Examining Four Experimental Paradigms for EEG-Based Sleep Quality Evaluation with Domain Adaptation

Wenrui Mu and Bao-Liang Lu* *Senior Member, IEEE*

Abstract— **With the quick development of dry electrode electroencephalography (EEG) acquisition technology, EEG-based sleep quality evaluation attracts more attention for its objective and quantitative merits. However, there hasn't been a standard experimental paradigm. This situation hinders the development of sleep quality evaluation method and technique. In this paper, we experimentally examine the performance of four typical experimental paradigms for EEG-based sleep quality evaluation and develop a new EEG dataset recorded by dry-electrode headset. To eliminate individual variation caused by subjects, we evaluate the four experimental paradigms using domain adaptation (DA) methods. Experimental results demonstrate that a relaxing paradigm is more effective than other attention concentration paradigms and achieves the average accuracy of 76.01%. Domain Adversarial Neural Network outperforms other DA methods and obtains 18.69% improvement on accuracy compared with transfer component analysis.**

I. INTRODUCTION

Sleep, an indispensable physiological state of human daily life, has been a hot topic in both healthcare and research field for decades. As the study on sleep is advancing, sleep quality has been proven as an intrinsic mental indicator. Besides, in various fields such as aviation, surgery and public transport, subjects need an instant and accurate last-sleep quality evaluation to ensure a vigorous mental state. Therefore, sleep quality evaluation shows not only academic significance, but also practical value.

The existing approaches for sleep quality evaluation can be classified into subjective methods and objective methods. Subjective methods such as Pittsburgh Sleep Quality Index (PSQI) [1] or Epworth Sleep Scale (ESS) [2] judge sleep quality via self-report questionnaires, which cannot meet the accuracy requirement for last-night sleep quality evaluation. Objective methods such as Polysomnography (PSG) [3] require complex and persistent monitoring. For various real-world applications, an instant and wearable detecting system is more desirable and more practical to objectively measure last-night sleep quality. For instance, the last-night sleep quality of high-speed train drivers needs to be measured instantly before every daily routine.

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002501, the National Natural Science Foundation of China under Grants 61673266 and 6197020261, the Fundamental Research Funds for the Central Universities, and the 111 Project.

W. Mu and B. L. Lu are with the Center for Brain-Like Computing and Machine Intelligence, Department of Computer Science and Engineering, the Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, the Brain Science and Technology Research Center, and Qing Yuan Research Institute, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China. *Corresponding author: Bao-Liang Lu (bllu@sjtu.edu.cn).

With considerable progress on dry EEG data acquisition technique, EEG-based sleep quality evaluation has become a promising approach because of its objectivity and rapidity. In recent years, Wang *et al.* evaluated last-night sleep quality from resting EEG signals under 4, 6, 8 hours sleep conditions [4]. Zhang *et al.* modified the acquisition procedure and improve the feasibility [5]. Tong *et al.* demonstrated that domain adaptation, which can exploit shared structure underlying different subjects and transmit the knowledge, is an effective method for tackling domain shift problem caused by individual differences in EEG data [6].

Almost all previous studies used wet-electrode devices for acquiring EEG data. However, the preparation process is laborious and limits portability. Dry-electrode devices are more practical, but individual differences are magnified during acquisition, which severely reduces accuracy and limits generalization ability. In this work, we explore the possibility to eliminate individual differences in dry-electrode devices.

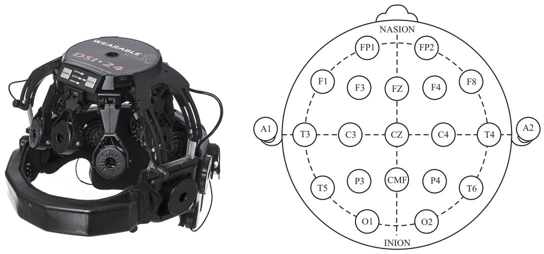
In addition, although research on EEG-based sleep quality evaluation is advancing, there hasn't been an EEG-based sleep-quality-evaluation experimental paradigm that is generally accepted. Previous experiments are non-standard — usually adopted a classical paradigm proposed in psychology, so different paradigm efficiencies still need to be assessed.

In this work, we develop a new EEG dataset with labels of three degrees of sleep deprivation (0h, 4h, 8h) under four paradigms: Closing-Staring (CS) paradigm, Stroop paradigm, Numerical Attention (NA) paradigm and Resting paradigm. Then we apply DA methods to eliminate personal characteristics. Afterwards, we analyze the performance of four experimental paradigms and discuss the relationship between mental state and the performance of experimental paradigms.

II. EXPERIMENTS

In this work, EEG signals were recorded with DSI-24, a dry electrode EEG headset with 21 channels, at a sampling rate of 300 Hz. Fig. 1 shows details on DSI-24. There were a total of 10 subjects (5 males and 5 females, age range: 19-23, mean: 21.50, std: 1.31) that participated in the experiments. All of the subjects kept a regular schedule without any sleep disorder. Before experiments started, they took a tutorial to clarify with the whole process.

According to National Sleep Foundation (NSF) [7], eight hours is the threshold of a sufficient sleep. Therefore, we arranged subjects to participate in three experiments with 8-hour sleep time (normal routine in last 24 hours), 4-hour sleep time (4 hours later than normal sleep time in last 24 hours) and 0-hour sleep time (sleep deprivation in last



(a) DSI-24 (b) 21 sensors positioned according to the 10-20 International System

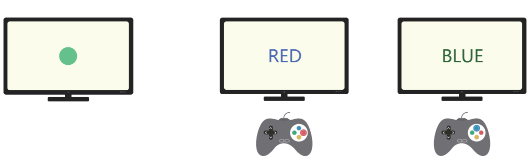
Fig. 1: EEG signal acquisition device

24 hours) respectively. To avoid the influence from prior experiment, there was a one-week interval between sections. For reliability of sleep time, a wearable smartband was provided for each subject to detect his/her real sleep time. For 4-hour sleep time, actual sleep time distributed with a mean of 4.43 h and a standard deviation of 0.60 h, and for 8-hour sleep time, it distributed with a mean of 7.67 h and a standard deviation of 0.50 h.

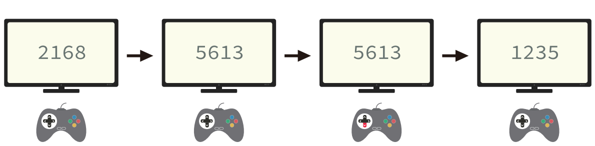
The whole experiment consisted of four paradigms: Closing-Staring, Stroop, Numerical Attention and Resting. Every paradigm was segregated by a short term rest. To reduce body movements, we provided subjects a handle to interact. The keystroke and performance in particular paradigms were recorded. But since the purpose of paradigms was to activate cognition and attention, we didn't employ these data as features. Fig. 3 is two scenes, showing the environment during the experiment.

A. Closing-Staring Paradigm

In this paradigm, subjects need to open eyes and stare at a green dot in the middle of screen when they heard the prompt of start. After 60 seconds, subjects will hear a same prompt and need to close eyes. The whole paradigm consists of three same loops, so this session lasts for 360 (3×2×60) seconds in total. Fig. 2 (a) illustrates the screen sketch.



(a) Closing-Staring (b) Stroop Paradigm



(c) Numerical Attention

Fig. 2: Screen sketches of Closing-Staring, Stroop and Numerical Attention paradigms

B. Stroop Paradigm

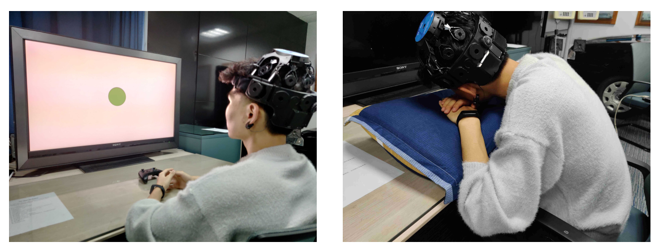
Stroop Paradigm needs subjects to name the ink color of a word with a mismatch between the ink color and the word. It is a classical psychological paradigm to reveal cognitive interference caused by attentional bias [8]. Subjects need to press literal color as soon as possible to proceed next recognition. A new word will appear if subjects press the correct button. Otherwise, the beep will ring to prompt subjects. This session lasts for 180 seconds. Fig. 2 (b) illustrates the screen sketch.

C. Numerical Attention Paradigm

For Numerical Attention paradigm [9], four-digit numbers flash on the screen for 50 ms in every second. Subjects need to press the button if the number showing on the screen is same with the prior one. This session lasts for 180 seconds. Fig. 2 (c) illustrates the screen sketch.

D. Resting Paradigm

Apparently, resting paradigm would cause fatigue and influence the accuracy of other paradigms. Therefore, we arrange it as the end of the whole experiment. We provide a pillow for subjects to ensure a comfortable and relax rest state. For avoiding pressuring to the electrode, we also prescribe subjects sleep posture as Fig. 3 (b): leaning head upon one side hand. This session lasts for 300 seconds.



(a) Closing-Staring (b) Resting

Fig. 3: The scenes of two experimental paradigms.

III. METHOD

A. Data Processing

To enhance the validity of data, we preprocessed the raw EEG signals for eliminating fluctuations induced by hardware. Firstly, EEG signals were processed with baseline correction, and then filtered noise with 1-49 Hz band-pass using Curry 7.

Differential entropy (DE) feature, which can reflect energy change of EEG, is proven to be a superior effective feature to detect fatigue and evaluate sleep quality compared with the conventional power spectral density (PSD) features [10]. Since EEG signals can be approximated as a random variable which obeys the Guassian distribution $N\left(\mu, \sigma^2\right)$, the DE feature can be simplified as the following formulation.

$$h(X) = -\int_{-\infty}^{\infty} f(x)\log(f(x))\mathrm{d}x = \frac{1}{2}\log 2\pi e\sigma^2$$
$$\text{where } f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(x-\mu)^2}{2\sigma^2}$$

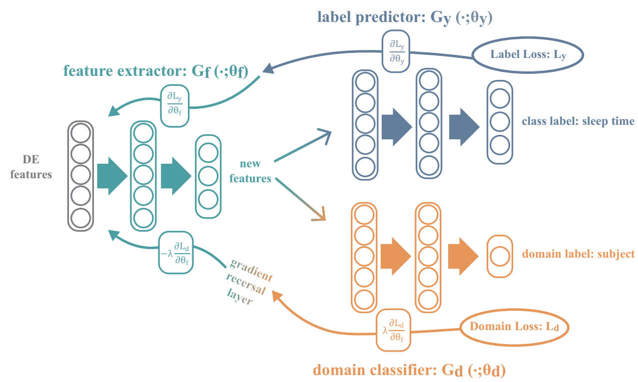


Fig. 4: The network structure of DANN

According to the definition of the DE feature mentioned above, we used Short Term Fourier Transform (STFT) with one-second time window without overlapping to extract the DE features in five frequency bands: delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma (31-49 Hz) [11]. Then we jointed features in five bands as the training input for training models. Thus, EEG signals converted to a 105 dimensional vector (21 channels×5 bands DE features) in every second.

B. Domain Adaptation Methods

In this paper, we explored three DA models to evaluate four experimental paradigms: transfer component analysis (TCA), Multisource Domain Adversarial Network (MDAN) and Domain Adversarial Neural Network (DANN).

1) Transfer Component Analysis (TCA): Transfer component analysis aims to find a low-dimensional feature subspace which can reduce the distance between the marginal distributions and preserve the important properties of source and target domains [12]. Maximum mean discrepancy (MMD) between source domain and target domain is the metric of distribution discrepancies in TCA, so the aimed projection to the subspace is computed by minimizing MMD in a reproducing kernel Hilbert space (RKHS).

2) Multisource Domain Adversarial Network (MDAN): MDAN is a domain adaptation method with adversarial neural networks, under the setting of multiple source domains with labeled instances and one target domain with unlabeled instances [13]. There are three components in MDAN Network architecture: a feature extractor, a domain classifier, and a paradigm learner. Its main idea is to reformulate the generalization bound by a minimax saddle point problem and optimize it via adversarial training.

3) Domain Adversarial Neural Network (DANN): DANN is another domain adaptation with deep architectures [14]. As Fig. 4 depicts, DANN consists of three components: a feature extractor G_f, a label predictor G_y, and a domain classifier G_d. The feature extractor aims to map the input x into a new feature space which can achieve goals in both the label predictor and the domain classifier. The label predictor aims to keep discriminativeness. The domain classifier, existing an adversarial relationship with the feature extractor, aims to keep domain invariance: the classifier cannot provide the correct predictions of the domain, so that the internal representation of the neural network contains no discriminative information about the origin of the input.

IV. RESULTS AND DISCUSSION

A. Model Evaluation

For the consistency of evaluation, same validation process is applied in three models. We aim to classify EEG signals with three degrees of sleep deprivation and adopt a leave-one-subject-out cross validation scheme. TCA is the baseline model. Fig. 5 represents the complete classification accuracies. Table 1 gives details of the best performance model, DANN.

From Fig. 5, we can obtain the following observations: (1) Deep neural networks achieve remarkable improvements in all paradigms. Compared with the baseline model, the average accuracies of MDAN and DANN are 59.80% (13.85% rise) and 64.64% (18.69% rise) respectively. (2) The mean standard deviation also reduces by deep neural networks, (TCA: 11.61%, MDAN: 7.99%, DANN: 8.10%), which means the stability is improved. (3) On some specific subjects(#3, #6), evaluations are inaccurate in all methods. We infer that neurophysiological signals have dramatic individual differences in them, which broadens the gap between other subjects' EEG signals and degrades model performances.

To summarize, classical DA method (TCA) fails to extract common features for dry-electrode EEG signals, but the expected projection can be elicited by multilayer neural networks. The experimental results demonstrate that DANN significantly outperforms other models for sleep quality evaluation.

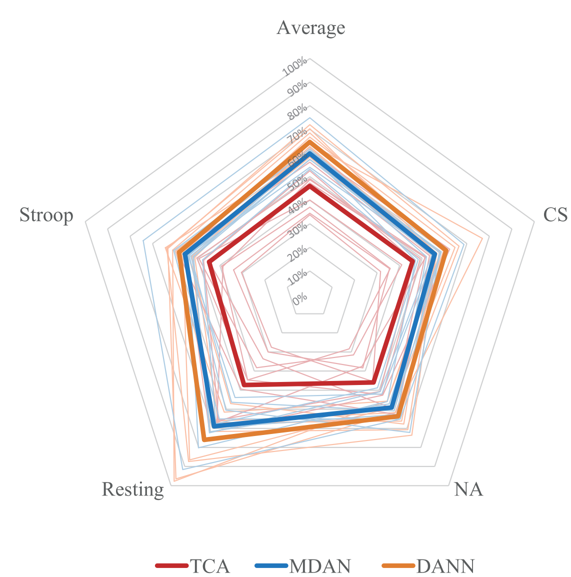


Fig. 5: The classification accuracies for four experimental paradigms and their average accuracies. Thin lines represent accuracy of each subjects, and bold lines correspond to their average accuracies.

TABLE I: Accuracy(%) of four paradigms in DANN

Subject	CS	Stroop	NA	Resting
1	**76.92**	51.15	61.83	66.11
2	57.88	55.82	64.29	68.67
3	54.21	54.10	67.77	57.00
4	56.59	59.55	66.25	67.11
5	56.23	52.38	70.71	71.67
6	58.24	**64.29**	55.74	61.56
7	59.52	61.00	55.82	97.67
8	56.41	63.67	**73.58**	87.33
9	66.48	63.18	61.63	**96.44**
10	64.65	56.54	59.11	86.56
Mean±Std	60.71±6.87	58.17±**4.83**	63.67±6.00	**76.01**±14.71

B. Evaluation of Experimental Paradigms

The accuracy propensities of four experimental paradigms in all models are identical: Resting is superior to other paradigms, and Stroop shows the poorest performance. NA gains a slender advantage with CS.

For three cognition paradigms, there is no obvious accuracy distinction among them. Nonetheless, there is a slight accuracy increase from Stroop to CS to NA in all models. In Stroop paradigm, subjects need to counter their intuitions and receive negative feedback if wrong. On the contrary, in NA paradigm, they can concentrate and make judgments by subconsciousness. Meanwhile, there is no prompt so that subjects can gain a sense of achievement without concerning about mistakes. Besides, resting is the most relaxed paradigm and has the obvious advantage of accuracy. This phenomena implies that with relief of the obstacle during paradigm, mental state can reflect fatigue degree more clearly. The energy in neural patterns also supports that inference.

As beta band can reflect variables of attention and event-related brain potentials [15], we depict the average energy distribution in beta band of EEG DE features for four experimental paradigms in Fig. 6. The differentiation of the topographic neural patterns demonstrates that the neural signatures corresponding to the sleep quality do exist. With the decline of sleep quality, the activation levels of prefrontal area and temporal area decrease in all four experimental paradigms, which is probably caused by the lack of attention.

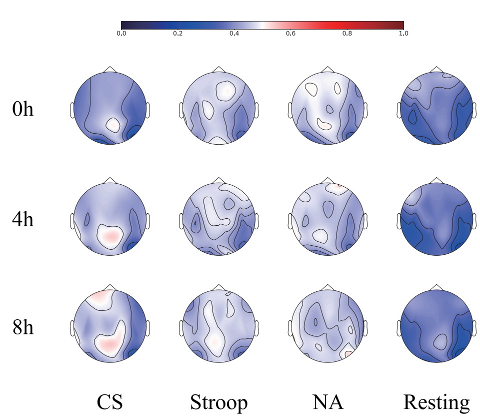


Fig. 6: Topographic neural patterns of four paradigms

Compared with attention concentrated, mental state under the relaxed wakefulness can reflect last-night sleep quality more clearly. However, in practical scenarios, the resting paradigm may be restricted and increase the degree of fatigue. Therefore, a motivational paradigm without obstacle is recommended instead. By this way, subjects can keep conscious and achieve a relatively precise evaluation.

V. CONCLUSIONS

In this paper, we have experimentally examined four experimental paradigms for EEG-based sleep quality evaluation. We have also analyzed the accuracies and neural patterns of four paradigms. The experimental results demonstrate that relaxed state outperforms other mental state in accuracy, and the activation levels of prefrontal area and temporal area positively correlate with sleep quality. Domain adaption with deep neural networks are efficient.

REFERENCES

[1] D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The pittsburgh sleep quality index: a new instrument for psychiatric practice and research," *Psychiatry Research*, vol. 28, no. 2, pp. 193–213, 1989.

[2] M. W. Johns, "A new method for measuring daytime sleepiness: the epworth sleepiness scale," *Sleep*, vol. 14, no. 6, pp. 540–545, 1991.

[3] Y.-K. Teng, L.-C. Chiang, K.-H. Lue, S.-W. Chang, L. Wang, S.-P. Lee, H. Ting, and S.-D. Lee, "Poor sleep quality measured by polysomnography in non-obese asthmatic children with or without moderate to severe obstructive sleep apnea," *Sleep Medicine*, vol. 15, no. 9, pp. 1062–1067, 2014.

[4] L.-L. Wang, W.-L. Zheng, H.-W. Ma, and B.-L. Lu, "Measuring sleep quality from eeg with machine learning approaches," in *IJCNN'16*. IEEE, 2016, pp. 905–912.

[5] X.-Z. Zhang, W.-L. Zheng, and B.-L. Lu, "Eeg-based sleep quality evaluation with deep transfer learning," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 543–552.

[6] J.-J. Tong, Y. Luo, B.-Q. Ma, W.-L. Zheng, B.-L. Lu, X.-Q. Song, and S.-W. Ma, "Sleep quality estimation with adversarial domain adaptation: From laboratory to real scenario," in *IJCNN'18*. IEEE, 2018, pp. 1–8.

[7] M. Hirshkowitz, K. Whiton, S. M. Albert, C. Alessi, O. Bruni, L. DonCarlos, N. Hazen, J. Herman, E. S. Katz, L. Kheirandish-Gozal *et al.*, "National sleep foundation's sleep time duration recommendations: methodology and results summary," *Sleep Health*, vol. 1, no. 1, pp. 40–43, 2015.

[8] J. M. G. Williams, A. Mathews, and C. MacLeod, "The emotional stroop task and psychopathology." *Psychological Bulletin*, vol. 120, no. 1, p. 3, 1996.

[9] J. S. Joseph, M. M. Chun, and K. Nakayama, "Attentional requirements in a 'preattentive'feature search task," *Nature*, vol. 387, no. 6635, p. 805, 1997.

[10] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for eeg-based vigilance estimation," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2013, pp. 6627–6630.

[11] K. Mann, P. Bäcker, and J. Röschke, "Dynamical properties of the sleep eeg in different frequency bands," *International Journal of Neuroscience*, vol. 73, no. 3-4, pp. 161–169, 1993.

[12] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[13] H. Zhao, S. Zhang, G. Wu, and G. J. Gordon, "Multiple source domain adaptation with adversarial learning," 2018.

[14] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[15] T. Egner and J. H. Gruzelier, "Eeg biofeedback of low beta band components: frequency-specific effects on variables of attention and event-related brain potentials," *Clinical Neurophysiology*, vol. 115, no. 1, pp. 131–139, 2004.