &= \int P(y|w, \pi_n, \Delta x_n) N(w|\mu, \Sigma) dw \\ &= \pi_n \times E_{N(w|\mu,\Sigma)}[\sigma(y w^T \Delta x_n)] + (1 - \pi_n) \times E_{N(w|\mu,\Sigma)}[\sigma(-y w^T \Delta x_n)]. \end{aligned} \tag{11}$$

Let $R_1 = E_{N(w|\mu,\Sigma)}[\sigma(yw^T\Delta x_n)]$, we calculate R_1 by the second order Taylor approximation of $\sigma(yw^T\Delta x_n)$ at μ. Then we have $R_2 = E_{N(w|\mu,\Sigma)}[\sigma(-yw^T\Delta x_n)] = 1 - R_1$, and the normalization constant $P(y|\Delta x_n)$ can be represented as follows,

$$R = P(y|\Delta x_n) = \int f(\pi_n)\text{Beta}(\pi_n|\alpha_n,\beta_n)d\pi = \frac{\alpha_n R_1 + \beta_n R_2}{\alpha_n + \beta_n}. \tag{12}$$

According to Bayesian theorem, the posterior distribution of π_n is $P(\pi_n|y,\Delta x_n) = \frac{f(\pi_n)\text{Beta}(\pi_n|\alpha_n,\beta_n)}{R}$, the moments $E[\pi_n]$ and $E[\pi_n^2]$ w.r.t. to $P(\pi_n|y,\Delta x_n)$ can be computed as follows:

$$E_{P(\pi_n|y,\Delta x_n)}[\pi_n] = \frac{R_1(\alpha_n+1)\alpha_n + R_2\alpha_n\beta_n}{R(\alpha_n+\beta_n+1)(\alpha_n+\beta_n)}, \tag{13a}$$

$$E_{P(\pi_n|y,\Delta x_n)}[\pi_n^2] = \frac{\alpha_n(\alpha_n+1)[R_1(\alpha_n+2) + R_2\beta_n]}{R(\alpha_n+\beta_n+2)(\alpha_n+\beta_n+1)(\alpha_n+\beta_n)}, \tag{13b}$$

where $n = 1,2,\cdots,N$. See appendix for detailed derivations. Then we can update the hyperparameter (α_n,β_n) of π_n as follows:

$$\alpha_n^{new} = \frac{(E[\pi_n] - E[\pi_n^2])E[\pi_n]}{E[\pi_n^2] - (E[\pi_n])^2}, \tag{14a}$$

$$\beta_n^{new} = \frac{(E[\pi_n] - E[\pi_n^2])(1 - E[\pi_n])}{E[\pi_n^2] - (E[\pi_n])^2}. \tag{14b}$$

where we omit the subscript $P(\pi_n|y,\Delta x_n)$ of the expectation operator for simplicity.

4.3 Online GBMM for the Calibration of SWORE

According to the above analysis, we summarize an Online GBMM for SWORE in Algorithm 1. It is notable that both the weight update and channel reliability update can be completed following analytic rules (Eq.(9), (10),(14a), (14b)). As a result of the efficient posterior updating procedure, Online GBMM enables SWORE naturally to handle streaming preferences.

Algorithm 1 Online Generalized Bayesian Moment Matching for SWORE

1: **Sequential Input:** Brain Dynamic Preference $\Delta x_n = x_{t+1}^n - x_t^n$ and y.
2: **if** the first iteration **then**
3: *Initialization:* initialize hyperparameters (μ,Σ) and $\{\alpha_n,\beta_n\}_{n=1}^N$.
4: **end if**
5: *Mean-field Assumption:* define $q(w)q(\pi)$ in the same form as the prior.
6: *GBMM for $q(w)$:* update the posterior parameters (μ,Σ) using Eq.(9),(10).
7: *GBMM for $q(\pi)$:* update the posterior parameters (α_n,β_n) using Eq.(14a), (14b).
8: *Replacement:* replace the prior $P(w)P(\pi)$ with the approximate posterior $q(w)q(\pi)$.
9: **Sequential Output:** Weight $w = \mu$, Channel reliability $\pi_n = \frac{\alpha_n}{\alpha_n+\beta_n}$, $n = 1,2,\ldots,N$.

5 Online Mental Fatigue Evaluation

In this section, we apply the SWORE model (Eq.(6)) to perform online mental fatigue monitoring. First, we introduce the data augmentation tricks to address the low data volume in the online scenario. Then, we propose to maintain a Brain Dynamic Table (BDtable), which sequentially stores the representative EEG signals. Finally, we summarize the whole framework for online mental fatigue evaluation.

5.1 Blank-out Noise Model for Data Augmentation

Due to the limited size of available trials, the learning model is prone to be overfitting during the training process. Therefore, we adopt the data augmentation trick in this paper. Data augmentation trick[6] replaces the original EEG signals with T corrupted versions from the predefined corrupting distribution $P(\Delta\tilde{x}|\Delta x_n)$. For simplicity, we are going to focus on the blank-out noise model (a.k.a dropout) as the corrupting distribution, which randomly omits subsets of neurons (or features). More precisely,

$$P(\Delta\tilde{x}_l|\Delta x_l;\theta) = \begin{cases} \theta & \Delta\tilde{x}_l = 0 \\ 1-\theta & \Delta\tilde{x}_l = \Delta x_l \end{cases}, \tag{15}$$

where $\Delta\tilde{x}_l \in \{0, \Delta x_l\}\ \forall l = 1, 2, \cdots, d$ and d is the feature dimension.

Note that each dimension of the input Δx_n is corrupted independently. Eq.(15) is also a promising technique to break up the complex co-adaptations, caused by high correlation among different dimensions of the EEG signals (either in time domain or frequency domain). Since the presence of any particular dimension is unreliable, each dimension cannot rely on other specific dimensions to correct its mistakes. It must perform well in a wide variety of different contexts provided by the other dimensions.

Therefore, adding data augmentation (i.e., Eq.(15)) before the line 5 of Algorithm 1 can improve the generalization of our SWORE in more complex situations (See Fig. 10).

5.2 Online Reservoir Sampling for BDtable

Our SWORE model requires brain dynamics related preferences, which are constructed using the current EEG signals and previously observed ones, for an update. Accordingly, Brain Dynamic table (BDtable) is introduced to store the EEG signals, which can help to calibrate our evaluations and guide the model update process. Considering the requirement of the high computational efficiency in online applications, BDtable should be a good summarization of previous seen EEG signals.

Since no prior knowledge about each subject is available, we propose to build BDtable with random sampling, namely each element of the BDtable is uniformly sampled from the EEG signals seen so far. Particularly, reservoir sampling is proven to meet the requirement on BDtable (Vitter, 1985). It is then carried out to sequentially maintain the BDtable following Algorithm 2, where S denotes the number of BDtable.

[6] Although the data augmentation procedure generates a corrupted dataset with a larger size, the final computational cost (scaling linearly with T) is acceptable benefiting from the efficient updating rules.

Algorithm 2 Online Reservoir Sampling

while $t < S$ **do**
  1) store the t-th EEG signal to the BDtable.
end while
for $t > S$ **to** $+\infty$ **do**
  2) save the t-th EEG signal to the BDtable with probability S/t, else discard it.
  3) use it to replace one EEG signal randomly sampled from the BDtable.
end for

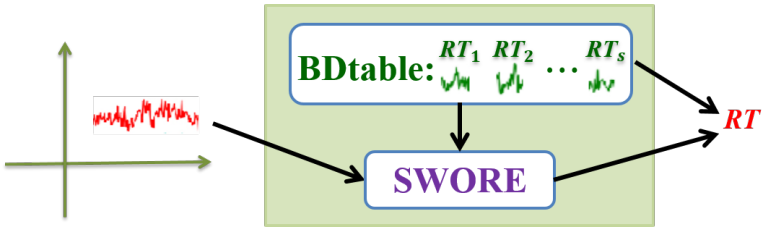

Figure 3: Online Mental Fatigue Evaluation.

5.3 Whole Framework for Online Mental Fatigue Evaluation

Assume the SWORE model $\{w, \pi_{1:N}\}$[7] and the BDtable $\{x_i^{1:N}, RT_i\}_{i=1:S}$ are both updated to time t-1 following Algorithm 1 and Algorithm 2, respectively. Then, the online mental fatigue monitoring refers to predicting RT_t with the EEG signals $x_t^{1:N}$, extracted at time t, using the up-to-date SWORE model and BDtable.

As stated in Section 3.1, the relative ordinal structure of RT is revealed through the brain dynamic related preferences, which are maximally preserved by our SWORE model $\{w, \pi_{1:N}\}$. Particularly, the relative ordinal structure of RT is consistent with that of $w^T x$ according to the definition of the SWORE model, namely

$$RT_i > RT_j \Longleftrightarrow \begin{cases} w^T x_i^n > w^T x_j^n & \pi_n > 0.5, \\ w^T x_i^n < w^T x_j^n & \pi_n < 0.5. \end{cases}$$

or, we can simply formulate it as

$$RT_i > RT_j \Longleftrightarrow \mathrm{sgn}(\pi_n - 0.5) w^T x_i^n > \mathrm{sgn}(\pi_n - 0.5) w^T x_j^n, \tag{16}$$

where i, j denote the index of different trials, and n represents the index of channels.

Then, we compare the newly collected EEG signals $x_t^{1:N}$ to the reference EEG signals stored in the BDtable following Eq.(16) and derive N full ranking lists over $S+1$ trials regarding each channel, respectively.

$$\begin{cases} \mathrm{sgn}(\pi_1 - 0.5) w^T \times \{x_t^1, x_1^1, x_2^1, \ldots, x_S^1\} & n = 1, \\ \mathrm{sgn}(\pi_2 - 0.5) w^T \times \{x_t^2, x_1^2, x_2^2, \ldots, x_S^2\} & n = 2, \\ \qquad \cdots & \\ \mathrm{sgn}(\pi_N - 0.5) w^T \times \{x_t^N, x_1^N, x_2^N, \ldots, x_S^N\} & n = N. \end{cases} \tag{17}$$

[7] The parameters $\{w, \pi_{1:N}\}$ are used to represent the SWORE model, since Eq.(6) is fully determined by $\{w, \pi_{1:N}\}$. Meanwhile, we omit the subscript t-1 for convenience.

Let $\text{Sort}(x_t^n)$ output the ranking position of the estimated RT_t for the EEG signals x_t^n in terms of the n-th channel, compared to EEG signals stored in the BDtable. The ranking position of the estimated RT_t over all N channels can be derived by aggregating the results from all channels while considering the channel reliability. Namely,

$$\text{Sort}(x_t^{1:N}) = \sum_{n=1}^{N} \frac{|2\pi_n - 1| \times \text{Sort}(x_t^n)}{\sum_{k=1}^{N} |2\pi_k - 1|}. \tag{18}$$

Correspondingly, we sort the S reaction times $\{RT_i\}_{i=1:S}$ stored in the BDtable and derive the full ranking list. Following the consistency of the relative ordinal structure between RTs and brain dynamics related preferences (Eq.(16)), the corresponding RT_t for the newly recorded EEG signals $x_t^{1:N}$ can be estimated by the average of the RTs with the ranking position being close to $\text{Sort}(x_t^{1:N})$.

Overall, the whole framework of online mental fatigue evaluation are summarized in the following (See Fig. 3). In the first S trials, we build the BDtable with the S EEG signals and their corresponding RTs and then initialize the SWORE. For a newly collected EEG signal $x_t^{1:N}$, the SWORE model conducts the indirect mental fatigue evaluation by giving a coarse estimation of RT_t following Eq.(18). When the real reaction time RT_t is available, we calibrate the SWORE model following Algorithm 1 and online update the BDtable following Algorithm 2.

5.4 Complexity Analysis of the Whole Framework

In this section, we analyze the space complexity and computational complexity of our framework. In particular, let d, S, and N denote the dimension of the feature vector, the size of BDtable, and the number of channels, respectively. Meanwhile, we use T to denote the number of data augmentation in Eq.15 and M to denote the number of sequential trials.

Space Complexity: the storage of the online system consists of two parts: $\mathcal{O}(d+N)$ for the model parameters $(\mu, \Sigma, \{\alpha_n, \beta_n\}_{n=1}^N)$[8] and $\mathcal{O}(SNd)$ for the BDtable $\{x_i^{1:N}, RT_i\}_{i=1:S}$. Therefore, the overall space complexity is $\mathcal{O}(SNd)$.

Computational Complexity: we analyze the complexity of the online system from the aspects of prediction and calibration, respectively.

- **Prediction Complexity.** The predication consists of two steps, i.e., Eq(17) and Eq.(18). The computational complexity is $\mathcal{O}(SNd)$ for Eq(17) and $\mathcal{O}(SN\log S)$ for Eq(18). Therefore, the overall computational complexity of predication for M sequential trials is $\mathcal{O}(MSN(d+\log S))$.
- **Calibration Complexity.** According to Algorithm 1, the most time-consuming step is the matrix multiplication in Eq.(10) with $\mathcal{O}(d^3)$ time complexity. Therefore, the overall time complexity for M calibration steps is $\mathcal{O}(MTSNd^3)$. It would further decrease to $\mathcal{O}(MTSNd)$ if a diagonal covariance matrix is adopted.

[8] Σ is simplified to be a diagonal matrix in the experiment for the sake of simplicity.

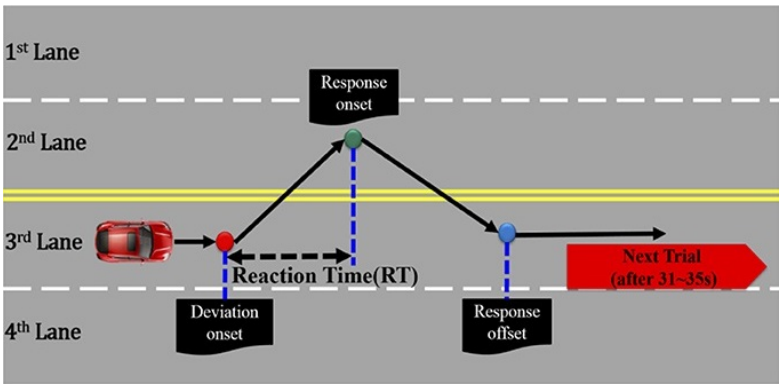

Figure 4: Event-related lane-departure driving paradigm

According to our analysis, the proposed online fatigue monitoring system is both space- and time-efficient since they are both linear related to each factor.

6 Numerical Experiments

In this section, we first introduced the experimental setup of mental fatigue monitoring. Then, we explored the reliability of our SWORE model in online mental fatigue evaluation tasks. Furthermore, we analyzed the parameter sensitivity and the model uncertainty of SWORE w.r.t. the proposed Online GBMM algorithm.

6.1 Experiment Setup

Data Collection: This paper uses the EEG data introduced in Huang et al. (2009). Forty healthy male adults aged 20-30 years were recruited to participate in the sustained-attention driving experiment in a virtual driving simulating environment (Fig. 4). All subjects participated in the sustained-attention driving experiment for 1.5h in the afternoon (13:00–14:00) after lunch. At the beginning of the experiment, a 5 min pre-test was performed to ensure that every subject understood the instructions and they did not suffer from simulator-induced nausea. During this 90 min sustained attention driving task, the experimental paradigm simulated a nighttime driving situation on a four-lane highway and lane-departure events were randomly triggered to make the car drift away from the original cruising lane toward the left side or the right side. Each participant was instructed to quickly compensate for this perturbation by steering the wheel. A complete trial in this study, including 1s baseline, deviation onset, response onset, and response offset, is shown in Fig. 4. EEG signals were recorded simultaneously. The next trial occurs within an interval of 5-10s after the completion of the current trial in which the subject has to drive back to the centerline of the third car lane. If a subject fell asleep during the experiment, there was no feedback to alert him up. For each trial t, the

10s EEG signals $\{x_{n,t}\}_{n=1}^{N}$ from N(=33) [9] different EEG channels before the deviation onset were recorded simultaneously and the corresponding reaction time RT_t was also collected afterward.

A wired EEG cap with 33 Ag/AgCl electrodes, including 30 EEG electrodes, two reference electrodes (A1 and A2), and one vehicle position channel (VP), was used to record the electrical activity of the brain from the scalp during the driving task. The EEG electrodes were placed according to a modified international 10-20 system. The contact impedance between all electrodes and the skin was kept $<$ 5k. The EEG recordings amplified by Scan SynAmps2 Express system (Compumedics Ltd., VIC, Australia) were digitized at 500Hz (resolution: 16bits). Before data analysis, the raw EEG data were prepossessed by the following steps: first, we used a digital band-pass (1-50 Hz) zero-phase FIR filter (the eegfilt.m routine from the EEGLAB toolbox) to remove the power line noise and low-frequency drift. Second, the signals are downsampled to 250Hz to reduce the volume of data. Finally, we did a manual removal of some artifacts such as random and persistent disturbance from body motion, eye movement, eye blinking, muscle activity, EEG channel malfunction, and environmental noise.

Data Preprocessing and Preferences Construction: following Huang et al. (2015); Pan et al. (2020), we preprocessed the EEG signals as follows. Considering the time delay among the channels in the time domain, Fourier transforms (Welch, 1967) has been applied to EEG signals to transform time-series into the frequency domain. Further, to avoid overhead computation, EEG power within 0-30Hz has been selected, which is considered to be the most relevant to the RTs (Huang et al., 2015).

Two types of preferences were constructed following Pan et al. (2020). The shaking state preferences $\mathscr{Y}_1$ were constructed with RT comparisons (RT_0, RT_1), namely $(RT_0 < RT_1)$, where $RT_0 < \min(RT_0 + \tau_1, \tau_2 * RT_0) < RT_1$, and vice versa for $RT_0 > RT_1$. The steady state preferences $\mathscr{Y}_2$ were constructed with RT comparisons (RT_0, RT_1), namely $(RT_0 < RT_1)$, where satisfies $RT_0 < RT_1 < \min(RT_0 + \tau_3, \tau_4 * RT_0)$, and vice versa when $RT_0 > RT_1$. It is notable that $\tau_1 > \tau_3 > 0$ and $\tau_2 > \tau_4 > 1$ control the sensitivity of mental fatigue evaluation. And we empirically set $\tau_1 = 0.15; \tau_2 = 1.2; \tau_3 = 0.1; \tau_4 = 1.1$ for all participants in our experiment.

In terms of the scenarios where RT is not available, other well-studied power spectral features can also be adopted as fatigue indicators. For example, dynamic time warping used in (Wang et al., 2018; Bose et al., 2019), proved to be consistent with RT, can also be adopted for constructing the brain dynamics preferences.

Evaluation Metric: Since mental fatigue monitoring is formulated as an ordinal regression task, the efficacy of SWORE can be evaluated via the consistency between its prediction and RTs in terms of order maintenance. In particular, we adopted the Wilcoxon-Mann-Whitney statistics (Yan et al., 2003) and calculated the prediction ac-

[9] It consists of 30 EEG channels, two reference channels, and one vehicle position channel. We did not eliminate the three non-EEG channel beforehand to demonstrate that our SWORE can automatically remove this kind of non-informative EEG channel during the training.

curacy of each trial as follows:

$$ACC = \frac{1}{S}\sum_{s=1}^{S}\mathbb{I}(y_s = \hat{y}_s), \quad \hat{y}_s = \mathrm{sgn}\Big(\sum_{n=1}^{N}\hat{y}_s^{(n)}\big[\mathbb{I}(\pi_n > \kappa) - \mathbb{I}(\pi_n < 1-\kappa)\big]\Big),$$

where S is the number of EEG signals stored in the BDtable and N is the number of channels. $\mathbb{I}()$ is an indicator that returns one if the argument is valid and returns zero otherwise. y_s is the ground truth order between the new trial and the s-th EEG signal in BDtable (See Eq.(3)), which is derived from RT or other fatigue indicators. $\hat{y}_s^{(n)} = 2\sigma(w^T\Delta x_{s,n}) - 1$ is the predicted order for the brain dynamic preference $(x_0^n, x_1^n)_s$ from the n-th channel: 1 denotes an up and -1 denotes a down. Then, we derive the final order over by aggregating the predictions of all N channels via majority voting. We only consider the prediction from reliable channels only, where a channel is recognized as a reliable channel if it satisfies $\pi_n > \kappa$ or $\pi_n < 1-\kappa$. κ is set to 0.85 in this paper.

$ACC \in [0,1]$ denotes the consistency between the prediction and the ordering of the response time, the higher the better. $ACC = 1$ means the learning model can correctly capture the level of current mental fatigue based on the reference EEG signals. $ACC = 0.5$ means the learning model can capture nothing about the level of current mental fatigue. $ACC = 0$ means the learning model can capture the level of current mental fatigue based on the reference EEG signals but in the reverse order. Ranking-based metric ACC is a relaxation of Mean Absolute Error (MAE). ACC is more robust to extreme RTs and local perturbation around RT compared to MAE. The optima status derived by MAE is also the optima for ACC, while it does not valid vice verse. Therefore, achieving the optima of ACC should be much easier than that of MAE.

Baselines: we only consider the data-driven mental fatigue evaluation approaches in previous literature. Among the regression methods, Support Vector Regression (SVR) (Bose et al., 2019) and neural network-based regression (Pan et al., 2020) are verified to achieve superior performance. Since neural network-based regression requires considerable large samples for training, we consider the SVR (Bose et al., 2019) where a small amount of samples is sufficient for training. In terms of classification methods, CArank (Pan et al., 2020), which requires large training samples, is not suitable for our online scenario. Alternatively, we consider Support Vector Machine (SVM), Random Forest, and Logistics Ordinal Regression (LOR). LOR is a special case of SWORE which only models the shaking state ($\mathscr{Y}_1$).

All baselines, i.e., SVR, SVM, Random Forest, LOR, and our SWORE were implemented with Matlab. In particular, we adopt the RBF kernel for SVR and SVM following (Bose et al., 2019). Since SVR SVM, Random Forest, and LOR have no mechanism to evaluate the channel state beforehand, we simply concatenate the EEG signals from all channels as the feature vector. It is proved to achieve superior performance than aggregating the output from different channels using majority voting. For all the methods, we use the first 20 trials for pretraining. In terms of SVR, SVM, and Random Forest, we fixed these models after pretraining since no efficient online calibration methods are available. In terms of LOR and SWORE, they can be efficient online calibrated with the update strategy introduced in Sect. 4. The size of BDtable is set to 10 for LOR and SWORE, i.e., $S = 10$, in Algorithm 2. For the sake of fair

comparison, we calculate the prediction accuracy of all methods regarding each trial using the same dynamic updated BDtable, respectively. Note two kinds of LOR, i.e., LOR, denoting LOR without online calibration, and Online LOR, denoting LOR with online calibration, are considered for better comparison.

6.2 Comparison with Offline Regression/Classification Methods

Following the online mental fatigue evaluation framework proposed in Fig. 3, we explored the reliability of SWORE in the online monitoring scenario. Specifically, we leveraged the pre-recorded 20 trials to pre-train the embryonic SVR, SVM, Random Forest, LOR, and SWORE, respectively. In terms of SWORE, we randomly initialized μ in $[-10^{-2}, 10^{-2}]$, Σ in $[\mathbf{0}, 10^{-4} \times \mathbf{I}]$ and set $\alpha_n = \beta_n = 5$ according to our parameter sensitivity analysis in Section 6.5. The data augmentation size T is set to 1 during pretraining and 3 during online updating according to Fig. 10 since we encountered a sufficient and insufficient scenario, respectively. The brain dynamics table size is fixed to 10. Therefore, we sequentially get the coarse estimation of RT for each new trial, collect the prediction accuracy, and update the SWORE model when the truth RT is available. We ran our SWORE model following the procedure in Fig. 3 for 100 times and recorded the prediction accuracy for each EEG signal accordingly.

Table 1: Comparison of average prediction accuracy (with 95% confidence interval) over 100 independent running.

Method		Average Prediction Accuracy
Offline Regression	SVR	$69.1 \pm 0.36\%$
Offline Classification	SVM	$67.4 \pm 0.26\%$
	Random Forest	$67.7 \pm 0.39\%$
	LOR	$63.1 \pm 0.35\%$
Online Classification	LOR	$72.6 \pm 0.33\%$
	SWORE	$\mathbf{76.0 \pm 0.30}\%$

From Table 1, we can find that: (1) Offline Regression method, i.e., SVR, achieves higher prediction accuracy than Offline Regression methods, i.e., SVR, Random Forest, and LOR. It is consistent with the result in Sec. 6.4, which makes sense since a classification task is easier to overfit on a small dataset than a regression task. (2) Online calibration is necessary for reliable mental fatigue evaluation. Both online methods, i.e., SWORE and LOR, achieve significant improvement over the offline regression/classification tasks. (3) Our SWORE achieves the highest prediction performance than all baselines.

6.3 Online Mental Fatigue Evaluation on One Participant

According to the experiment results in Sect. 6.2, the offline regression method, i.e., SVR, achieves superior prediction accuracy than offline classification methods, i.e., SVM, Random Forest, and offline LOR. In the following, we will only consider the

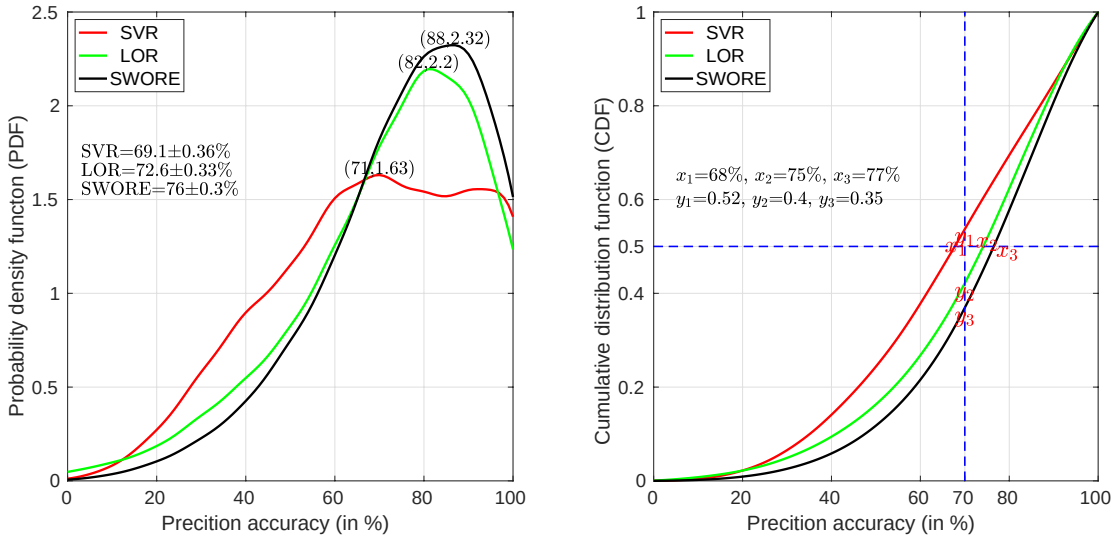

Figure 5: The PDF & CDF of online prediction accuracy with regards to each trial. We only collected the result of the first participant for a showcase.

comparison between the offline regression method SVR and online classification methods, i.e., Online LOR and SWORE. In particular, we conducted more detailed comparisons following the experiment setting in Sect. 6.2.

PDF & CDF of online prediction accuracy: we estimated the probability density function (PDF) and cumulative distribution function (CDF) regarding the prediction accuracy of each EEG signal (See Fig. 5).

From Fig. 5, we can observe that: (1) SWORE gives the most reliable evaluation (76% average prediction accuracy) with the lowest variance for any new EEG signal, comparing to LOR (72.6%) and SVR (69.1%). (2) In terms of half of the samples ($y = 0.5$), SWORE gives the prediction accuracy of more than 77%, while it is 75% for LOR and 68% for SVR, respectively. (3) SWORE gives the prediction accuracy of more than $x = 70\%$ for 65% $(1 - y_3)$ of samples. SVR and LOR are much worse, where only 60% $(1 - y_2)$ and 48% $(1 - y_1)$ of samples can be predicted respectively when requiring the prediction accuracy of more than $x = 70\%$.

To further demonstrate our claim, we conducted the two-sample t-Test under the assumption of equal means without assuming equal variances using the ttest2 function in Matlab. In particular, we conducted the testing between any two of SWORE, LOR, and SVR at the 5% significance level. The comparison results are listed in Table 2.

Equal means without assuming equal variances	Test decision	P-value
SWORE VS. LOR	Reject	2.38e-42
SWORE VS. SVR	Reject	1.039e-153
LOR VS. SVR	Reject	1.03e-37

Table 2: t-Test between at the 5% significance level

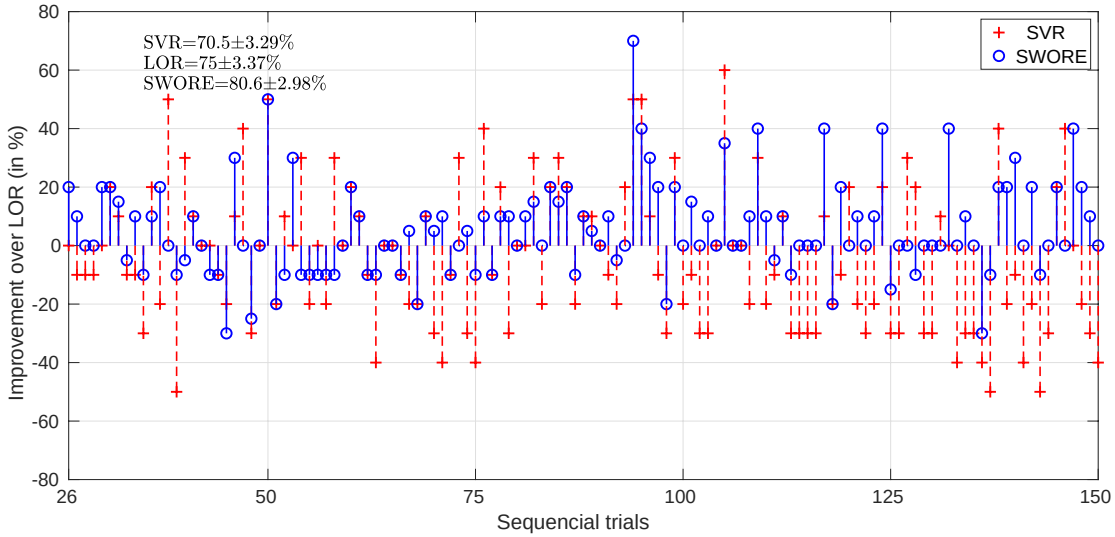

Figure 6: Online Mental Fatigue Evaluation. We used the Online prediction accuracy of LOR as the base, and ploted the improvement of SVR and SWORE over LOR for each trial, respectively. Note the first 25 trials are used to pre-train each model. We only collected the result of the first participant for a showcase.

All test results in Table 2 indicate that t-Test rejects all three null hypotheses at the 5% significance level event without assuming equal variances. Namely, there exists a significant difference between the results of any two of SWORE, LOR, and SVR.

Showcase of online mental fatigue evaluation: To give an intuitive comparison, we calculated the prediction accuracy of each trail for one random experiment with three methods and then plotted performance improvement of SWORE and SVR compared to LOR (See Fig. 6).

Fig. 6 shows that (1) SWORE consistently achieves superior or at least comparable performance compared to LOR and SVR, which demonstrates our claim that channel reliability indeed affects the efficacy of the learning model. (2) In terms of (regression-based) SVR, it suffers from high generalization errors for new EEG signals compared to (classification-based) SWORE and LOR. (3) SWORE achieves an average prediction accuracy of 80.6% for each trial in one random experiment, which is higher than that of LOR (75%) and SVR (70.5%).

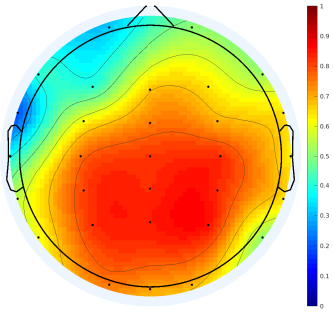

Figure 7: Topoplot visualization of the channels' relative contribution.

Channel reliability estimation: Following our analysis in Remark 3, the model parameter π_n reveals the reliability of the n-th channel, $n = 1, 2, \ldots, N$. Therefore, we visualize the estimated π_n for each channel in Fig. 8. Note the 33-channel EEG data in our experiment consists of 30 EEG channels, two reference channels, i.e., A1 and A2, and the vehicle position channel, i.e., VP. We also visualized the relative contribution of each channel via the 30-channel layout of

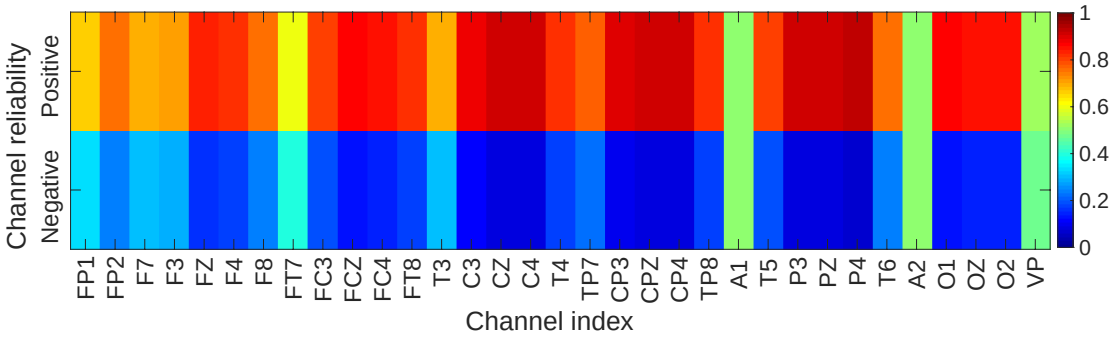

Figure 8: Channel Reliability Estimation and Non-relevant Channel Detection. The color bar denotes the estimated channel reliability. A channel with comparable values in the two rows is considered as a task non-relevant channel. See Remark 3 for more details. We only collect the result of the first participant for a showcase.

Topoplot in Fig. 7. Accordingly to our analysis in Remark 5, both positive and negative channels are all considered as informative and contribute equally to SWORE. We introduce a new metric R ($R = 2 * |\pi - 0.5|$) to denote the contribution of each channel w.r.t. the SWORE model. Similar to π, R ranges between 0 and 1, but a higher value (i.e., $\nearrow 1$) indicates the corresponding EEG channel is more informative.

From Fig. 7 and Fig. 8, we find that (1) the relative contributions of different channels are different, which is consistent with our motivation that different regions of the human brain perform different functions. (2) The majority of channels ($29/33$) are considered as reliable channels. All the three known nonrelevant channels, i.e., A1, A2, and VP, can be automatically detected and removed during the learning process (Pan et al., 2020).

Model Reliability with fewer Channels: Fewer EEG channels is always preferred in the online scenario, which means less computation cost, small storage, and minimal impact on the drivers. Therefore, we explored the reliability of SWORE when fewer channels are available. According to Fig. 8, we retrained SWORE with $5, 10, 20$ randomly selected reliable channels respectively, and compared them with the original SWORE model using all channels in Fig. 9. We collected the prediction accuracy of all variants of SWORE using the same dynamic updated BDtable for a fair comparison.

Fig. 9 shows that all SWORE variants achieve comparable prediction accuracy in terms of overall average or each trial. It proves that: (1) the superior performance of our SWORE model is hardly affected when fewer EEG channels are used. (2) The reliable channels detected by SWORE are trustworthy.

6.4 Ablation study for exploring the efficacy of our contributions

In this paper, we have made contributions, i.e., C1-C4, from different perspectives. To be specific, C1 denotes the robust multi-channel aggregation method in Eq. 4; C2 denotes the gradient flattening in Eq. 5; C3 denotes the efficient online updating strategy

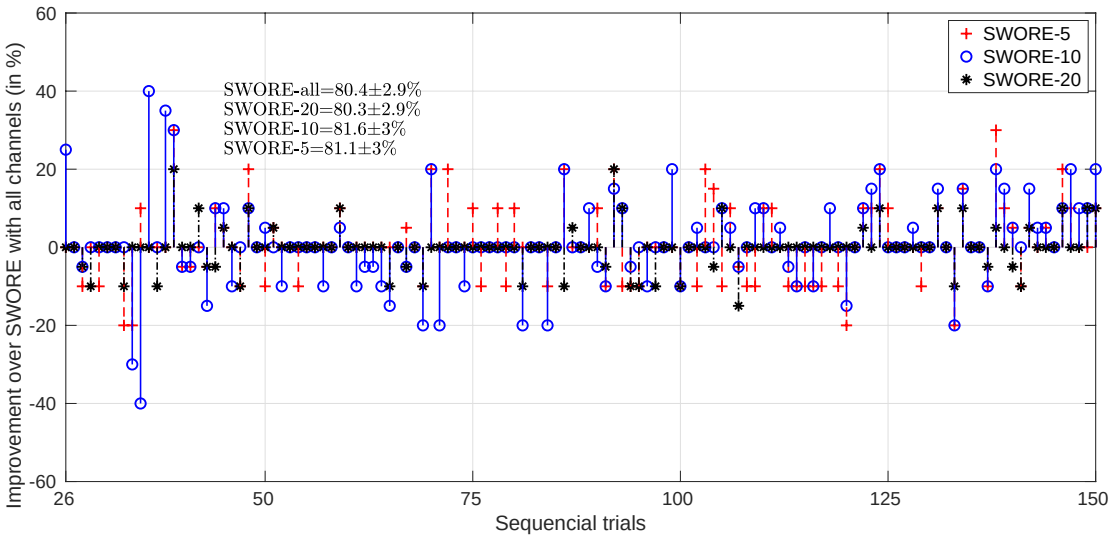

Figure 9: Performance Comparison with Reduced Number of Channels. SWORE-K denotes we retrain SWORE with only K randomly selected reliable channels following the estimated channel reliability shown in Fig. 8. SWORE-all denotes SWORE trained with all channels. We only collect the result of the first participant for a showcase.

(i.e., Online GBMM) in Sect. 4; C4 denotes the blank-out noise mode for data augmentation in Eq. 15.

To explore the efficacy of the listed four contributions, we introduce five new baselines, namely, LOR w/o C3: LOR without online calibration; SWORE w/o C3: SWORE without online calibration; SWORE w/o C2&C4: SWORE without modeling steady-state and using data augmentation; SWORE w/o C2: SWORE without modeling steady-state; SWORE w/o C4: SWORE without using data augmentation. We ran all methods for 100 times on the first participant and calculated the mean and standard deviation in Table 3.

Table 3: Comparison of average prediction accuracy (with 95% confidence interval) for different SWORE variants over 100 independent running.

Method	Calibration Option	Average Prediction Accuracy
LOR	w/o C3	$63.1 \pm 0.35\%$
	Full	$72.6 \pm 0.33\%$
SWORE	w/o C3	$64.5 \pm 0.33\%$
	w/o C2	$75.7 \pm 0.30\%$
	w/o C4	$74.6 \pm 0.32\%$
	w/o C2 & C4	$74.1 \pm 0.32\%$
	Full	$\mathbf{76.0 \pm 0.30}\%$

From Table 3, we can find that: (1) Online calibration (C3) is necessary for reliable mental fatigue evaluation. SWORE and LOR achieve significant improvement (around 10%) over their offline versions via adopting efficient online calibration, respectively.

(2) Comparing SWORE w/o C2&C4 to LOR, the robust multi-channel aggregation strategy (C1) enables higher predication accuracy by automatically eliminating the task-nonrelevant channels. (3) The data augmentation (C4) and the gradient flattening (C2) are all useful for further improving the model performance, comparing SWORE w/o C2 and SWORE w/o C4 to SWORE w/o C2&C4. (4) The importance ranking of all four contributions on the first participant is C3>C1>C4>C2.

6.5 Offline Analysis of Parameter Sensitivity and Model Uncertainty

In this section, we explored the parameter sensitivity of SWORE w.r.t. the hyperparameters (μ, Σ) and (α, β), respectively. In particular, we generated the offline brain dynamic preferences as follows: (1) the trials of each participant were randomly divided into two parts: 50% for training and 50% for test; (2) Offline Brain dynamic preferences were constructed according to the pairwise comparisons between the RTs regardless of their sequential property.

Sensitivity analysis w.r.t. hyperparameters (μ, Σ) For the sake of simplicity, we considered diagonal covariance matrix here. Specifically, we randomly initialized μ in $[-10^{-a}, 10^{-a}]$ and Σ in $[\mathbf{0}, 10^{-b}\mathbf{I}]$. The value of a and b are set within $\{0, 2, 4\}$, respectively. Further, we adopted a non-informative prior for π_n, namely $\alpha_n = \beta_n = 5$ to eliminate the effects of noisy channels. The data augmentation size T is set to 1 since the training data is sufficient. The testing performance of SWORE under different parameter setting are presented in Table 4.

Table. 4 shows that: (1) SWORE consistently performs very well with testing accuracy greater than 70% on all participants under small initialization $(2 < a, b < 4)$ for (μ, Σ). Because SWORE model suffers from spurious overflow/underflow problems with large initialization at each updating step, due to the high dimension feature $(L = 492)$ and the exponential operator (within the sigmoid function). (2) Although the performance of SWORE has minor differences for different participants, SWORE is robust to the small initialization and shows comparable performance for the same participant under different initialization.

Sensitivity analysis w.r.t. hyperparameters (α, β) To explore the effects of hyperparameter (α_n, β_n) w.r.t. SWORE model, we randomly initialized α_n, β_n in $\{1, 3, 5\}$, respectively. We randomly initialized μ in $[-10^{-2}, 10^{-2}]$ and Σ in $[\mathbf{0}, 10^{-4} \times \mathbf{I}]$. The corrupting size T was set to 1 as previous. The performance of SWORE on the testing data are reported in Table. 4.

It is worth to note that SWORE is insensitive to the initialization of hyperparameters (α, β). Particularly, SWORE achieves comparable performance for each participant under different initialization of (α, β). SWORE consistently performs very well on all forty participants, regardless of different initialization for (α, β).

Sensitivity analysis w.r.t. data augmentation size T According to Table. 4, we randomly initialized μ in $[-10^{-2}, 10^{-2}]$, Σ in $[\mathbf{0}, 10^{-4} \times \mathbf{I}]$, and initialized hyperparameters

Table 4: Test accuracy (in %, the larger the better) w.r.t. hyperparameter (μ, Σ) and (α, β) with dropout rate $\theta = 0.5$, data augmentation number $T = 1$. The best parameter settings are marked in gray. Some parameter settings do not consistently perform very well and may fail on some participants (marked in bold).

Test	(a,b) with $\mu = 10^{-a}$ and $\Sigma = 10^{-b}\mathbf{I}$. (α,β) fixed to (5,5)									(α,β) with (μ,Σ) fixed to $(10^{-2}, 10^{-4})$								
ACC	(0,0)	(2,0)	(4,0)	(0,2)	(2,2)	(4,2)	(0,4)	(2,4)	(4,4)	(1,1)	(1,3)	(1,5)	(3,1)	(3,3)	(3,5)	(5,1)	(5,3)	(5,5)
P1	50.00	50.00	79.89	77.23	78.29	78.75	50.00	78.71	79.28	78.48	78.26	78.29	78.71	78.52	78.37	78.64	78.83	78.71
P2	50.00	50.00	50.00	79.55	81.48	83.00	50.00	82.03	81.85	81.62	82.13	82.09	82.03	81.75	82.17	82.04	81.98	82.03
P3	50.00	50.00	50.00	85.32	84.54	84.29	70.07	82.89	83.14	83.47	83.47	83.47	82.89	83.02	83.55	83.55	83.72	82.89
P4	50.00	50.00	50.00	72.16	76.45	75.72	50.00	73.26	73.26	73.34	73.65	73.68	73.26	73.42	73.65	73.38	73.46	73.26
P5	50.00	50.00	50.00	85.34	85.52	85.21	50.00	85.07	85.00	85.17	85.14	85.13	85.07	85.06	85.17	85.06	85.07	85.07
P6	50.00	50.00	50.00	86.76	83.25	84.76	50.00	84.56	84.31	84.23	84.31	84.31	84.56	84.64	84.40	85.13	85.05	84.56
P7	50.00	50.00	50.00	75.44	75.21	75.10	50.00	75.18	75.31	75.19	75.12	75.19	75.18	75.22	75.10	75.19	75.23	75.18
P8	50.00	50.00	50.00	84.38	84.06	84.10	50.00	84.70	84.62	84.30	84.54	84.5	84.70	84.66	84.46	84.61	84.53	84.70
P9	50.00	50.00	50.00	83.12	83.01	83.12	50.00	83.26	83.46	83.25	82.86	82.87	83.26	83.22	83.00	82.85	83.14	83.26
P10	50.00	50.00	50.00	**79.17**	88.00	88.51	50.00	89.50	89.47	89.14	88.56	88.56	89.50	89.21	88.44	88.43	88.41	89.50
P11	50.00	50.00	50.00	80.45	75.99	76.39	50.00	77.49	77.42	77.06	76.14	76.14	77.49	77.29	76.26	77.11	77.09	77.49
P12	50.00	50.00	50.00	79.97	80.18	80.28	50.00	80.09	80.09	80.13	80.09	80.13	80.09	80.13	80.09	79.94	79.94	80.09
P13	50.00	50.00	50.00	81.09	80.75	80.75	50.00	81.52	81.52	81.33	81.23	81.23	81.52	81.33	81.14	81.04	81.18	81.52
P14	50.00	50.00	50.00	**50.00**	78.58	78.65	50.00	80.03	80.07	79.70	80.00	80.03	80.03	79.70	80.03	79.46	79.49	80.03
P15	50.00	50.00	50.00	89.58	89.92	89.86	50.00	89.93	89.97	89.95	89.95	89.95	89.93	89.90	89.95	90.04	89.97	89.93
P16	50.00	50.00	50.00	73.02	72.75	72.59	50.00	72.41	72.17	72.44	72.37	72.37	72.41	72.42	72.33	72.50	72.42	72.41
P17	50.00	50.00	50.00	**50.00**	76.99	77.63	50.00	78.05	78.09	78.09	77.94	77.79	78.05	78.24	77.45	77.86	77.75	78.05
P18	50.00	50.00	50.00	**50.00**	**78.03**	85.38	50.00	89.36	93.52	88.38	87.82	87.79	89.36	88.88	88.00	87.31	87.92	89.36
P19	50.00	50.00	50.00	77.97	77.94	77.80	50.00	77.92	77.89	77.69	77.62	77.61	77.92	77.92	77.83	77.68	77.89	77.92
P20	50.00	50.00	50.00	80.78	79.96	79.80	50.00	80.32	80.48	80.29	80.29	80.28	80.32	80.29	80.30	80.33	80.34	80.32
P21	50.00	50.00	50.00	**50.00**	**69.92**	73.51	50.00	75.74	78.23	79.13	78.89	78.90	75.74	74.38	79.05	79.29	79.37	75.74
P22	50.00	50.00	50.00	78.42	77.96	78.08	50.00	78.05	78.22	77.96	77.95	77.96	78.05	77.99	77.98	77.95	77.96	78.05
P23	50.00	50.00	50.00	84.72	84.34	84.47	50.00	84.66	84.58	84.55	84.55	84.55	84.66	84.64	84.55	84.80	84.69	84.66
P24	50.00	50.00	50.00	80.08	80.12	80.11	50.00	80.04	80.09	79.98	79.99	79.99	80.04	80.01	79.98	80.12	80.09	80.04
P25	50.00	50.00	50.00	82.12	82.18	82.26	50.00	82.17	82.39	82.18	82.20	82.21	82.17	82.31	81.99	82.18	81.99	82.17
P26	50.00	50.00	50.00	86.71	86.52	86.55	50.00	86.61	86.63	86.61	86.61	86.61	86.61	86.60	86.61	86.59	86.60	86.61
P27	50.00	50.00	50.00	81.17	**77.35**	82.60	50.00	82.71	83.20	82.79	82.94	82.93	82.71	82.77	82.86	82.97	83.00	82.71
P28	50.00	50.00	50.00	85.36	**64.69**	85.40	50.00	85.34	83.73	85.37	85.34	85.29	85.34	85.31	85.33	85.23	85.32	85.34
P29	50.00	50.00	50.00	84.12	84.06	84.13	50.00	83.87	83.86	83.83	83.85	83.85	83.87	83.85	83.86	83.85	83.84	83.87
P30	50.00	50.00	50.00	**50.00**	82.30	84.32	50.00	84.27	84.40	84.07	84.08	84.14	84.27	84.19	84.03	84.08	84.08	84.27
P31	50.00	50.00	50.00	82.28	83.80	83.33	50.00	83.55	83.60	83.55	83.53	83.52	83.55	83.57	83.52	83.51	83.53	83.55
P32	50.00	50.00	50.00	84.54	86.02	85.19	50.00	85.69	86.69	85.66	86.62	86.56	85.69	85.65	86.54	86.46	86.51	85.69
P33	50.00	65.27	50.00	80.05	80.59	80.62	50.00	80.90	81.34	80.83	81.00	80.98	80.90	80.83	80.79	81.26	81.18	80.90
P34	50.00	50.00	69.92	87.27	86.98	87.37	50.00	87.65	87.65	87.47	87.65	87.65	87.65	87.59	87.62	87.40	87.47	87.65
P35	50.00	50.00	50.00	74.24	75.32	74.28	50.00	74.77	74.95	74.81	74.77	74.74	74.77	74.76	74.69	74.90	74.82	74.77
P36	50.00	50.00	50.00	86.17	85.58	85.42	50.00	85.55	85.58	85.50	85.55	85.55	85.55	85.52	85.50	85.47	85.52	85.55
P37	50.00	50.00	50.00	90.96	89.81	90.25	50.00	90.20	90.64	89.81	89.43	89.43	90.20	90.03	89.49	89.98	89.92	90.20
P38	50.00	50.00	50.00	90.30	90.06	90.14	50.00	90.52	90.40	90.48	90.28	90.28	90.52	90.52	90.28	90.44	90.48	90.52
P39	50.00	50.00	50.00	85.09	84.65	84.65	50.00	84.90	84.98	84.94	84.90	84.90	84.90	84.98	84.98	84.68	84.79	84.90
P40	50.00	50.00	50.00	75.80	75.90	75.86	50.00	75.93	75.96	75.93	75.93	75.92	75.93	75.92	75.91	75.86	75.89	75.93

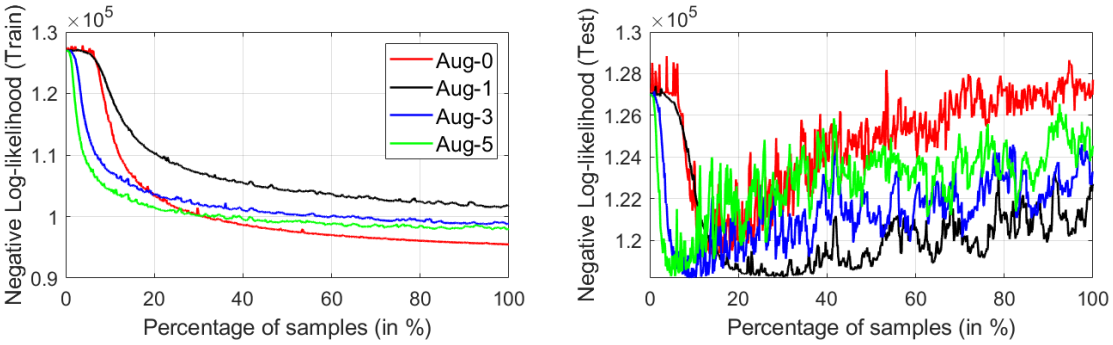

Figure 10: The negative Log-likelihood of brain dynamic preferences on training and test dataset w.r.t. different level of data augmentation size. Note that "Aug-N" denotes the data augmentation size T is set to N.

(α, β) to $(5,5)$. Then, we collected the negative Log-likelihood of brain dynamic preferences on training and test dataset (See Fig. 10) with data augmentation size T being set to $\{0,1,3,5\}$, respectively. Note that we only showed the results of the first participant due to space concern.

From Fig. 10, we can observe that: (1) the SWORE model is prone to be overfitting on the original EEG signal, since the dimensions in the EEG signals (either in time domain or frequency domain) are closely related to each other. See Section 5.1 for more details explanation. (2) The feature corruption trick $(T = 1)$ achieves the best performance comparing to other setting, including the data augmentation methods $(T > 1)$. The larger the data augmentation size T, the worse the generalization performance of SWORE is. (3) It is interesting to note that SWORE with data augmentation methods $(T > 1)$ performs extremely good with only a few samples (less than 20% training data), but SWORE starts to overfitting when updated with more samples.

Here, we empirically analyzed the stability of Online GBMM algorithm. According to our sensitivity analysis w.r.t. hyperparameters (μ, Σ) (Table. 4) and (α, β) (Table. 4), we randomly initialized μ in $[-10^{-2}, 10^{-2}]$, Σ in $[\mathbf{0}, 10^{-4} \times \mathbf{I}]$. Further, we initialized hyperparameters (α, β) to $(5,5)$. The corrupting size T is set to 1. Then, we repeated the Online GBMM algorithm on the training data for 20 times and summarized the prediction accuracy on the test data Fig. 11.

Stability analysis of Online GBMM algorithm It can be observed from Fig. 11 that: (1) the test accuracies of each participant are quite stable in different runnings. (2) SWORE consistently achieves high generalization performance (test accuracy above 80%) on 26 over forty participants with 95% confidence. Note that the performance of each specific participant can be further improved by tailor-designed brain dynamic preferences for each participant.

6.6 Online Mental Fatigue Evaluation on Forty Participants

Following the online experiment setting in Sect.6.3, we explored the reliability of SWORE on forty participants in the online monitoring scenario. Similarly, we leveraged the

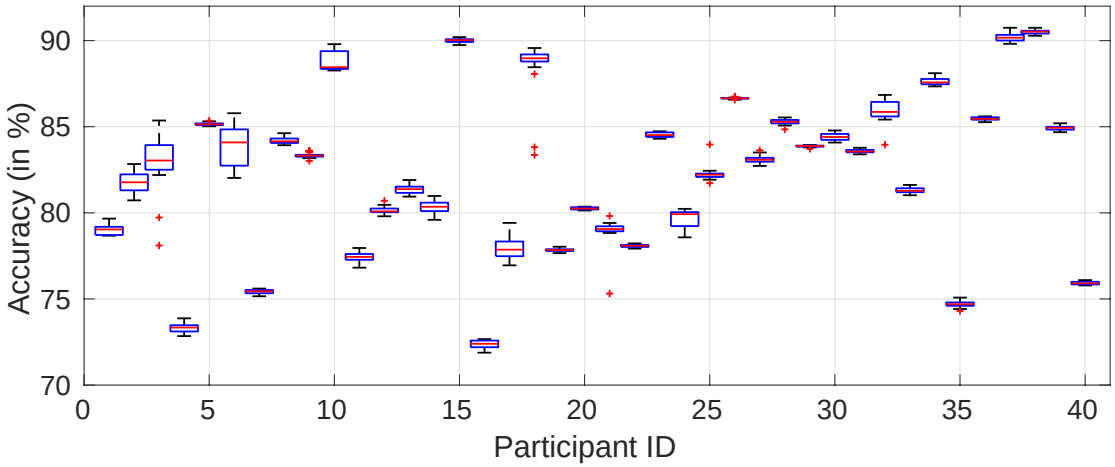

Figure 11: Box plot of the prediction accuracy on the test dataset. The symbol "+" denotes the outliers.

Table 5: Comparison of average prediction accuracy (in %) with 95% confidence interval: for all baselines on forty participants over 100 independent running. Best results are marked in gray.

ACC	P1	P2	P3	P4	P5	P6	P7	P8
SVR	69.1±0.36	76.9±0.30	74.0±0.31	70.4±0.35	72.7±0.34	70.6±0.39	51.5±0.47	76.6±0.31
LOR	72.6±0.33	78.3±0.31	73.1±0.37	73.8±0.33	73.7±0.36	74.5±0.32	75.1±0.33	74.8±0.32
SWORE	76.0±0.30	79.9±0.29	76.5±0.31	75.9±0.31	76.7±0.31	74.6±0.33	78.3±0.32	76.9±0.30
ACC	P9	P10	P11	P12	P13	P14	P15	P16
SVR	69.7±0.36	71.0±0.37	73.7±0.37	74.4±0.33	73.7±0.32	72.0±0.35	74.2±0.36	45.8±0.44
LOR	73.2±0.35	73.3±0.36	76.5±0.32	78.3±0.30	74.3±0.34	72.6±0.35	72.5±0.35	76.3±0.32
SWORE	75.7±0.33	75.2±0.32	78.2±0.30	78.4±0.31	74.1±0.36	74.0±0.32	74.0±0.36	79.9±0.31
ACC	P17	P18	P19	P20	P21	P22	P23	P24
SVR	63.3±0.41	65.6±0.43	67.0±0.38	68.8±0.36	47.4±0.45	63.7±0.40	50.9±0.38	75.3±0.35
LOR	79.8±0.29	74.4±0.34	77.6±0.32	73.7±0.36	76.6±0.34	73.1±0.34	84.2±0.27	84.6±0.24
SWORE	80.0±0.29	76.7±0.31	79.0±0.31	76.7±0.35	76.4±0.32	75.2±0.35	84.0±0.29	84.2±0.25
ACC	P25	P26	P27	P28	P29	P30	P31	P32
SVR	71.7±0.34	72.8±0.32	75.2±0.33	69.3±0.39	72.6±0.36	79.0±0.32	70.1±0.39	39.1±0.41
LOR	71.1±0.37	76.8±0.31	72.3±0.35	76.3±0.33	79.7±0.30	78.9±0.30	77.1±0.33	80.8±0.31
SWORE	75.2±0.32	77.8±0.29	74.0±0.32	76.9±0.31	80.0±0.29	79.5±0.28	77.5±0.30	81.1±0.28
ACC	P33	P34	P35	P36	P37	P38	P39	P40
SVR	72.6±0.35	59.3±0.41	70.8±0.35	75.5±0.30	78.9±0.29	74.0±0.32	74.1±0.35	56.1±0.43
LOR	74.5±0.33	80.2±0.28	77.1±0.31	74.9±0.31	79.1±0.28	79.4±0.29	79.5±0.29	76.5±0.33
SWORE	77.7±0.31	80.9±0.29	77.3±0.31	76.3±0.31	79.8±0.28	81.0±0.28	80.6±0.28	78.3±0.32

pre-recorded 25 trials to pre-train the embryonic SWORE, LOR, and SVR model, respectively. Then, we ran SWORE and other baselines independently for 100 times and calculated the average prediction accuracy for all forty participants in Table 5.

It can be observed from Table 5 that: (1) SWORE can give the most reliable evaluation with the lowest variance for the new EEG signal, comparing to LOR and SVR. In particular, SWORE achieves the highest average prediction accuracy on $34/40$ participants and comparable results on the rest participants. (2) SWORE achieves consistent reliable evaluations on different participants. There are 35 participants for SWORE whose average prediction accuracy is above 75%, while there are 7 participants for SVR and 22 participants for LOR, respectively. (3) The non-online method, i.e., SVR, is not trustworthy, since it does not consider the non-stationary properties of brain dynamics. There are 7 participants, i.e., P7, P16, P21, P23, P32, P34, and P40, on which SVR achieves an average prediction accuracy below 60%. On the contrary, LOR and SWORE, equipped with efficient online calibration strategies, can consistently achieve an average prediction accuracy above 75% on the same participants.

7 Concluding Remarks

This paper takes an initial step to calibrate prediction models on non-stationary brain dynamics. We proposed a Self-Weight Ordinal REgression (SWORE) model with Brain Dynamics table (BDtable) for online mental fatigue monitoring. SWORE can aggregate the information from multiple noisy channels based on the brain dynamic preferences; while BDtable is used to online calibrate the SWORE model utilizing a generalized Bayesian moment matching algorithm. Empirical results demonstrate that the proposed framework achieves significantly better performance than baseline approaches like SVR and LOR. As a direction for future research, we are committed to assess the feasibility of performing the online mental-fatigue monitoring system with EEG signals and other mental fatigue indicators.

A Proof for Therorem 1

Assume $f(w)$ is the marginalized likelihood of preference y and almost twice differentiable. Upon updating this preference, the posterior parameters $(\mu^{new}, \Sigma^{new})$ of weight w can be estimated as:

$$\mu^{new} \approx \mu + \Sigma \times \frac{d \log f(w)}{dw}\Big|_{w=\mu}, \tag{19a}$$

$$\Sigma^{new} \approx \Sigma + \Sigma \times \frac{d^2 \log f(w)}{dw dw^T}\Big|_{w=\mu} \times \Sigma. \tag{19b}$$

Proof. Before introducing our proof, we first introduce the lemma 1 as a building block.

Lemma 1. *((Weng and Lin, 2011)) Let z be a random vector, where each entry is independent and $z_i \sim \mathcal{N}(0,1)$, $i = 1, 2, \cdots, L$. Suppose that $f(z)$ is the likelihood function*

and almost twice differentiable. Then the first and second order moments of the posterior distribution can be estimated as

$$E[z] = E[\frac{\nabla f(z)}{f(z)}], \tag{20a}$$

$$E[z_i z_j] = I_{ij} + E[\frac{\nabla^2 f(z)}{f(z)}]_{ij}, \quad i,j = 1,\cdots,L \tag{20b}$$

where $I_{ij} = 1$ if $i = j$ and 0 otherwise, and $\left[.\right]_{ij}$ indicates the (i,j) component of a matrix.

Based on the Lemma 1, we give the detailed proof in the following. In terms of the posterior parameter μ^{new}, we have

$$\begin{aligned}
\mu^{new} = E_w[w] = \mu + \sqrt{\Sigma} \times E_z[z] &\overset{①}{=} \mu + \sqrt{\Sigma} \times E_z\left[\frac{\nabla f(\mu + \sqrt{\Sigma} z)}{f(\mu + \sqrt{\Sigma} z)}\right] \\
&\overset{②}{\approx} \mu + \sqrt{\Sigma} \times \frac{d \log f(\mu + \sqrt{\Sigma} z)}{dz}\Big|_{z=\mathbf{0}} \overset{③}{=} \mu + \sqrt{\Sigma} \times \frac{dw}{dz} \times \frac{d \log f(\mu + \sqrt{\Sigma} z)}{dw}\Big|_{w=\mu} \\
&= \mu + \Sigma \times \frac{d \log f(w)}{dw}\Big|_{w=\mu}.
\end{aligned}$$

where ① follows the Eq.(20a) of Lemma 1. ② sets $z = 0$. Such a substitution is reasonable as we expect that the posterior density of z to be concentrated on 0. ③ follows the chain rule.

In terms of the posterior parameter Σ^{new}, we have

$$\begin{aligned}
\Sigma^{new} &= \mathrm{Var}(w) = \sqrt{\Sigma} \times (E_z[zz^T] - E_z[z]E_z^T[z]) \times \sqrt{\Sigma} \\
&\overset{①}{=} \sqrt{\Sigma} \times \Big(\mathbf{I} + E_z\big[\frac{\nabla^2 f(\mu + \sqrt{\Sigma} z)}{f(\mu + \sqrt{\Sigma} z)}\big] - E_z\big[\frac{d \log f(\mu + \sqrt{\Sigma} z)}{dz} \frac{d \log f(\mu + \sqrt{\Sigma} z)}{dz^T}\big]\Big) \times \sqrt{\Sigma} \\
&\overset{②}{\approx} \sqrt{\Sigma} \times \Big(\mathbf{I} + \frac{\nabla^2 f(\mu + \sqrt{\Sigma} z)}{f(\mu + \sqrt{\Sigma} z)}\Big|_{z=\mathbf{0}} - \frac{d \log f(\mu + \sqrt{\Sigma} z)}{dz}\Big|_{z=\mathbf{0}} \frac{d \log f(\mu + \sqrt{\Sigma} z)}{dz^T}\Big|_{z=\mathbf{0}}\Big) \times \sqrt{\Sigma} \\
&\overset{③}{=} \Sigma + \sqrt{\Sigma} \times \frac{d^2 \log f(\mu + \sqrt{\Sigma} z)}{dz dz^T}\Big|_{z=\mathbf{0}} \times \sqrt{\Sigma} \\
&\overset{④}{=} \Sigma + \sqrt{\Sigma} \times \frac{dw}{dz} \times \frac{d^2 \log f(\mu + \sqrt{\Sigma} z)}{dw dw^T}\Big|_{w=\mu} \times \frac{dw^T}{dz^T} \times \sqrt{\Sigma} \\
&= \Sigma + \Sigma \times \frac{d^2 \log f(w)}{dw dw^T}\Big|_{w=\mu} \times \Sigma.
\end{aligned}$$

where ① follows the Eq.(20b) of Lemma (1). ② sets $z = 0$. Such a substitution is reasonable as we expect that the posterior density of z to be concentrated on 0. ④

follows the chain rule. We give the proof for ③ as follows,

$$
\begin{aligned}
&\Big[\frac{d^2 \log f(\mu+\sqrt{\Sigma}z)}{dz dz^T}\Big]_{ij} = \frac{\partial}{\partial z_j}\Big(\frac{\partial f(\mu+\sqrt{\Sigma}z)/\partial z_i}{f(\mu+\sqrt{\Sigma}z)}\Big) \\
&= \frac{\frac{\partial^2 f(\mu+\sqrt{\Sigma}z)}{\partial z_i \partial z_j} f(\mu+\sqrt{\Sigma}z) - \frac{\partial f(\mu+\sqrt{\Sigma}z)}{\partial z_i} \times \frac{\partial f(\mu+\sqrt{\Sigma}z)}{\partial z_j}}{f^2(\mu+\sqrt{\Sigma}z)} \\
&= \Big[\frac{\nabla^2 f(\mu+\sqrt{\Sigma}z)}{f(\mu+\sqrt{\Sigma}z)}\Big]_{ij} - \frac{\partial \log f(\mu+\sqrt{\Sigma}z)}{\partial z_i} \times \frac{\partial \log f(\mu+\sqrt{\Sigma}z)}{\partial z_j} \\
&= \Big[\frac{\nabla^2 f(\mu+\sqrt{\Sigma}z)}{f(\mu+\sqrt{\Sigma}z)}\Big]_{ij} - \Big[\frac{d \log f(\mu+\sqrt{\Sigma}z)}{dz}\Big]_i \times \Big[\frac{d \log f(\mu+\sqrt{\Sigma}z)}{dz^T}\Big]_j .
\end{aligned}
$$

□

B Second-order Taylor approximation for $E_{\mathcal{N}(w|\mu,\Sigma)}\big[\sigma\big(w^T \Delta x_n\big)\big]$

Suppose $R_1 = E_{\mathcal{N}(w|\mu,\Sigma)}\sigma(w^T \Delta x_n)$, the second-order Taylor approximation of R_1 at μ can be represented as follows:

$$
R_1 = E_{\mathcal{N}(w|\mu,\Sigma)}\sigma(w^T \Delta x_n) = \sigma(\mu^T \Delta x_n)\Big[1 + \frac{1}{2}[1-\sigma(\mu^T \Delta x_n)][1-2\sigma(\mu^T \Delta x_n)]\Delta x_n^T \Sigma \Delta x_n\Big].
$$

Then we set $R_1 = \max(R_1, \kappa_2)$, where κ_2 is a small positive value to ensure a positive R_1.

C Posterior Moments of Beta distribution

$$
\begin{aligned}
E[\pi_n] &= \int \pi_n P(\pi_n|y,\Delta x_n) d\pi_n = \int \pi_n \frac{P(y|\pi_n,\Delta x_n)\mathrm{Beta}(\pi_n|\alpha_n,\beta_n)}{R} d\pi_n \\
&= \frac{1}{R}\int \pi_n[\pi_n R_1 + (1-\pi_n)R_2]\mathrm{Beta}(\pi_n|\alpha_n,\beta_n) d\pi_n \\
&= \frac{R_1 - R_2}{R}\int \pi_n^2 \mathrm{Beta}(\pi_n|\alpha_n,\beta_n) d\pi_n + \frac{R_2}{R}\int \pi_n \mathrm{Beta}(\pi_n|\alpha_n,\beta_n) d\pi_n \\
&= \frac{R_1 - R_2}{R}\frac{(\alpha_n+1)\alpha_n}{(\alpha_n+\beta_n+1)(\alpha_n+\beta_n)} + \frac{R_2}{R}\frac{\alpha_n}{\alpha_n+\beta_n} \\
&= \frac{R_1(\alpha_n+1)\alpha_n + R_2\alpha_n\beta_n}{R(\alpha_n+\beta_n+1)(\alpha_n+\beta_n)}.
\end{aligned}
$$

$$
\begin{aligned}
E[\pi_n^2] &= \int \pi_n^2 P(\pi_n|y,\Delta x_n)d\pi_n = \int \pi_n^2 \frac{P(y|\pi_n,\Delta x_n)\text{Beta}(\pi_n|\alpha_n,\beta_n)}{R}d\pi_n \\
&= \frac{1}{R}\int \pi_n^2[\pi_n R_1 + (1-\pi_n)R_2]\text{Beta}(\pi_n|\alpha_n,\beta_n)d\pi_n \\
&= \frac{R_1-R_2}{R}\int \pi_n^3 \text{Beta}(\pi_n|\alpha_n,\beta_n)d\pi_n + \frac{R_2}{R}\int \pi_n^2 \text{Beta}(\pi_n|\alpha_n,\beta_n)d\pi_n \\
&= \frac{R_1-R_2}{R}\frac{(\alpha_n+2)(\alpha_n+1)\alpha_n}{(\alpha_n+\beta_n+2)(\alpha_n+\beta_n+1)(\alpha_n+\beta_n)} + \frac{R_2}{R}\frac{(\alpha_n+1)\alpha_n}{(\alpha_n+\beta_n+1)(\alpha_n+\beta_n)} \\
&= \frac{\alpha_n(\alpha_n+1)[R_1(\alpha_n+2)+R_2\beta_n]}{R(\alpha_n+\beta_n+2)(\alpha_n+\beta_n+1)(\alpha_n+\beta_n)}.
\end{aligned}
$$

D The updating rules for hyperparameter $(\alpha_n^{new}, \beta_n^{new})$

In terms of the sufficient moments with regard to the posterior distribution $q(\pi_n|\alpha_n^{new},\beta_n^{new})$, we have

$$
\begin{aligned}
E[\pi_n] &= \int \pi_n q(\pi_n|\alpha_n^{new},\beta_n^{new})d\pi_n = \frac{\alpha_n^{new}}{\alpha_n^{new}+\beta_n^{new}}, \\
E[\pi_n^2] &= \int \pi_n^2 q(\pi_n|\alpha_n^{new},\beta_n^{new})d\pi_n = \frac{\alpha_n^{new}(\alpha_n^{new}+1)}{(\alpha_n^{new}+\beta_n^{new}+1)(\alpha_n^{new}+\beta_n^{new})}.
\end{aligned}
$$

According to the above equations, we have

$$
\begin{aligned}
E[\pi_n] - E[\pi_n^2] &= \frac{\alpha_n^{new}\beta_n^{new}}{(\alpha_n^{new}+\beta_n^{new}+1)(\alpha_n^{new}+\beta_n^{new})}, \\
E[\pi_n^2] - (E[\pi_n])^2 &= \frac{\alpha_n^{new}\beta_n^{new}}{(\alpha_n^{new}+\beta_n^{new}+1)(\alpha_n^{new}+\beta_n^{new})^2}.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\alpha_n^{new} &= \frac{(E[\pi_n]-E[\pi_n^2])E[\pi_n]}{E[\pi_n^2]-(E[\pi_n])^2}, \\
\beta_n^{new} &= \frac{(E[\pi_n]-E[\pi_n^2])(1-E[\pi_n])}{E[\pi_n^2]-(E[\pi_n])^2}.
\end{aligned}
$$

References

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. (2012). Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75.

Bose, R., Wang, H., Dragomir, A., Thakor, N., Bezerianos, A., and Li, J. (2019). Regression based continuous driving fatigue estimation: Towards practical implementation. *IEEE Transactions on Cognitive and Developmental Systems*.

Chai, R., Naik, G. R., Nguyen, T. N., Ling, S. H., Tran, Y., Craig, A., and Nguyen, H. T. (2016). Driver fatigue classification with independent component by entropy rate bound minimization analysis in an eeg-based system. *IEEE journal of biomedical and health informatics*, 21(3):715–724.

Colosio, M., Shestakova, A., Nikulin, V. V., Blagovechtchenski, E., and Klucharev, V. (2017). Neural mechanisms of cognitive dissonance (revised): An eeg study. *Journal of Neuroscience*, pages 3209–16.

Congedo, M., Barachant, A., and Bhatia, R. (2017). Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174.

Cui, Y. and Wu, D. (2017). Eeg-based driver drowsiness estimation using convolutional neural networks. In *International Conference on Neural Information Processing*, pages 822–832. Springer.

Cui, Y., Xu, Y., and Wu, D. (2019). Eeg-based driver drowsiness estimation using feature weighted episodic training. *IEEE transactions on neural systems and rehabilitation engineering*, 27(11):2263–2273.

Fallahi, M., Motamedzade, M., Heidarimoghadam, R., Soltanian, A. R., and Miyake, S. (2016). Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study. *Applied ergonomics*, 52:95–103.

Gharagozlou, F., Saraji, G. N., Mazloumi, A., Nahvi, A., Nasrabadi, A. M., Foroushani, A. R., Kheradmand, A. A., Ashouri, M., and Samavati, M. (2015). Detecting driver mental fatigue based on eeg alpha power changes during simulated driving. *Iranian journal of public health*, 44(12):1693.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Graimann, B., Allison, B., and Pfurtscheller, G. (2009). Brain–computer interfaces: A gentle introduction. In *Brain-computer interfaces*, pages 1–27. Springer.

Gurudath, N. and Riley, H. B. (2014). Drowsy driving detection by eeg analysis using wavelet transform and k-means clustering. *Procedia Computer Science*, 34:400–409.

Homan, R. W., Herman, J., and Purdy, P. (1987). Cerebral location of international 10–20 system electrode placement. *Electroencephalography and clinical neurophysiology*, 66(4):376–382.

Huang, C.-S., Pal, N. R., Chuang, C.-H., and Lin, C.-T. (2015). Identifying changes in eeg information transfer during drowsy driving by transfer entropy. *Frontiers in human neuroscience*, 9:570.

Huang, R.-S., Jung, T.-P., and Makeig, S. (2009). Tonic changes in eeg power spectra during simulated driving. In *International Conference on Foundations of Augmented Cognition*, pages 394–403, Berlin, Heidelberg. Springer.

Jagannath, M. and Balasubramanian, V. (2014). Assessment of early onset of driver fatigue using multimodal fatigue measures in a static simulator. *Applied ergonomics*, 45(4):1140–1147.

Jaini, P., Chen, Z., Carbajal, P., Law, E., Middleton, L., Regan, K., Schaekermann, M., Trimponias, G., Tung, J., and Poupart, P. (2017). Online bayesian transfer learning for sequential data modeling. In *International Conference on Learning Representations*.

Jap, B. T., Lal, S., Fischer, P., and Bekiaris, E. (2009). Using eeg spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, 36(2):2352–2359.

Kar, S., Bhagat, M., and Routray, A. (2010). Eeg signal analysis for the assessment and quantification of driver's fatigue. *Transportation research part F: traffic psychology and behaviour*, 13(5):297–306.

Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Curio, G., Hagemann, K., Bruns, A., Schrauf, M., Kincses, W., et al. (2007). Improving human performance in a real operating environment through real-time mental workload detection. *Toward Brain-Computer Interfacing*, 409422.

Lal, S. K., Craig, A., Boord, P., Kirkup, L., and Nguyen, H. (2003). Development of an algorithm for an eeg-based driver fatigue countermeasure. *Journal of safety Research*, 34(3):321–328.

Laurent, F., Valderrama, M., Besserve, M., Guillard, M., Lachaux, J.-P., Martinerie, J., and Florence, G. (2013). Multimodal information improves the rapid detection of mental fatigue. *Biomedical Signal Processing and Control*, 8(4):400–408.

Li, G., Li, B., Wang, G., Zhang, J., and Wang, J. (2017). A new method for human mental fatigue detection with several eeg channels. *Journal of Medical and Biological Engineering*, 37(2):240–247.

Lin, C.-T., Chang, C.-J., Lin, B.-S., Hung, S.-H., Chao, C.-F., and Wang, I.-J. (2010). A real-time wireless brain–computer interface system for drowsiness detection. *IEEE transactions on biomedical circuits and systems*, 4(4):214–222.

Lin, C.-T., Tsai, S.-F., and Ko, L.-W. (2013). Eeg-based learning system for online motion sickness level estimation in a dynamic vehicle environment. *IEEE transactions on neural networks and learning systems*, 24(10):1689–1700.

Liu, Y.-T., Lin, Y.-Y., Wu, S.-L., Chuang, C.-H., and Lin, C.-T. (2016). Brain dynamics in predicting driving fatigue using a recurrent self-evolving fuzzy neural network. *IEEE transactions on neural networks and learning systems*, 27(2):347–360.

Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., and Blankertz, B. (2008). Machine learning for real-time single-trial eeg-analysis: from brain–computer interfacing to mental state monitoring. *Journal of neuroscience methods*, 167(1):82–90.

Nguyen, T., Ahn, S., Jang, H., Jun, S. C., and Kim, J. G. (2017). Utilization of a combined eeg/nirs system to predict driver drowsiness. *Scientific reports*, 7:43933.

Palanivel Rajan, S. and Dinesh, T. (2015). Systematic review on wearable driver vigilance system with future research directions. *International Journal of Applied Engineering Research*, 10(1):627–32.

Pan, Y., Tsang, I. W., Singh, A. K., Lin, C.-T., and Sugiyama, M. (2020). Stochastic multichannel ranking with brain dynamics preferences. *Neural Computation*, 32(8):1499–1530.

Ratcliff, R., Philiastides, M. G., and Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the eeg. *Proceedings of the National Academy of Sciences*, 106(16):6539–6544.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322.

Resalat, S. N. and Saba, V. (2015). A practical method for driver sleepiness detection by processing the eeg signals stimulated with external flickering light. *Signal, Image and Video Processing*, 9(8):1751–1757.

Richer, R., Zhao, N., Amores, J., Eskofier, B. M., and Paradiso, J. A. (2018). Real-time mental state recognition using a wearable eeg. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5495–5498. IEEE.

Sahayadhas, A., Sundaraj, K., and Murugappan, M. (2012). Detecting driver drowsiness based on sensors: a review. *Sensors*, 12(12):16937–16953.

Sauvet, F., Bougard, C., Coroenne, M., Lely, L., Van Beers, P., Elbaz, M., Guillard, M., Leger, D., and Chennaoui, M. (2014). In-flight automatic detection of vigilance states using a single eeg channel. *IEEE Transactions on Biomedical Engineering*, 61(12):2840–2847.

Schlkopf, B., Smola, A. J., and Bach, F. (2018). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. the MIT press.

Soon, C. S., Brass, M., Heinze, H.-J., and Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature neuroscience*, 11(5):543.

Teplan, M. et al. (2002). Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11.

Van Cutsem, J., Marcora, S., De Pauw, K., Bailey, S., Meeusen, R., and Roelands, B. (2017). The effects of mental fatigue on physical performance: a systematic review. *Sports medicine*, 47(8):1569–1588.

Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57.

Wang, H., Dragomir, A., Abbasi, N. I., Li, J., Thakor, N. V., and Bezerianos, A. (2018). A novel real-time driving fatigue detection system based on wireless dry eeg. *Cognitive neurodynamics*, 12(4):365–376.

Wang, S., Zhang, Y., Wu, C., Darvas, F., and Chaovalitwongse, W. A. (2015). Online prediction of driver distraction based on brain activity patterns. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):136–150.

Wei, C.-S., Lin, Y.-P., Wang, Y.-T., Lin, C.-T., and Jung, T.-P. (2018). A subject-transfer framework for obviating inter-and intra-subject variability in eeg-based drowsiness detection. *NeuroImage*, 174:407–419.

Welch, P. (1967). The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73.

Weng, R. C. and Lin, C.-J. (2011). A bayesian approximation method for online ranking. *Journal of Machine Learning Research*, 12(Jan):267–300.

Woodroofe, M. (1989). Very weak expansions for sequentially designed experiments: linear models. *The annals of statistics*, 17(3):1087–1102.

Xu, J., Min, J., and Hu, J. (2018). Real-time eye tracking for the assessment of driver fatigue. *Healthcare technology letters*, 5(2):54–58.

Yan, L., Dodier, R. H., Mozer, M., and Wolniewicz, R. H. (2003). Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 848–855, Washington D.C. AAAI Press 2003.

Yarkoni, T., Barch, D. M., Gray, J. R., Conturo, T. E., and Braver, T. S. (2009). Bold correlates of trial-by-trial reaction time variability in gray and white matter: a multi-study fmri analysis. *PLoS One*, 4(1):e4257.

Zhou, K., Xue, G.-R., Zha, H., and Yu, Y. (2008). Learning to rank with ties. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–282, Singapore. ACM.