Stochastic Multi-Channel Ranking with Brain Dynamics Preferences

Yuangang Pan[1] **Ivor W. Tsang**[1] **Avinash K Singh**[1] **Chin-Teng Lin**[1] **Masashi Sugiyama**[2,3]

[1]Centre for Artificial Intelligence, University of Technology Sydney.

[2]Center for Advanced Intelligence Project, RIKEN.

[3]Graduate School of Frontier Sciences, The University of Tokyo.

Keywords: Mental Fatigue Monitoring, Brain Dynamics Preference, Stochastic Expectation-Maximum, electroencephalogram (EEG)

Abstract

Driver's cognitive state of mental fatigue significantly affects driving performance and more importantly public safety. Previous studies leverage the response time (RT) as the metric for mental fatigue and aim at estimating the exact value of RT using electroencephalogram (EEG) signals within a regression model. However, due to the easily corrupted and also non-smooth properties of RTs during data collection, methods focusing on predicting the exact value of a noisy measurement (RT) generally suffer from poor generalization performance. Considering that human reaction time (RT) is the reflection of brain dynamics preference (BDP) rather than a single regression output of EEG signals, a novel Channel-reliability Aware Ranking (CArank) model is proposed for multi-channel ranking problem. CArank learns from BDPs using EEG data robustly and aims at preserving the ordering corresponding to RTs. In particular, a transition matrix is introduced to characterize the reliability of each channel used in EEG data, which helps in learning with BDPs only from informative EEG channels. To handle large-scale EEG signals, a Stochastic-Generalized Expectation Maximum (SGEM) algorithm is proposed to update CArank in an online fashion. Comprehensive empirical analysis on EEG signals from 40 participants shows that our CArank achieves substantial improvements in reliability while simultaneously detecting noisy or less informative EEG channels.

1 Introduction

As reported by sleep health report (Adams et al., 2017), mental fatigue is a major cause in $33\% - 45\%$ of all road accidents. In general, mental fatigue (Boksem and Tops, 2008) refers to the inability to maintain optimal cognitive performance in continuous task of the high demand of cognitive activity. Such inability in the context of driver could lead to accidents with severe consequences (Adams et al., 2017). Individuals may find themselves in a mental fatigue state because of lack of sleep, continuous driving for long-time, midnight driving, monotonous driving, and driving during the influence of sleeping drugs or sleep disorders (Ji et al., 2004; Ting et al., 2008).

In response to these critical issues, several methods (Cook et al., 2007; Blankertz et al., 2009; Fazli et al., 2009; Wascher et al., 2014; Tian et al., 2018; Kaji et al., 2019) have been proposed to estimate and predict the mental fatigue based on electroencephalography (EEG) and reaction time (RT) (Fig. 1(a)). Some of these methods, however, performed considerably well for some participants but failed for others due to lack of generalization. There are several challenges behind such poor generalization and one of such problems is how to use RT effectively. RT is easily affected by the instrumental error, mind wandering or any other task non-related factors. A previous study (Wei et al., 2015) tried to overcome this problem by adopting different techniques to smooth RTs, but still failed to make it work for all participants. Note that humans' RT is usually the result of preference (Izuma and Adolphs, 2013) in brain dynamics during the task, rather than just a single value. Such preferences of humans can be affected by different cognition (Möckel et al., 2015) like mind-wandering (Lin et al., 2016), and/or a lower level of attention (Chuang et al., 2018). Therefore, the relationship between the EEG signals and RTs including the extreme/abnormal RTs should be taken care in the way that reflects human Brain Dynamics Preferences (BDPs) by the developed technique itself.

Another important problem lies in the heterogeneous channels extracted from different brain regions, which are normally responsible for different functionalities. There was an attempt to choose different brain regions (Wascher et al., 2014) for a method during mental fatigue evaluation but these regions of the brain are not necessarily the same for all participants (Gramann et al., 2006). For example, Wascher et al. (2014) heuristically used frontal theta to represent a different level of mental fatigue for all participants. In such a case the reliability of the learning model would inevitably degrade because of possibly noisy or less informative channels chosen, on different brain regions, by the method. Some previous work (de Naurois et al., 2017), attempted to solve this issue using artificial neural network models but still failed to provide convincing results. The aforementioned works impel us to pursue a purely data-driven approach to predict mental fatigue while getting rid of the low versatility caused by various heuristic tricks.

To overcome the above-mentioned problems, we first formulate the mental fatigue monitoring task into a multi-channel ranking problem and solve it with the proposed Channel-reliability Aware ranking (CArank) model. In particular, CArank could learn from BDPs using EEG data robustly, while effectively preserve the exact ordering of RTs (Fig. 1(b)). This approach surprisingly improves over defects of previous models and their performance caused by noisy and extreme RTs. Furthermore, our model also

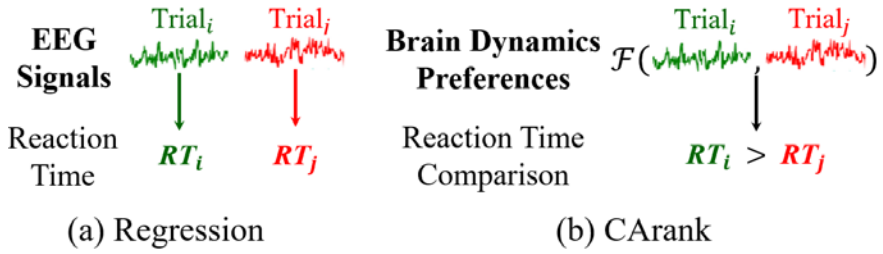

Figure 1: (a) Regression model with EEG signals. (b) CArank model with brain dynamics preferences.

proposes to use a transition matrix to evaluate the high confidence sources among heterogeneous EEG channels, which contributes highly toward task performance. In order to handle large-scale EEG signals and obtain higher generalization, a stochastic generalized expectation maximum (SGEM) algorithm is further proposed. More precisely, we make the following key contributions:

- We formulate the mental fatigue monitoring task into a multi-channel ranking problem and tackle it with the Channel-reliability Aware ranking (CArank) model. CArank is a purely data-driven approach to detect mental fatigue while evaluating channel reliability.
- We propose a stochastic generalized expectation-maximum algorithm for CArank, which extends CArank to large-scale applications.
- We conduct empirical experiments on EEG signals from 40 participants to demonstrate the superior reliability of our CArank in terms of mental fatigue monitoring.

The sequel of this paper is organized as follows. Section 2 introduces the background of mental fatigue monitoring and motivates the practice of using brain dynamics preferences. In Section 3, we proposes the multi-channel ranking problem and introduces our channel-reliability aware ranking to solve it. Section 4 describes a stochastic generalized expectation-maximization algorithm. Section 5 demonstrates the reliability of the proposed CArank with EEG signals from forty participants. Section 6 concludes the paper and envisions future work.

2 Background

In this section, we first introduce some preliminary knowledge about mental fatigue monitoring and then discuss our motivation for learning from brain dynamics preferences.

The reaction time is an intuitive indicator used to assess human mental fatigue. Therefore, a common practice for mental fatigue monitoring is to find a robust mapping for humans' reaction time (RT) to an emergent situation using the EEG signals recorded beforehand (Lal et al., 2003; Kohlmorgen et al., 2007; Jap et al., 2009).

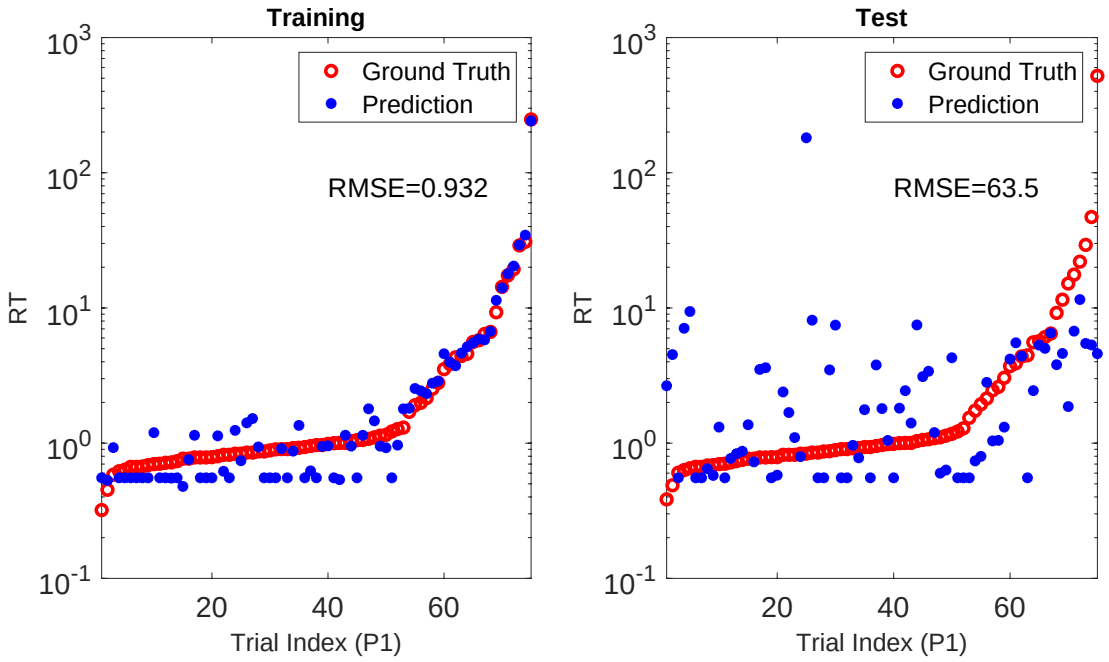

Figure 2: Overfitting of the two-layer deep regression model for mental fatigue monitoring. EEG signals from multiple channels are simply concatenated. The difference between the ground truth and the prediction is calculated with the root mean squared error (RMSE). We only collect the result from the first participant for a showcase.

2.1 Overfitting of the Regression Model

A natural way to forecast the RT with the EEG signals is to formulate it as a regression task (Fig. 2), namely finding a (non)linear mapping (e.g., neural networks, SVR) from the EEG signals x to the corresponding RT. However, due to the easily corrupted properties of the EEG signals and the existence of extreme values in RTs during data collection (Wei et al., 2015; Huang et al., 2015), focusing on predicting the exact value of a noisy and non-smooth measurement (i.e., RT) is easier to create a near-perfectly fitted model with poor generalization performance (see Fig. 2, Table 1 and Fig. 5).

This creates a dilemma: it requires a reliable learning model to predict RT with the complex EEG signals (indeed, it is exactly our target) but it is not required to excessively approximate the exact value of RT, especially the extreme values. Here comes the problem: how to find an efficient way to learn from the noisy RT/non-smooth while the exact value is not necessary.

2.2 Consistency of the Ordinal Regression Model

Instead of using regression, we propose to transform the problem into an ordinal regression problem. In particular, the RTs are defined in the totally ordered space R. The totally ordered space owns its structure meanings, which are preserved by the pairwise comparisons between the RTs. The pairwise comparisons indeed preserve the whole

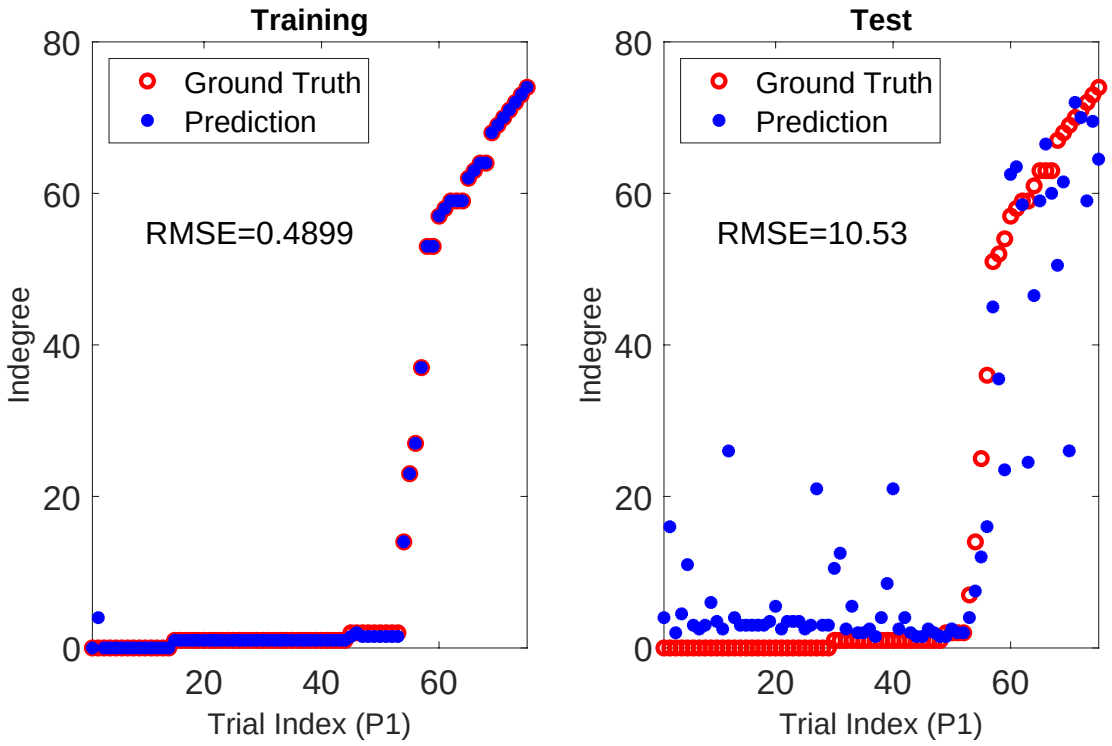

Figure 3: Consistency of the two-layer ordinal regression model using brain dynamics preferences. EEG signals from multiple channels are simply concatenated. Indegree sequences for the ground truth and the prediction are calculated, respectively. The root-mean-squared error (RMSE) was also measured between the indegree sequences of the ground truth and the prediction. We only collect the result from the first participant for a showcase.

relative structure information between the RTs while ignoring their absolute numerical information. Therefore, predicting the orderings of the pairwise comparisons may be regarded as a relaxed alternative of the previous regression model (see Fig. 3).

We showcase our motivation using a naive ordinal regression model for mental fatigue monitoring and show the results in Fig. 3. It shows that even the naive ordinal regression model could capture some meaningful results comparing to the regression model. In particular, the relative structure information between the RTs is kind of preserved, namely the boundary between large RTs and small RTs is clear. Meanwhile, large RTs could serve as an indicator of mental fatigue monitoring.

Reliability Issues Caused by Heterogeneous Channels However, a naive ordinal regression method still suffers from overfitting, mainly because of the simple concatenation of the EEG signals. Since the EEG signals are from heterogeneous channels, it would inevitably lead to degradation of the model's generalization if we simply concatenate the EEG signals without discriminating the reliability of each channel.

3 Model and Methodology

In this section, we first formulate the mental fatigue monitoring task into a multi-channel ranking problem. Further, we extend the ordinal classification model for brain dynamics preferences and introduce a transition matrix to evaluate the channel reliability. Then, the channel-reliability aware ranking model is proposed to tackle the multi-channel ranking problem.

Note that we used the term "preference" intentionally to show that brain dynamics keep changing with regard to human behaviours and it happens because the human brain prefers one decision over others (Ekman and Davidson, 1994; Izuma and Adolphs, 2013; Franks, 2019). Therefore, we prefer to call it "preference" rather than "classification". Our task is then referred to as the "Brain Dynamics Preference (BDP)".

3.1 Multi-Channel Ranking

According to our analysis, our aim is then to correctly preserve the whole orderings between the pairwise RT comparisons (Fig. 1(b)). In particular, the collection of the pairwise RT comparisons $\mathscr{D}$, named as preference propositions, can be constructed as follows:

$$\mathscr{D} = \{(T_i, T_j) | T_i, T_j \in \mathscr{T}, i \neq j\}. \tag{1}$$

where $\mathscr{T}$ is the set of reaction times. Note that the ground truth of each pairwise RT comparison is accessible since RTs are known. Since the connection between RT and BDP is based on human intuition. Therefore, the ground truth of the pairwise RT comparison is called as "preference proposition" in this paper, with regards to BDP.

For brevity of notations, we adopt the new notation to represent the preference propositions in the following. Namely,

$$\mathscr{D} = \{\rho_m : (T_{m,1}, T_{m,2})\}_{m=1}^{M}, \tag{2}$$

where M denotes the number of preference propositions and $\rho_m (\in \mathscr{D})$ denotes the m-th preference proposition. There are usually two types of preference propositions: (1) $\rho_m = 1/-1$: the orderings between the RTs are significant, i.e., $T_{m,1} \geq T_{m,2}$ or $T_{m,1} \leq T_{m,2}$; (1) $\rho_m = 0$: the RTs in each comparison are comparable, i.e., $T_{m,1} \approx T_{m,2}$.

Then, the BDP could be constructed for each proposition using the corresponding pairwise EEG signals recorded from each channel, respectively. Namely

$$\text{preference propositions} \quad \rho_m : (T_{m,1}, T_{m,2}) \qquad \Longleftrightarrow \qquad \text{BDP} \quad (x_{n,m}^1, x_{n,m}^2), \tag{3}$$

where $n = 1, 2, \ldots, N$. The BDP $(x_{n,m}^1, x_{n,m}^2)$ denotes the EEG signals recorded within the n-th channel for each preference proposition ρ_m $\forall m = 1, 2, \ldots, M$.

Multi-Channel Ranking (MCR) In summary, our problem is formulated as predicting the preference propositions (the ordering of the pairwise RT comparisons) by aggregating the BDPs from multiple channels, namely

$$f(\{x_{n,m}^1, x_{n,m}^2\}_{n=1}^{N}) \longrightarrow \rho_m, \quad \forall m = 1, 2, \ldots, M, \tag{4}$$

3.2 Beyond Ordinal Classification

Note that a preference proposition ρ has three states: $1, 0, -1$, denoting win ($T_1 > T_2$), tie ($T_1 \approx T_2$) and loss ($T_1 < T_2$), respectively[1]. It means that classical classification models, e.g., support vector machine, ordinal logistic regression, are infeasible for our problem, due to lack of a normalized probability definition for three states. The softmax function, which models different states equally, also does not serve as a good candidate, failing to capture the intrinsic connection of these two types of preference propositions.

Therefore, we tailor-define a normalized probability for the three states while considering the two types of preference propositions, namely first normalizing the probability over states $(1, -1)$ (exclusively to the preference proposition) to 1, then generalizing the probability definition to state 0. To be specific, it can be mathematically formulated as

$$P(\rho|w, x^1, x^2) = \begin{cases} \sigma(w^T\Delta x)[1 - \kappa(w^T\Delta x)] & \rho = 1, \\ \kappa(w^T\Delta x) & \rho = 0, \\ \sigma(-w^T\Delta x)[1 - \kappa(w^T\Delta x)] & \rho = -1, \end{cases} \tag{5}$$

where $\sigma(z) = 1/(1+e^{-z})$ is the sigmoid function and $\sigma(-z) = 1 - \sigma(z)$. Let Δx denote the subtraction $(x^1 - x^2)$ between the BDP (x^1, x^2). Following Weng and Lin (2011), the probability of a tie is modelled as the geometric mean between a win and a loss, namely $\kappa(w^T\Delta x) = \sqrt{\sigma(w^T\Delta x)\sigma(-w^T\Delta x)}$. Note that we consider the linear mapping $w^T\Delta x$ here since the EEG data are usually high-dimensional with low sample size.

Remark 1 (Extension to Deep Models)**.** *For the sake of clarity, we elaborate our three-states ordinal classification with a linear formulation (Eq.* (5)*). In the case of a deep learning model, the raw feature x in Eq.* (5) *could be replaced with the output of the last layer of the encoder.*

3.3 Channel Reliability

Considering different functions of different regions in the human brain, relative contributions of different channels to human RT may vary a lot. For example, the information conveyed by positive channels is positively related to the RT; while negative channels may convey the information which is negatively related to the RT. There are also some noisy (non-relevant) channels which are independent of the learning task. Therefore, if we directly model the EEG preferences recorded in each channel without any distinctions between the channels regarding the channel state (i.e., positive, noisy and negative), the model's reliability would inevitably degrade. Note that a channel is called "noise" if the current algorithms could not extract useful brain information with EEG signals from this channel (Alharbi, 2018; Lin et al., 2018).

In the following, a transition matrix Π_n is introduced to characterize the reliability of each channel n w.r.t. the learning task. Let ρ denote the preference proposition and $\rho^{(n)}$ denote the prediction from the n-th channel. ρ and $\rho^{(n)}$ are all defined on a finite

[1] In the following, we omitted the subscripts for simplicity.

state space $S = \{1, 0, -1\}$. Then we have

$$\Pi_n = P(\rho|\rho^{(n)}) = \begin{bmatrix} \pi_{11}^n & \pi_{12}^n & \pi_{13}^n \\ \pi_{21}^n & \pi_{22}^n & \pi_{23}^n \\ \pi_{31}^n & \pi_{32}^n & \pi_{33}^n \end{bmatrix}, \tag{6}$$

where $P_{i,j}(\rho|\rho^{(n)}) = P(\rho = S_j|\rho^{(n)} = S_i)$. According to the definition of the transition matrix, Π_n should satisfy the following three constraints: (1) each entry of Π_n should be constrained in $[0, 1]$; (2) each raw of Π_n should be summed up to be 1; (3) each column of Π_n should be summed up to be 1.

However, it is usually costly and redundant to estimate Π_n (Eq. (6)) directly. In the following, we consider to impose more constraints to Eq. (6), so as to simplify the inference while enhancing interpretability. (1) The transition between states $(1, -1)$ is constrained to be symmetric, since states $(1, -1)$ are exclusive to the preference proposition wherein the orderings between the RTs are significant, i.e., $P(\rho = 1|\rho^{(n)} = -1) = P(\rho = -1|\rho^{(n)} = 1)$. (2) Since the equal case between two real values are hard to measure when conducting prediction, the transition from the significant RT pairwise comparisons to comparable RT ones are not considered[2], i.e., $P(\rho = 0|\rho^{(n)} = \{1, -1\}) = 0$. Therefore, a simplified transition matrix can be represented as follows:

$$\Pi_n = P(\rho|\rho^{(n)}) = \begin{bmatrix} \pi_n & 0 & (1-\pi_n) \\ 0 & 1 & 0 \\ (1-\pi_n) & 0 & \pi_n \end{bmatrix}, \tag{7}$$

The parameter π_n in the transition matrix Π_n (Eq. (7)) actually indicates the reliability of the n-th channel $\forall n = 1, 2, \ldots, N$. It additionally helps to divide the channels into three states: (a) positive channels with π_n close to 1, the ranking model (Eq. (5)) can extract enough information from the n-th channel, and exactly predict the state of the preference proposition. (b) Noisy channels with π_n approximating to 0.5, the ranking model cannot extract any useful information from the n-th channel. (c) Negative channels with π_n close to 0, the ranking model can extract enough information from the n-th channel, but the prediction states are exactly opposite to the proposition states.

The identified positive and negative channels are all considered as informative EEG channels, which helps in learning reliable models for the corresponding task.

3.4 Channel-reliability Aware Ranking (CArank)

With the incorporation of transition matrix Π_n (Eq. (7)) on top of the introduced three states learning to rank model (Eq. (5)), the likelihood function for each preference proposition ρ can be represented as

$$\begin{aligned} P(\rho|w, \Pi_n, x_n^1, x_n^2) &= \mathbb{E}_{\rho^{(n)}}\left[P(\rho|\rho^{(n)})P(\rho^{(n)}|w, x_n^1, x_n^2)\right] \\ &= \begin{cases} [\pi_n\sigma(w^T\Delta x_n) + (1-\pi_n)\sigma(-w^T\Delta x_n)][1-\kappa(w^T\Delta x_n)] & \rho = 1, \\ \kappa(w^T\Delta x_n) & \rho = 0, \\ [(1-\pi_n)\sigma(w^T\Delta x_n) + \pi_n\sigma(-w^T\Delta x_n)][1-\kappa(w^T\Delta x_n)] & \rho = -1. \end{cases} \end{aligned} \tag{8}$$

[2] A promising approach to generalize the transition matrix Π_n (Eq. (7)) is to introduce the concept of the confidence region to measure the equal cases (Pregibon et al., 1981).

where the subscripts m, indicating the index of preference proposition, are omitted for the sake of simplicity.

Let $\mathscr{D}$ denote the collection of preference propositions and X represent the recorded EEG signals from N different channels. We further extend Eq. (8) to a Bayesian formulation. A Gaussian prior is introduced for w (i.e., $w \sim N(\mu, \Sigma)$). Since the transition matrix Π_n only depends on the parameter π_n, we focus on estimating the parameter π_n $\forall n = 1, 2, \ldots, N$ in the following. Let π denote $\{\pi_n\}_{n=1}^N$, and we introduce a Beta prior for each π_n (i.e., $\pi \sim B(\alpha, \beta) = \prod_{n=1}^N B(\alpha_n, \beta_n)$). Then, our Channel-reliability Aware Ranking (CArank) model (Eq. (9)) for multi-channel ranking problem (Eq. (4)) can be represented as

$$\begin{aligned} P(\mathscr{D}, w, \pi | X) &= P_0(\pi) P_0(w) P(\mathscr{D} | w, \pi, X) \\ &= B(\pi | \alpha, \beta) N(w | \mu, \Sigma) \prod_{m=1}^{M} \prod_{n=1}^{N} P(\rho_m | w, \pi_n, \Delta x_{n,m}). \end{aligned} \tag{9}$$

Let M denote the number of preference propositions, i.e., $|\mathscr{D}| = M$. The variable n iterates over the channels. m iterate over preference propositions. Due to the symmetry of the state probability (Eq. (5)) and transition matrix (Eq. (7)) w.r.t. states 1 and -1, the resulting marginal likelihood (Eq. (8)) and the corresponding Bayesian formulation (Eq. (9)) remains symmetric w.r.t. states 1 and -1.

Now our aim is to estimate the model parameters (w and π) by maximizing Eq. (9). In principle, any solution strategies for MAP estimation can be considered to solve this problem. See Section 4 for optimization details.

3.5 Reliability Analysis and Channel State Estimation

CArank (Eq. (9)) indeed trains a mixture of two complementary classifiers, which share the same parameter w. It is different from classical mixture models, since it clusters at the channel level instead of the sample level.

In particular, in terms of the positive channels with π_n close to 1, CArank relies on the first classifier to update the shared parameter w. In terms of the negative channels with π_n close to 0, Eq. (9) automatically switches to the opposite classifier which can extract correct information from the negative channels and update the shared parameter w accordingly. Further, CArank is robust to the noisy channels with π_n approximately equal to 0.5, because Eq. (9) gives up extracting information from the noisy channels by assigning a constant likelihood (i.e., 0.5) to each BDP. The estimated π_n can be leveraged as an indicator to detect noisy channels with $\pi_n \approx 0.5, \forall n = 1, 2, \ldots, N$. See Fig. 6 for more details.

4 Stochastic Generalized Expectation Maximization

In this section, we describe a generalized expectation-maximization (GEM) algorithm (Dempster et al., 1977) to solve the proposed CArank (Eq. (9)). Since the feasible region of π_n is restricted to $[0, 1]$, the gradient-based optimization methods would make our solution inaccurate and inefficient. The GEM algorithm is an efficient iterative

procedure to compute the MAP solution in the presence of latent variables ($\rho_m^{(n)}$ in Eq. (9)). GEM avoids directly calculating the derivative to the expectation of latent variables, and resorts to a surrogate lower bound to optimize. Therefore, GEM, a silver bullet for MAP with latent variables, can significantly simplify the optimization over parameter π_n for Eq. (9).

4.1 GEM for CArank

For each preference proposition ρ_m, we introduce an auxiliary variable $\delta_m^{(n)} \in \{1,0\}$ for the n-th channel, representing the consistency between the preference proposition ρ_m and the prediction $\rho_m^{(n)}$ given by the n-th channel. Specifically, $\delta_m^{(n)} = 1$ denotes the prediction $\rho_m^{(n)}$ given by the first classifier is consistent with the preference proposition ρ_m, and $\delta_m^{(n)} = 0$ denotes the prediction $\rho_m^{(n)}$ estimated by the second classifier is consistent with the preference proposition ρ_m. We can therefore find an equivalent formulation of Eq. (8) for each preference proposition ρ_m involving the auxiliary variable $\Xi_m = \{\delta_m^{(n)}\}_{n=1}^N$.

$$P(\rho_m, \Xi_m | \pi, w, X) = \prod_{n=1}^{N} P(\rho_m, \delta_m^{(n)} | \pi_n, w, \Delta x_{n,m}) \tag{10}$$

$$= \begin{cases} \prod_{n=1}^N [\pi_n \sigma(w^T \Delta x_{n,m})]^{\delta_m^{(n)}} [(1-\pi_n)\sigma(-w^T \Delta x_{n,m})]^{1-\delta_m^{(n)}} [1 - \kappa(w^T \Delta x_{n,m})] & \rho_m = 1, \\ \prod_{n=1}^N \kappa(w^T \Delta x_{n,m}) & \rho_m = 0, \\ \prod_{n=1}^N [(1-\pi_n)\sigma(w^T \Delta x_{n,m})]^{\delta_m^{(n)}} [\pi_n \sigma(-w^T \Delta x_{n,m})]^{1-\delta_m^{(n)}} [1 - \kappa(w^T \Delta x_{n,m})] & \rho_m = -1. \end{cases}$$

This shows that we can deal with the joint distribution directly, which leads to significant simplifications for optimization. The complete log likelihood of CArank (Eq. (9)) can be written as

$$\log P(\mathscr{D}, \Xi, w, \pi | X) = \log P_0(\pi) + \log P_0(w) + \sum_{m=1}^{M} \sum_{n=1}^{N} \log P(\rho_m, \delta_m^{(n)} | w, \pi_n, \Delta x_{n,m}). \tag{11}$$

Expectation Step In the expectation step, we first calculate the expected value of the auxiliary variable $\delta_m^{(n)}$ w.r.t. its posterior distribution $P(\delta_m^{(n)} | \pi, w, \rho_m, x_{n,m})$ $\forall n = 1, 2, \ldots, N, \forall m = 1, 2, \ldots, M$:

$$\mathbb{E}[\delta_m^{(n)}] = \frac{P(\rho_m, \delta_m^{(n)} | w, \pi_n, \Delta x_{n,m})}{P(\rho_m | w, \pi_n, \Delta x_{n,m})} = \begin{cases} \left[1 + \frac{(1-\pi_n)\sigma(-w^T \Delta x_{n,m})}{\pi_n \sigma(w^T \Delta x_{n,m})}\right]^{-1} & \rho_m = 1, \\ 1 & \rho_m = 0, \\ \left[1 + \frac{\pi_n \sigma(-w^T \Delta x_{n,m})}{(1-\pi_n)\sigma(w^T \Delta x_{n,m})}\right]^{-1} & \rho_m = -1. \end{cases} \tag{12}$$

where $\mathbb{E}[\delta_m^{(n)}]$ denotes the degree of the consistency between the prediction $\rho_m^{(n)}$ and the preference proposition ρ_m. Then, the expectation of Eq. (9) w.r.t. the posterior distribu-

tion $P(\delta_m^{(n)}|\pi, w, \rho_m, x_{n,m})\ \forall n = 1,2,\ldots,N, \forall m = 1,2,\ldots,M$ can be represented as:

$$
\begin{aligned}
&\mathscr{L}(w,\pi) = \mathbb{E}[\log P(\mathscr{D}, \Xi, w, \pi|X)] \qquad (13)\\
&= \sum_{n=1}^{N}[(\alpha_n - 1)\log\pi_n + (\beta_n - 1)\log(1-\pi_n)] - \frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\\
&\quad + \sum_{m=1}^{M}\sum_{n=1}^{N}\Big[\mathbb{I}(\rho_m = 0)\log\kappa(w^T\Delta x_{n,m}) + \mathbb{I}(\rho_m \neq 0)\log[1-\kappa(w^T\Delta x_{n,m})]\\
&\quad + \mathbb{I}(\rho_m = 1)\big[\mathbb{E}[\delta_m^{(n)}]\log\pi_n\sigma(w^T\Delta x_{n,m}) + (1-\mathbb{E}[\delta_m^{(n)}])\log(1-\pi_n)\sigma(-w^T\Delta x_{n,m})\big]\\
&\quad + \mathbb{I}(\rho_m = -1)\big[\mathbb{E}[\delta_m^{(n)}]\log(1-\pi_n)\sigma(w^T\Delta x_{n,m}) + (1-\mathbb{E}[\delta_m^{(n)}])\log\pi_n\sigma(-w^T\Delta x_{n,m})\big]\Big],
\end{aligned}
$$

where $\mathbb{I}(*)$ is the indicator function, which equals 1 if the condition is true and 0 otherwise.

Generalized Maximization Step In the generalized maximization step, we increase the objective function Eq. (13) w.r.t. the model parameters π and w, respectively. In terms of π, we set the gradient of Eq. (13) w.r.t. π_n to zero and obtain the following estimate for π_n:

$$
\pi_n^{\text{new}} = \frac{\sum_{m=1}^{M}\Big[\mathbb{I}(\rho_m = 1)\mathbb{E}[\delta_m^{(n)}] + \mathbb{I}(\rho_m = -1)(1-\mathbb{E}[\delta_m^{(n)}])\Big] + \alpha_n - 1}{\sum_{m=1}^{M}[\mathbb{I}(\rho_m = 1) + \mathbb{I}(\rho_m = -1)] + \alpha_n + \beta_n - 2}, \qquad (14)
$$

where $n = 1,2,\ldots,N$.

In terms of w, due to the complexity of the sigmoid function, we cannot have a closed-form solution for w and we need to use gradient-based methods to optimize Eq. (13) w.r.t. w. In particular, the gradient function $g(w)$ can be represented as follows:

$$
\begin{aligned}
g(w) = -\Sigma^{-1}(w-\mu) + \sum_{m=1}^{M}\sum_{n=1}^{N}\Big[\big[\mathbb{I}(\rho_m = 0) + \frac{\mathbb{I}(\rho_m \neq 0)}{1-[\kappa(w^T\Delta x_{n,m})]^{-1}}\big]\frac{1-2\sigma(w^T\Delta x_{n,m})}{2}\\
+ \mathbb{I}(\rho_m \neq 0)(\mathbb{E}[\delta_m^{(n)}] - \sigma(w^T\Delta x_{n,m}))\Big]\Delta x_{n,m}. \qquad (15)
\end{aligned}
$$

Regarding the linear rank mapping, we adopt the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) (Byrd et al., 1995) to optimize w. w^{new} can be obtained with L-BFGS using $\mathscr{L}(w)$ and $g(w)$, namely

$$
w^{\text{new}} = \text{L-BFGS}(\mathscr{L}(w), g(w), \mathscr{D}). \qquad (16)
$$

The GEM algorithm (See algorithm 1) then iterates the E-step and the generalized M-step until convergence is achieved.

Remark 2 (Extension to Deep Models)**.** *The L-BFGS optimization method used in Eq.* (16) *aims to find the optimum w. It is easy to find its alternatives in deep learning literature, such as vanilla stochastic gradient descent (SGD) and its various variants (Kasai, 2017), if we replace the raw EEG feature x with neural embedding.*

Algorithm 1 Generalized Expectation-Maximization (GEM) algorithm for Channel-reliability Aware Ranking (CArank)

1: **Input:** collection of preference propositions $\mathscr{D}$, EEG signals X, maximum number of iterations T.
2: **Initialization:** hyperparameters$\{\alpha_n, \beta_n\}_{n=1}^{N}$ for π and (μ, Σ) for w.
3: **for** $t = 1, 2, \ldots, T$ or not convergence **do**
4: *E-step:* calculate the posterior expectation of auxiliary variable $\mathbb{E}[\delta_m^{(n)}]$ according to Eq. (12), $\forall m = 1, 2, \ldots M, \forall n = 1, 2, \ldots N$.
5: *M-step:* update π_n according to Eq. (14) $\forall n = 1, 2, \ldots N$ and update w according to Eq. (16).
6: **end for**
7: **Output:** channel reliability π and rank mapping w.

4.2 Stochastic GEM for CArank

The GEM approach introduced in Section 4.1 is inefficient for large-scale datasets, because we need to iteratively calculate the gradient w.r.t. parameters π and w over all samples during each generalized maximization step. Motivated by the stochastic approximation literature (Roche, 2011), we introduce a stochastic generalized expectation-maximization (SGEM) approach, which resorts to stochastic mini-batch optimization to learn the parameters. To be specific, SGEM approximates the updated π and w in batch EM with a single sample or mini-batch samples. Since mini-batch samples cannot be a perfect approximation to the whole dataset, we interpolate between the new and former estimators with a decreasing step-size[3] η_k, as in (Liang and Klein, 2009).

Sampling Step Before the t-th iteration, we randomly sample a mini-batch $\mathscr{D}^t$ from $\mathscr{D}$. The number of preference propositions in $\mathscr{D}^t$, denoted by M^t, is much smaller than the corresponding total dataset size M.

Expectation Step The expectation step remains similar. The only difference is that we need to calculate the posterior expectation of the auxiliary variable $\delta_m^{(n)}$ over the mini-batch $\mathscr{D}^t$.

Generalized Maximization Step In the generalized maximization step, We increase the objective function, calculated on the mini-batch $\mathscr{D}^t$, w.r.t. model parameters π and w. In terms of parameter π_n, since its marginal distribution belongs to the exponential family, we perform the stochastic update in the space of sufficient statistics (Cappé and

[3] Here, the step-size is set to $\eta_t = (t+2)^{-\tau_0}$, where t is the number of iterations and $0.5 < \tau_0 < 1$. The smaller the τ_0 is, the larger the update η_t is, and the more quickly we forget (decay) our old parameters. This can lead to swift progress but also generates instability.

Moulines, 2009). Let $\tilde{\phi}_n$ denote the noisy estimate of the sufficient statistic for π_n.

$$\tilde{\phi}_n = \frac{M}{M^t} \sum_{m \in \mathscr{D}^t} \left[\mathbb{I}(\rho_m = 1) \mathbb{E}[\delta_m^{(n)}] + \mathbb{I}(\rho_m = -1)(1 - \mathbb{E}[\delta_m^{(n)}]) \right], \tag{17a}$$

$$\phi_n^t = (1 - \eta_t)\phi_n^{t-1} + \eta_k \tilde{\phi}_n, \tag{17b}$$

$$\pi_n^{\text{new}} = \frac{\phi_n^t + \alpha_n - 1}{\sum_{m \in \mathscr{D}^t} [\mathbb{I}(\rho_m = 1)] + \mathbb{I}(\rho_m = -1)] + \alpha_n + \beta_n - 2}, \quad n = 1, 2, \ldots, N. \tag{17c}$$

In terms of parameter w, the above practice is infeasible due to its non-exponential marginal distribution. Inspired by the stochastic gradient EM algorithms in (Cappé and Moulines, 2009), we perform the stochastic update in the original space. First, a local optima regression weight w^t can be obtained via iterative optimization over the mini-batch $\mathscr{D}^t$, using L-BFGS. Then we interpolate between a local optimum and the former estimations to form a global approximation w.r.t. the parameter w.

$$w^t = \text{L-BFGS}(\mathscr{L}(w), g(w), \mathscr{D}^t), \tag{18a}$$

$$w^{\text{new}} = (1 - \eta_k) w^{\text{old}} + \eta_k w^t. \tag{18b}$$

Remark 3 (Convergence Analysis)**.** *The convergence issues of the proposed stochastic GEM algorithm are analogous to the discussion given by Cappé and Moulines (2009) for their stochastic gradient EM algorithms. The existence of such links is hardly surprising. In view of the discussions in Section 3 of Cappé and Moulines (2009), the online update rule (Eq.* (18b)*) could also be seen as a stochastic gradient recursion formula, namely* $w^{\text{new}} = w^{\text{old}} + \eta_k (w^t - w^{\text{old}})$.

5 Empirical Analysis

In this section, we demonstrate the reliability of the proposed CArank (Eq. (9)) with EEG signals from forty participants.

Experiment Paradigm: This paper utilized the 33-channels EEG data recorded in the previous study (Huang et al., 2015) from 40 adult participants while performing long sustained attention task. This data contains one intrinsic non-EEG channel, i.e., 33rd channel, which contains the information about only one axis in the direction of deviation. The experiment has been conducted using a virtual-reality dynamic driving simulator (Fig. 4D-E). The task involves driving on the four-lane highway while lane-departure events were randomly induced deviation toward the side of the road from the original position. Each participant was instructed to quickly respond to steer back to the original position. A complete trial in this study (Fig. 4A), includes 10s baseline, deviation onset, response onset, and response offset (Fig. 4B-C). The next trial occurs within an interval of 5-10s after finishing the current trial. Each participant completed T trials within 1.5h. For each trial i, the EEG signals $\{x_{n,i}\}_{n=1}^N$ from N different channels were recorded simultaneously and the corresponding reaction time RT_i was also collected afterward. If a participant fell asleep during the experiment, there was no feedback to wake him up.

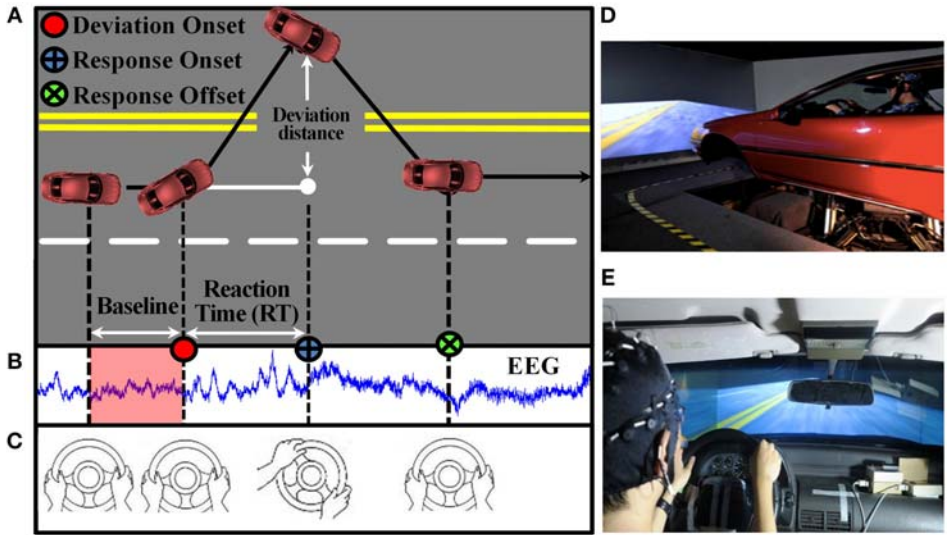

Figure 4: Sustained-attention driving task

In this paper, the 10s baseline (Fig. 4B) as the feature vector has been adopted, which is assumed to be long enough to detect any significant changes in brain activity (Zhang, 2000). This followed by exploring the relationship between the 10s baseline x $(\in R^k)$ and the preference proposition ρ_m under the following four assumptions: (a) different participants are independent during the data collection process; (b) Different EEG sensors used for recording are recorded independently from scalp without influencing other sensors (Homan et al., 1987; Teplan et al., 2002); (c) Different trials are conducted independently during the data collection process; (d) The collected reaction time is slightly corrupted by inherent (basically irremovable) sources of noise, but the ranking relationships are preserved to some extents.

Data Preprocessing: Brain dynamics preferences for each participant has been generated as follows: (a) the trials of each participant were randomly divided into two parts: 50% for training and 50% for test; (b) EEG preferences were constructed according to the pairwise comparisons between the RTs. To be specific, two types of RT comparisons could be constructed: (1) significant RT pairwise comparisons $(T_{m,1}, T_{m,2})$, where $T_{m,1} \gg T_{m,2}$ or $T_{m,2} \gg T_{m,1}$; (2) comparable RT pairwise comparisons $(T_{m,1}, T_{m,2})$, where $T_{m,1} \approx T_{m,2}$. Considering the time delay among the channels in the time domain, Fourier transform (Welch, 1967) has been applied to EEG signals to transform time-series into 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).

Baselines: We compared CArank with widely adopted methods: Regression and Classification methods under the multiple channel concatenation formulation and the multiple channel aggregation formulation, respectively. In particular, two (deep) regression models (Lin et al., 2014; Hajinoroozi et al., 2016) are considered: (1) Regression (C), the EEG signals from multiple channels are simply concatenated; (2) Regression

(A), the EEG signals from multiple channels are considered independently and the regression results are aggregated using majority voting afterward. Two (deep) ordinal classification model (Zarei, 2017; Liu et al., 2018; Zeng et al., 2018) are considered: (1) Classification(C), the EEG signals from multiple channels are simply concatenated; (2) Classification (A), the EEG signals from multiple channels are considered independently and the classification results are aggregated using majority voting afterward.

Metrics: First, we aggregate the predictions from different channels using a simple voting scheme, namely

$$\hat{\rho}_m = \text{sign}\left(\sum_{n=1}^{N} \rho_m^{(n)} \left[\mathbb{I}(\pi_n > \kappa) - \mathbb{I}(\pi_n < 1-\kappa) \right] \right),$$

where $\rho_m^{(n)}$ denotes the predicted state (1 means win and -1 means loss) for the pairwise RT comparison $(T_{m,1}, T_{m,2})$ by the n-th channel, using the brain dynamics preference $(x_{n,m}^1, x_{n,m}^2)$. $\hat{\rho}_m$ is the final estimated order for $(T_{m,1}, T_{m,2})$ by aggregating the predictions $\rho_m^{(n)}$ over all channels. $\mathbb{I}(*)$ is an indicator that returns one if the argument is valid and zero otherwise.

Then, we introduce two metrics to measure the performance of CArank model from different perspectives. First, we adapted the Wilcoxon-Mann-Whitney statistics (Yan et al., 2003) to evaluate the accuracy (in %, higher is better) over all pairs, namely

$$\text{Acc} = \frac{1}{\bar{M}} \sum_{m=1}^{M} \mathbb{I}(\rho_m = \hat{\rho}_m), \qquad \bar{M} = \sum_{m=1}^{M} \mathbb{I}(\rho_m \neq 0). \tag{19}$$

Further, we investigate the reliability of CArank in terms of preserving the global ordering w.r.t RTs. Note that a totally ordered set could be equally represented by a fully directed graph, where the fully directed graph can be further encoded by its degree sequence. We only consider the indegree sequence because the indegree and outdegree of a vertex can be uniquely determined when the overall degree of each vertex is fixed. The indegree of vertex v_i can be calculated as

$$\widehat{\text{Indeg}}(v_i) = \sum_{m \in N_1(v_i)} \mathbb{I}(\hat{\rho}_m = 1) + \sum_{m \in N_2(v_i)} \mathbb{I}(\hat{\rho}_m = -1) + \sum_{m \in N_1(v_i) \cup N_2(v_i)} 0.5 \times \mathbb{I}(\hat{\rho}_m = 0), \tag{20}$$

where $N_1(v_i), N_2(v_i)$ denote the index set of the pairwise comparisons with the RT of trial i (vertex v_i) appearing in the first and second position, respectively. Therefore, we first collected the indegree sequences (Becirovic, 2017) of the constructed directed graph using the predicted RTs and then measured the discrepancy between the predicted indegree sequences and ground truth using the root-mean-squared errors (smaller is better). Namely

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{i=1}^{T} [\text{Indeg}(v_i) - \widehat{\text{Indeg}}(v_i)]^2}. \tag{21}$$

where T denotes the number of trails for each participant. $\text{Indeg}(v_i)$ is the ground truth indegree of vertex v_i while $\widehat{\text{Indeg}}(v_i)$ is the predicted indegree of vertex v_i.

Note that we only trust the predictions from informative channels with reliability $\pi_n > \kappa$ or $\pi_n < 1-\kappa$. κ is set to 0.85 for all participants in our experiment. In terms of Regression (C)/ Classification (C), It is a simple regression/ classification problem, since the EEG signals from multiple heterogeneous channels are simply concatenated. We train a two-layer neural network for Regression (C)/ Classification (C), respectively. In terms of Regression (A)/ Classification (A), considering the high-dimensional feature with low sample size, we train a two-layer neural network shared by all channels and aggregate the results from different channels to calculate the final predictions using the majority voting scheme. Since there is no mechanism for Regression (C/A), Classification (C/A) to evaluate the channel state, we trust all the channels by default. Further, we only calculate the two metrics on the preference propositions wherein the orderings between the RT pair are significant, since it is hard to evaluate when the orderings between the RT pair are comparable.

Parameter Initialization: A two-layer neural network, with the hidden neuron size being 100, is implemented for our CArank. In terms of the channel reliability π_n, we aimed to eliminate the effects of noisy channels during the training process, and therefore initialized the channel reliability π_n to 0.5, $\forall n = 1,2,\ldots,N$. The L2 norm is used, which equals to adopt the standard Gaussian distribution for w, i.e., $w \sim N(\mathbf{0},\mathbf{1})$. In terms of the hyperparameters (α_n,β_n), as we intended to eliminate the effects of noisy channels, we adopted a strong non-informative prior for π_n, namely $\alpha_n = \beta_n = 100$, $\forall n = 1,2,\ldots,N$, according to Bishop (2006). The Adam method is used to optimize the weight[4] w. In terms of the maximum iteration number, we set MaxIter $= 7$ in our experiment to ensure the algorithm converged for each participant. The minibatch size is set to 256 and the learning rate is 0.001. For the sake of a fair comparison, we also implemented a two-layer neural network, with the hidden layer being 100, for each baseline. In terms of Regression (C/A), the commonly mean square error (MSE) is adopted as the loss function. In terms of Classification (C/A), the negative log-likelihood (Eq. (13)) is adopted as the loss function, except that π_n is fixed to 1, $\forall n = 1,2,\ldots,N$.

5.1 Empirical Results of CArank on Brain Dynamics Preferences

The Wilcoxon-Mann-Whitney statistics (Eq. (19)) of all baselines and CArank on the test BDPs are presented in Table 1. In terms of Regression (C/A), we first collected the predicted RTs on the test EEG signals and then calculated the Wilcoxon-Mann-Whitney statistics of the predicted RTs w.r.t. the ground truth.

From Table 1, we observe that: (1) **CArank** $>$ **other baselines**. CArank exhibits consistent improvements over other baselines. In particular, CArank achieves the highest test accuracy on 33 participants. This is consistent with our motivation that classification served as a relaxed alternative for regression, can effectively circumvent the overfitting caused by non-smooth RTs and preserve the ordering w.r.t. RTs. Meanwhile, our channel-reliability aware formulation could also eliminate the effects of the EEG signals from a noisy channel during the training process, compared with using simple concatenation.

[4] In terms of the L-BFGS implementation, a Matlab code could be downloaded from (Granzow, 2017).

Table 1: Test accuracy (in %). Higher is better, the best is marked in gray

Participant		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Test ACC	Regression (C)	71.63	79.21	80.22	72.39	83.65	68.38	60.31	54.99	77.98	59.01	82.72	89.80	79.56	85.45	68.60	65.88	54.30	50.58	68.65	61.80
	Regression (A)	71.71	72.97	79.81	70.90	82.80	57.42	61.88	60.96	66.38	52.96	79.37	73.87	67.70	80.54	66.03	54.47	51.01	65.07	62.33	54.80
	Classification (C)	76.85	82.48	82.40	74.77	83.12	65.69	76.12	70.84	83.02	63.74	76.41	85.08	77.74	88.03	69.09	71.80	58.44	77.31	80.85	63.56
	Classification (A)	79.97	77.61	79.87	68.69	82.55	63.86	49.85	51.47	51.78	53.03	75.79	79.69	66.40	89.39	68.10	53.07	50.00	52.81	61.19	52.02
	CArank	82.29	80.97	83.78	77.50	87.42	76.62	82.34	79.16	91.40	78.25	81.74	84.17	83.23	90.53	76.66	80.40	88.69	81.13	80.42	78.35
Participant		**P21**	**P22**	**P23**	**P24**	**P25**	**P26**	**P27**	**P28**	**P29**	**P30**	**P31**	**P32**	**P33**	**P34**	**P35**	**P36**	**P37**	**P38**	**P39**	**P40**
Test ACC	Regression (C)	69.84	50.58	80.73	56.85	78.72	67.76	65.06	84.75	79.59	82.59	63.41	66.46	56.78	61.81	66.70	87.21	81.98	57.71	84.41	67.48
	Regression (A)	53.44	58.27	78.29	54.25	77.46	53.31	51.33	77.73	69.92	77.06	58.46	64.09	53.09	59.69	72.45	85.64	73.21	62.83	50.55	46.35
	Classification (C)	68.22	79.82	84.36	68.10	84.28	69.60	77.09	86.46	82.11	86.85	74.22	85.05	60.49	71.58	73.03	90.40	83.51	75.62	80.37	69.07
	Classification (A)	49.86	74.65	72.45	59.46	75.35	49.80	52.20	76.89	51.88	73.62	59.50	61.30	50.00	53.07	60.46	90.24	72.15	60.79	78.30	65.76
	CArank	72.83	85.33	82.70	89.35	84.57	76.52	85.02	83.58	86.56	85.64	92.74	85.74	79.24	84.77	90.53	90.96	86.05	77.12	93.48	75.56

Table 2: Test RMSE (in #). Smaller is better, the best is marked in gray

Participant		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Test RMSE	Regression (C)	13.11	17.40	13.06	17.92	25.98	22.14	22.46	42.16	30.63	18.16	12.04	5.40	13.88	14.03	27.87	12.47	17.93	45.85	43.96	36.64
	Regression (A)	11.92	19.12	13.46	18.27	24.65	26.14	21.17	38.24	37.44	20.53	11.82	11.56	19.15	16.71	26.91	14.82	17.30	35.59	46.23	41.59
	Classification (C)	10.53	15.35	12.51	16.78	27.36	22.85	16.32	33.48	25.18	17.45	15.21	8.64	15.58	12.63	25.24	10.54	15.98	27.56	31.11	35.76
	Classification (A)	9.20	18.28	13.54	18.35	27.15	22.83	25.98	44.48	49.43	20.36	14.61	9.65	18.83	12.23	25.20	15.19	17.42	42.07	48.42	44.38
	CArank	8.66	16.97	12.64	16.27	20.71	19.97	13.98	25.71	12.52	13.70	12.73	9.41	12.43	10.06	23.95	8.99	5.74	25.00	31.39	28.00
Participant		**P21**	**P22**	**P23**	**P24**	**P25**	**P26**	**P27**	**P28**	**P29**	**P30**	**P31**	**P32**	**P33**	**P34**	**P35**	**P36**	**P37**	**P38**	**P39**	**P40**
Test RMSE	Regression (C)	38.13	25.26	26.88	40.08	38.27	36.48	38.22	26.56	36.39	31.98	26.09	22.99	24.67	25.53	13.85	9.94	13.74	52.71	6.66	18.75
	Regression (A)	48.77	22.10	27.19	41.04	37.01	46.19	46.01	31.61	41.87	36.54	26.77	23.33	23.51	25.03	12.49	9.23	17.07	46.63	17.52	24.70
	Classification (C)	40.79	13.76	23.97	31.98	30.39	36.36	28.85	24.44	33.09	26.21	19.45	13.54	22.04	20.48	13.41	7.22	13.46	39.68	9.65	17.33
	Classification (A)	51.50	15.99	30.35	37.33	37.94	47.82	44.97	37.25	57.80	42.96	26.74	23.69	24.75	26.88	19.49	6.59	16.26	46.13	7.61	16.97
	CArank	37.77	11.77	25.49	16.44	29.67	30.72	19.38	26.32	26.00	28.49	8.00	11.65	12.94	12.06	5.34	7.03	11.17	36.77	3.83	15.14

(2) **Classification** $>$ **Regression**. In particular, the test accuracies of classification-based methods for most participants are higher to their regression-based counterparts. Namely, Classification(C) outperforms Regression(C) on 33 participants, while Classification(A) outperforms Regression(A) on 26 participants. This observation is consistent with our statement that regression-based models are easily overfitting, especially when extreme values (RTs in our problem) exist. (3) **Concatenation** $>$ **Aggregation**. It is interesting to note that the test accuracy based on multiple channel aggregation is significantly inferior to their counterparts based on simple feature concatenation. Specifically, Regression(C) outperforms Regression(C) on 33 participants, while Classification(C) outperforms Classification(A) on 38 participants. This is quite impressive but reasonable. Since a shared regression/ classification model is trained in the case of the multiple channel aggregation formulation, the generalization performance would inevitable degenerate when learning with noisy noisy channels. However, the noisy channels universally exist and at least one noisy channel is detected for each participant according to Table 6.

To further investigate the reliability of CArank in terms of preserving the global ordering corresponding to RTs, we first collected the indegree sequences according to Eq. (20) using the predicted RTs and then measured the indegree discrepancy between the calculated indegree sequences and the ground truth using the root-mean-squared error (Eq. (21)). The RMSE for all participants are shown in Table.2. From Table.2, we could draw similar conclusions: (1) our CArank consistently achieves lower RMSE compared to other baselines. In particular, CArank achieves the lowest test RMSE on 32 over 40 participants. (2) Excepted for our CArank, Classification (C) shows better performance over the rest baselines. This is reasonable, since classification is robust to extreme RTs while concatenation approach is less affected by the noisy channels compared to simple aggregation. (3) The difference between other baselines methods becomes ambiguous. This is because RMSE assigned higher punishment to an estimation with a larger error.

To further explore the superiority of our CArank, we visualized Table 2 using the indegree sequences. For the sake of intuitive interpretation, we particularly showcase participants P9, P13, P22, P24, P31 with the most representative performance in Fig. 5. Regarding the rest participants, our CArank also achieves superior performance with the lowest RMSE (See Table 2).

From Fig. 5, we observe that: (1) Overall, the indegree sequences predicted by CArank closely align to the ground truth with slight fluctuates (small RMSE); while the indegree sequences predicted by other baselines fluctuates significantly and fails to maintain the trend with the ground truth (large RMS). (2) The points located in the northeast denote the trials with high RTs (or called extremely RTs). The indegree sequences predicted by CArank show slighter fluctuates, compared to that of other baselines. It denotes that CArank could accurately detect the mental fatigue associated with higher RTs. However, other baselines either show large fluctuates (e.g., P9, P13, P24), leading to a high false-negative rate, or completely fails to maintain the trend, leading to a high error rate. (3) The points located in the southwest denote the trials with small RTs. The indegree sequences predicted by other baselines show large fluctuates (e.g., P22), a high false-positive rate. (4) It is worth noting that the indegree sequences predicted by Regression(C/A) usually fluctuates heavily for low indegree trials (small RTs)

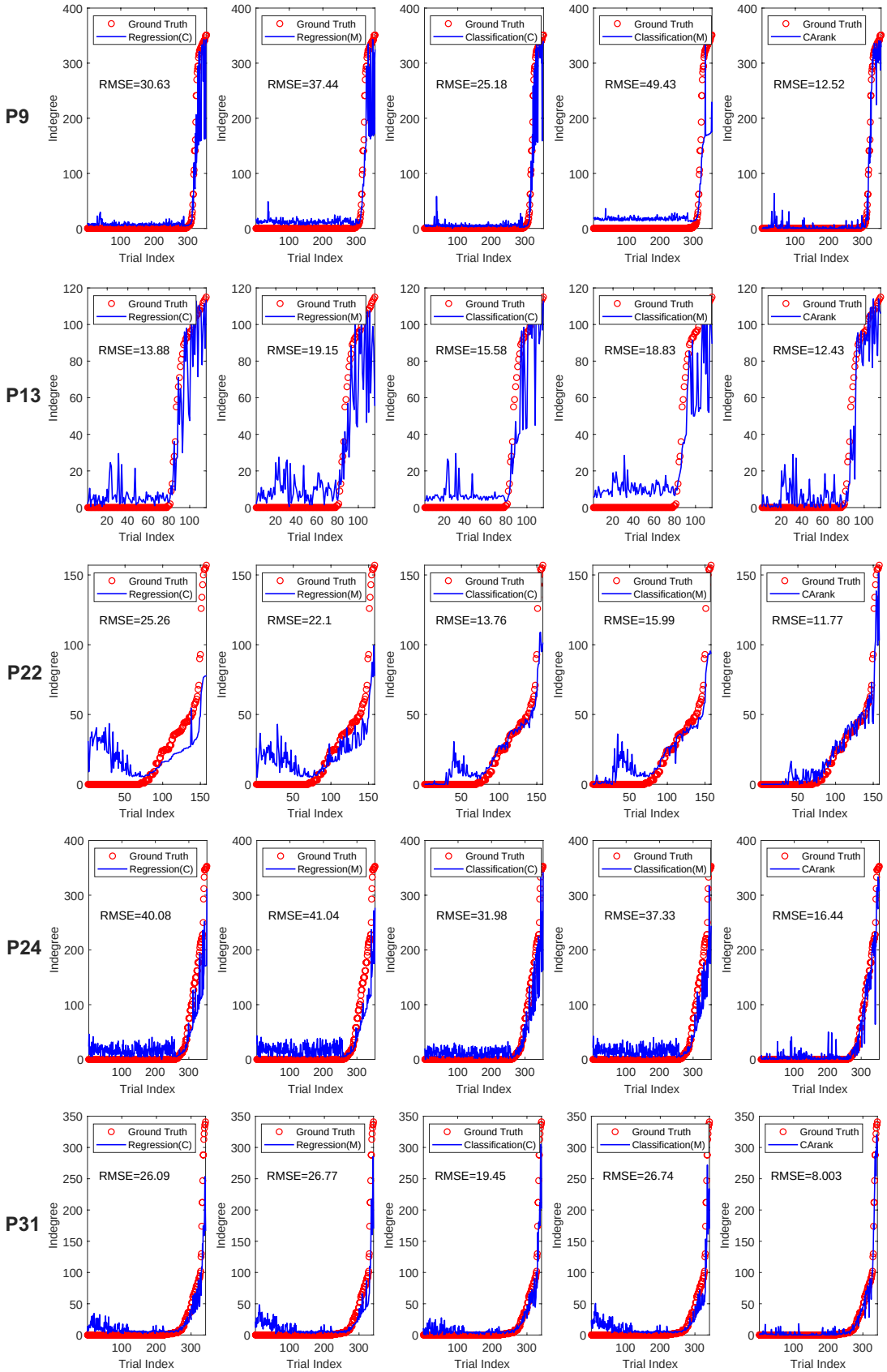

Figure 5: Indegree sequence for CArank and other baselines (closer is better). The root-mean-squared error (RMSE) was also measured according to Eq. (21).

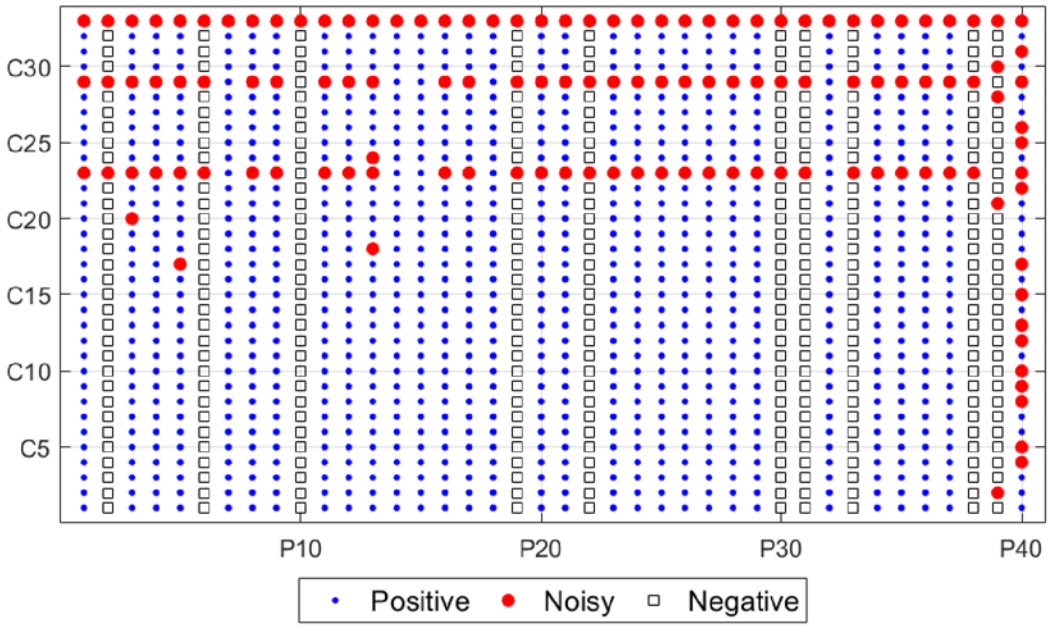

Figure 6: Reliability of different channels for forty participants estimated by CArank. Each column denotes the states of 33 channels for each participant. The channels with estimated reliability $0.15 \leq \pi_n \leq 0.85$ are considered as noisy channels marked in red.

and high indegree trials (large RTs). It means that Regression (C/A) over-estimates the RTs with small values and under-estimates the RTs with large values. It is consistent with our claim that the regression-based model is not suitable for the tasks with the non-smooth response variable (RT). (5) Meanwhile, a simple classification using multi-channel aggregation, i.e., Classification (M), also shows heavy fluctuations, since it lacks an effective mechanism to aggregate the predication from multiple channels. Classification (C) shows better performance but is just as bound to be overfitting, since Classification (C) also could not eliminate the effects of noisy channels during the training process.

5.2 Noisy Channel Detection

We also investigated the reliability of our CArank from the perspective of noisy channel detection. According to our analysis, the parameter π_n in the transition matrix Π_n actually indicates the channel reliability. Hereafter, we leverage π_n as the channel reliability indicator to detect noisy channels. Fig. 6 lists the noisy channels (marked in red) detected with $0.15 \leq \pi_n \leq 0.85$, $\forall n = 1, 2, \ldots, N$.

Fig. 6 shows that: (a) the noisy channels universally exist among the EEG signals. At least one noisy channel is detected for each participant. For example, the 33-th channel is recognized as the noisy channel by CArank for almost all participants. It is reasonable since the 33-th channel is generally acknowledged as the non-relevant channel to any tasks (Lin et al., 2014); (b) For each participant, most channels are reliable, which ensures we can always find enough support to training our CArank; and (c) The detected noisy channels varies from participant to participant, and do not possess the

transitivity property between participants. Because the noise can arise due to (i) intrinsic non-informative EEG channel, e.g. the 33-th channel (for all participants); (ii) channels for lateral mastoid references, e.g. the 23-th and 29-th channel (for majority participants) (Chatrian et al., 1985); and (iii) improper experimentation or artifacts (for P13, P39, P40)(Lin et al., 2018).

6 Conclusion

This work proposes a CArank model to assess the state of mental fatigue. The efficacy of CArank model was demonstrated using EEG data collected in a sustained driving task from 40 participants. This model has been further combined with a stochastic-generalized expectation-maximization (SGEM) algorithm to provide an efficient update in the large-scale setting. CArank model utilized a unique methodology with a relaxed alternative, i.e. ordinal classification, to circumvent overfitting to the extreme values of RTs. It has been demonstrated that the overall performance of CArank can be significantly improved with the introduction of the transition matrix, which enables the technique to evaluate the reliability of informative EEG channels while detecting noisy EEG channels. Empirical results show that CArank delivers significant improvements over simple classification and regression methods in terms of global ranking preservation.

In this work, the cooperation mechanism among channels is simplified as a weighted majority voting system, while different trials are viewed independently. We intend to further formulate it with more complex mechanisms, such as the Markov decision process (MDP), to conduct learning and decision making simultaneously. Some previous (Chen et al., 2016, 2015) studied the decision making process among crowd (noisy) workers, which is promising to our setting to investigate the cooperation mechanism among noisy channels. Efforts are underway to apply this approach in future work.

References

Adams, R. J., Appleton, S. L., Taylor, A. W., Gill, T. K., Lang, C., McEvoy, R. D., and Antic, N. A. (2017). Sleep health of australian adults in 2016: results of the 2016 sleep health foundation national survey. *Sleep Health: Journal of the National Sleep Foundation*, 3(1):35–42.

Alharbi, N. (2018). A novel approach for noise removal and distinction of EEG recordings. *Biomedical signal processing and control*, 39:23–33.

Becirovic, E. (2017). On social choice in social networks.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blankertz, B., Tangermann, M., Vidaurre, C., Dickhaus, T., Sannelli, C., Popescu, F., Fazli, S., Danóczy, M., Curio, G., and Müller, K.-R. (2009). Detecting mental states by machine learning techniques: the berlin brain–computer interface. In *Brain-computer interfaces*, pages 113–135. Springer.

Boksem, M. A. and Tops, M. (2008). Mental fatigue: costs and benefits. *Brain research reviews*, 59(1):125–139.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Cappé, O. and Moulines, E. (2009). On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613.

Chatrian, G., Lettich, E., and Nelson, P. (1985). Ten percent electrode system for topographic studies of spontaneous and evoked EEG activities. *American Journal of EEG technology*, 25(2):83–92.

Chen, X., Jiao, K., and Lin, Q. (2016). Bayesian decision process for cost-efficient dynamic ranking via crowdsourcing. *Journal of Machine Learning Research*, 17(217):1–40.

Chen, X., Lin, Q., and Zhou, D. (2015). Statistical decision making for optimal budget allocation in crowd labeling. *Journal of Machine Learning Research*, 16(1):1–46.

Chuang, C.-H., Cao, Z., King, J.-T., Wu, B.-S., Wang, Y.-K., and Lin, C.-T. (2018). Brain electrodynamic and hemodynamic signatures against fatigue during driving. *Frontiers in neuroscience*, 12:181.

Cook, D. B., O'Connor, P. J., Lange, G., and Steffener, J. (2007). Functional neuroimaging correlates of mental fatigue induced by cognition among chronic fatigue syndrome patients and controls. *Neuroimage*, 36(1):108–122.

de Naurois, C. J., Bourdin, C., Stratulat, A., Diaz, E., and Vercher, J.-L. (2017). Detection and prediction of driver drowsiness using artificial neural network models. *Accident Analysis & Prevention*.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Ekman, P. E. and Davidson, R. J. (1994). *The nature of emotion: Fundamental questions*. Oxford University Press.

Fazli, S., Popescu, F., Danóczy, M., Blankertz, B., Müller, K.-R., and Grozea, C. (2009). Subject-independent mental state classification in single trials. *Neural networks*, 22(9):1305–1312.

Franks, D. D. (2019). *Neurosociology: Fundamentals and Current Findings*. Springer.

Gramann, K., Müller, H., Schönebeck, B., and Debus, G. (2006). The neural basis of ego-and allocentric reference frames in spatial navigation: Evidence from spatio-temporal coupled current density reconstruction. *Brain research*, 1118(1):116–129.

Granzow, B. (2017). A matlab implementation of L-BFGS-B. https://github.com/bgranzow/L-BFGS-B. Accessed July 4, 2017.

Hajinoroozi, M., Mao, Z., Jung, T.-P., Lin, C.-T., and Huang, Y. (2016). Eeg-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing: Image Communication*, 47:549–555.

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.

Izuma, K. and Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron*, 78(3):563–573.

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.

Ji, Q., Zhu, Z., and Lan, P. (2004). Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE transactions on vehicular technology*, 53(4):1052–1068.

Kaji, H., Iizuka, H., and Sugiyama, M. (2019). ECG-based concentration recognition with multi-task regression. *IEEE Transactions on Biomedical Engineering*, 66(1):101–110.

Kasai, H. (2017). Sgdlibrary: A matlab library for stochastic gradient descent algorithms. *arXiv preprint arXiv:1710.10951*.

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.

Liang, P. and Klein, D. (2009). Online em for unsupervised models. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 611–619. Association for Computational Linguistics.

Lin, C.-T., Chuang, C.-H., Huang, C.-S., Tsai, S.-F., Lu, S.-W., Chen, Y.-H., and Ko, L.-W. (2014). Wireless and wearable eeg system for evaluating driver vigilance. *IEEE Transactions on biomedical circuits and systems*, 8(2):165–176.

Lin, C.-T., Chuang, C.-H., Kerick, S., Mullen, T., Jung, T.-P., Ko, L.-W., Chen, S.-A., King, J.-T., and McDowell, K. (2016). Mind-wandering tends to occur under low perceptual demands during driving. *Scientific reports*, 6:21353.

Lin, C.-T., Huang, C.-S., Yang, W.-Y., Singh, A. K., Chuang, C.-H., and Wang, Y.-K. (2018). Real-time EEG signal enhancement using canonical correlation analysis and gaussian mixture clustering. *Journal of Healthcare Engineering*.

Liu, Y., Lan, Z., Khoo, H. H. G., Li, K. H. H., Sourina, O., and Mueller-Wittig, W. (2018). EEG-based evaluation of mental fatigue using machine learning algorithms. In *2018 International Conference on Cyberworlds (CW)*, pages 276–279. IEEE.

Möckel, T., Beste, C., and Wascher, E. (2015). The effects of time on task in response selection-an erp study of mental fatigue. *Scientific reports*, 5:10113.

Pregibon, D. et al. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724.

Roche, A. (2011). Em algorithm and variants: An informal tutorial. *arXiv preprint arXiv:1105.1476*.

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

Tian, S., Wang, Y., Dong, G., Pei, W., and Chen, H. (2018). Mental fatigue estimation using EEG in a vigilance task and resting states. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1980–1983. IEEE.

Ting, P.-H., Hwang, J.-R., Doong, J.-L., and Jeng, M.-C. (2008). Driver fatigue and highway driving: A simulator study. *Physiology & behavior*, 94(3):448–453.

Wascher, E., Rasch, B., Sänger, J., Hoffmann, S., Schneider, D., Rinkenauer, G., Heuer, H., and Gutberlet, I. (2014). Frontal theta activity reflects distinct aspects of mental fatigue. *Biological psychology*, 96:57–65.

Wei, C.-S., Lin, Y.-P., Wang, Y.-T., Jung, T.-P., Bigdely-Shamlo, N., and Lin, C.-T. (2015). Selective transfer learning for EEG-based drowsiness detection. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 3229–3232. IEEE.

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.

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.

Zarei, R. (2017). *Developing enhanced classification methods for ECG and EEG signals*. PhD thesis, Victoria University.

Zeng, H., Yang, C., Dai, G., Qin, F., Zhang, J., and Kong, W. (2018). Eeg classification of driver mental states by deep learning. *Cognitive neurodynamics*, 12(6):597–606.

Zhang, G. P. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4):451–462.