Classification of Mental Workload Levels Using EEG and a Hybrid Model of Stacked Denoising AutoEncoder

Zixuan Cao
Engineering Research Center of Optical Instrument and System, Ministry of Education, Shanghai Key Lab of Modern Optical System
University of Shanghai for Science and Technology
Shanghai, China
173790710@st.usst.edu.cn

Zhong Yin*
Engineering Research Center of Optical Instrument and System, Ministry of Education, Shanghai Key Lab of Modern Optical System
University of Shanghai for Science and Technology
Shanghai, China
yinzhong@usst.edu.cn

Jianhua Zhang
OsloMet Artificial Intelligence Lab, Department of Computer Science
Oslo Metropolitan University
Olso, Norway
jianhuaz@oslomet.no

***Abstract*—Human-machine systems are increasingly being used in a variety of scenarios, so it is particularly important to determine the operator's mental workload (MW) levels in time. In this study, EEG was selected as the main indicator for MW levels. After we used Stacked Denoising AutoEncoder and K-nearest Neighbors algorithm to classify MW levels separately, we merged the two models by stacking. Finally, we proved its effectiveness by comparing this new model with the mainstream classifiers and the original two models.**

Keywords—Human-machine, mental workload, Stacked Denoising AutoEncoder, EEG, K-nearest Neighbors algorithm, stacking

I. INTRODUCTION

Human-machine (HM) systems are widely applied in complex control environments [1]. HM systems have the capability to stabilize the machine performance [2] by incorporating the operator's supervision, decision-making [3] and other capabilities. However, as a critical component of HM system, operator's performance is instable [4] because of the intention distraction or mental fatigue [5]. Such issues are major factors that cause many serious human factor accidents. Many reported works indicated the mental workload (MW) was closely related to the brain activity, mental resource utilization, stress and the working memory for information processing during tasks [6]. To this end, MW can be used to assess the operator's cognitive working state aiming at reducing the risk of human performance degradation.

The MW is vulnerable to many factors under human-machine interaction. There is currently no well-established definition of MW [7]. In literature, the MW can be considered as the amount of operator resources taken up by task requirements [8]. When the workload increases and exceeds the general working capacity of the operator, it causes excessive MW [9] and results in inability of information analysis and decision making for the operator [10]. On the other hand, low MW can cause the operator to become inefficient [11]. Therefore, an accurate and effective model is required to be designed for evaluating the MW and to help stabilize the human performance within the proper range. This is especially important for reducing the operation risks and increasing operational safety of the human-machine systems.

1.1 Related works for MW assessment using EEG

There are three main ways to assess MW: 1. Subjective measures 2. task performance measures 3. neurophysiological signals [12]. Subjective measures are also known as subjective rating scales, of which the two most widely practical methods are Subjective Workload Assessment Technique (SWAT) and National Aeronautics and Space Administration-Task Load Index (NASA-TLX) [9]. However, subjective measures lack objectivity and are limited by the low time resolution for data collection [3]. Task performance measures are not suitable for implementation in those task environments where the performance parameters are implicit and cannot be collected directly [13]. Different from the two classical methods, the neurophysiological signals, such as electroencephalogram (EEG), electrocardiogram (ECG), functional Near InfraRed Spectroscopy (fNIRS), event-related potential (ERP) [6], are easy to acquire and to be processed in an online fashion. Among them, EEG has high sensitivity, strong objectivity and easy implementation of task conditions. From the literature, EEG was also closely linked to the alertness and the fatigue of operators [14]-[15] who were engaged in the task environments for nuclear power plants [16] and driving [17].

Extensively reported works applied pattern recognition methods to analyze EEG signals. The pattern classifiers can improve the accuracy on modeling the mapping between the EEG signals and the human cognitive state. Based on a hierarchical Bayesian model, Wang et al. [18] designed an EEG-based workload classifier with correct recognition rate of 80%. In recent studies, Support Vector Machine (SVM) was used by Ke et al. [19] and they built a cross-task MW identification model for n-back tasks. By incorporating the EEG features of power spectrum within 3-15 Hz frequency band, Dornhege et al. [20] identified the tasks under different difficulties by using Linear Discriminant Analysis (LDA). Vuckovic et al. [21] found that the Learning Vector Quantization (LVQ) neural network achieved the best

* Corresponding Author: Zhong Yin, Address: Jungong Road 516, Yangpu district, Shanghai, China. E-mail: yinzhong@usst.edu.cn.

classification performance among three different neural network frameworks.

1.2 Overview of Stacked Denoising AutoEncoder

Although the pattern recognition methods can achieve higher MW recognition accuracy, the classical shallow machine learning methods have the difficulty in mining the hidden information associated with the operator's cognitive state variables. To this end, we attempt to implement the Stacked Denoising AutoEncoder (SDAE) to build a deep-learning based MW classifier. Unlike the general neural network framework, the deep learning model possesses deeper complex network structure [22]. It increases the number of hidden layers in the feedforward path. In particular, SDAE are formed by stacking multiple Denoising Auto-Encoders (DAEs). The DAE is a special Auto-Encoder (AE). It learns higher-level features from the raw data with superimposed noise [23]. The SDAE generates a network structure connected by multiple, consecutive DAEs. By feeding the noisy input data, the layer-by-layer training scheme was performed with the functionality for unsupervised feature de-noising [24].

For neurophysiological signals, the input features are usually accompanied with various types of noise. Therefore, it is very crucial to extract the noise-free EEG features by recovering the original signals. Since SDAE employed noise elimination during learning the weights of deep neural networks, its stability can be achieved by adding different types of noise into the training set [24]. The training procedure of the SDAE consists of two stages, i.e., pre-training and fine-tuning. The unsupervised layer-by-layer pre-training is carried out before fine-tuning the parameters of the entire network [22]. In recent years, SDAE is widely used in face recognition [25], defect detection [24] and other fields, many of which have been developed and commercialized successfully.

II. EXPERIMENTAL DATA

The data used in this experiment has been collected on the Automation-enhanced Cabin Air Management System (Auto-CAMS) in previous studies [26]. Auto-CAMS is a system that fulfills the needs of complex HM missions by simulating an air handling environment in an aircraft cabin [3]. The EEG signals of the experimental participants were collected by the Nihon Kohden biomedical signal processor and displayed in real time by Neu-rofax software.

2.1 Experimental participants and setup

Eight on-campus postgraduate students (male, aged 21-24 years) participated in the experiment as volunteers, by informed consent. Each underwent complete Auto-CAMS operation training prior to the experiment and was anonymized consecutively with a label from S1 to S8.

Auto-CAMS controls air quality through four subsystems: oxygen concentration, carbon dioxide concentration, pressure and temperature. Participant's task is to manually adjust the parameters of the failed subsystem and stabilize it within the target range. The system can change the complexity of the task by manipulating the Number Of Failed Subsystems (NOFS) to meet different MW requirements. The greater number of failed subsystems corresponds to the higher task complexity and MW level. Participants are required to conduct two experiments separately for two consecutive days, each of which is divided into 8 phases and lasts for 100 minutes. The first and eighth stages correspond to the condition of NOFS=0 and last for 5 minutes. The remaining six stages last for 15 minutes and correspond to the conditions of NOFS = 1, 3, 4, 4, 3, 1, respectively. For the participants, the stage of the NOFS=0 condition does not require them to operate, so these two stages can be used to verify whether the MW level is restored. The condition in which NOFS=2 is omitted is to prevent a situation in which excessive experimental time leads to exhaustion of participants and affection of their MW levels. This multi-day scheme with cycle task complexity can capture the MW level information more comprehensively and facilitate the subsequent algorithm research.

2.2 Data preprocessing

The 11 electrodes used to collect the EEG data are placed at the F3, F4, Fz, C3, C4, Cz, P3, P4, Pz, O1 and O2 positions as specified by the 10-20 system. Each phase of the experiment generates a dataset, which contains 450,000 data points corresponding to 11 channels. Each participant conducted two eight-phase experiments, so the corresponding number of datasets was 16.

Each dataset first eliminates eye movement artifacts and high frequency muscle noise using Independent Component Analysis (ICA) and 4th order Butterworth IIR filter (with low pass frequency of 40 Hz) respectively. For sufficient temporal resolution, the data is split in units of two seconds. Then Fast Fourier Transform (FFT) with a frequency resolution of 0.5 Hz is used to calculate the Power Spectral Density (PSD) features of each segment. Through the calculation of the four frequency bands of each channel: theta (4-8Hz), alpha (8-13Hz), beta (14-30Hz) and gamma (31-40Hz), we obtained a total of 44 frequency domain features. And 16 frequency domain features can be derived by calculating the power differences between the left and right hemispheres of the scalp. Finally, with the time-domain features of mean, variance, zero crossing rate, Shannon entropy, spectral entropy, kurtosis and skewness calculated by 11 channels, a total of 137 features were obtained.

The features of the second, fourth, fifth and seventh phases in the experiment were selected for the next study. The feature sets of the four phases on the same day of each person are connected in series to form a feature matrix of 1800×137, wherein the 900 data points corresponding to the second and seventh phases are determined to be low MW level, and the remaining data points are determined as high MW level. Finally, the 16 newly formed total datasets are subjected to standardized processing and adding labels of MW levels.

III. METHODOLOGY

3.1 Stacked Denoising AutoEncoder

The reproducibility of data in SDAE is achieved through its basic composition Autoencoder (AE). AE consists of a three-layer Multi-layer Perceptron. By training the equivalent transformation of the input, a hidden layer's different representation of it is obtained. In this study, the hidden layer dimension adopted by AE is smaller than the data dimension, so the reduced dimensionality expression of input can be obtained through training. The transformation between every two layers in AE is a linear transformation plus a non-linear activation. The input layer to hidden layer mapping is defined by the sigmoid activation function $f(z)=1/(1+e^{-z})$,

$$x_h = f(Wx + \theta) = \frac{1}{1 + e^{-(Wx+\theta)}} \tag{1}$$

Where $x \in R^D$ is defined as the input to AE and $x_h \in R^d$ is the vector of activation values in the hidden layer. The bias vector and weight matrix correspond to $\theta \in R^d$ and $W \in R^{d \times D}$ respectively. And *R* represents the real number, *D* and *d* represent matrix dimensionalities. Then the hidden layer to the output layer mapping by the function expressed as

$$x_o = f(W' x_h + \theta') = \frac{1}{1+e^{-(W' x_h + \theta')}} = \frac{1}{1+e^{-(W' \frac{1}{1+e^{-(Wx+\theta)}} + \theta')}} \tag{2}$$

$x_o \in R^D$ is the output of the AE, and the tied weights are usually used to make $W' = W^T$. The loss of input and output is represented by the square error cost function,

$$L(x, x_o) = \sum_{i=1}^{D} \left| x^{(i)} - x_o^{(i)} \right|^2 \tag{3}$$

The Back Propagation (BP) algorithm can use the above function to determine W, θ and θ'. It corrects weight by calculating the partial derivative of each neuron by the error function until the preset precision or maximum number of learnings is reached. The BP algorithm is a pivotal and efficient algorithm for multi-layer neural network training. The adjusted parameters are defined as W^*, θ^* and θ'^* and can be represented by the following formula, where N represents the number of training data points.

$$\{W^*, \theta^*, \theta'^*\} = \arg\min \frac{1}{N} \sum_{i=1}^{N} L(x, x_o) \tag{4}$$

$$= \arg\min \frac{1}{N} \sum_{i=1}^{N} L(x, f(W' f(Wx + \theta) + \theta'))$$

The input layer of the AE has the same number of neurons as the output layer, and the essential function of the AE is to extract the high-dimensional expression of the input. Therefore, after training an AE, the output of the hidden layer can be used as input to train a new AE, and the second high-dimensional expression of the original input can be obtained. Based on this method, the stack of layers forms a Stacked AutoEncoder (SAE), wherein the output of the ***n***th hidden layer can be expressed as

$$x_h^{(n)} = f(W^{(n)} \ldots f(W^{(2)} f(W^{(1)} x^{(0)} + \theta^{(1)}) + \theta^{(2)}) \ldots + \theta^{(n)}) \tag{5}$$

Finally add a top layer containing two nodes to form a network that can indicate the MW levels, the output of which is defined as

$$y = f(V x_h^{(n)} + v) = s(x^{(0)}) \tag{6}$$

where $y = [1\ 0]^T$ and $y = [0\ 1]^T$ represent the low MW level and high MW level, respectively. V is defined as the output weight matrix and v is the output bias vector.

In order to increase the ability of the SAE to reduce noise, we randomly set the input $x^{(i)}$ through a uniform distribution to 0 to cause interference before training, where the probability is *p* to elicit $\tilde{x}^{(i)}$ and $i = 1, 2, \ldots, D$.

$$\tilde{x}^{(i)} \sim p\left(\tilde{x}_0^{(i)} \middle| x^{(i)} \right) \tag{7}$$

The resulting network is SDAE, and the parameters are fine-tuned by the BP algorithm after layer-by-layer pre-training.

3.2 K-nearest Neighbors algorithm

The K-nearest Neighbors (KNN) algorithm first finds k training samples that are closest to the testing sample, and then predicts the category or value based on their information. The distance between the samples is generally calculated by the Euclidean Distance

$$d(x^{(i)}, x^{(j)}) = \sqrt{\sum_{p=1}^{n} (x_p^{(i)} - x_p^{(j)})^2} \tag{8}$$

Where $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$, $x_j^{(i)}$ represents the ***j***th attribute of sample i.The k samples closest to the testing sample can be obtained by sorting. The categories corresponding to this known sample set are counted, and the most numerous category is the category of the testing sample.

The choice of k value is critical to the accuracy of the model. The enumeration method is generally used: the sample error is calculated separately using different k values, and the k value corresponding to the minimum sample error is selected.

3.3 Stacking

Integrated learning is divided into three methods: bagging, boosting and stacking. Bagging and boosting correspond to parallel and serial calculation methods, respectively. Stacking is different from the above two, and its core idea is k-fold cross-certification. K-fold cross-certification can effectively avoid the over-fitting phenomenon caused by the limited number of samples, generally k=5.

Stacking is mainly divided into four steps. First, the training set is divided into five equal parts according to the sample order, and the first four of them are trained by the base classifier. Then the fifth part of training data and the testing data are predicted using the trained model. After the process is completed, the portion of the training set except the fourth part are selected for retraining and the trained model is used to predict the remaining data set. This process is repeated four times based on the different selected portion of the retraining. Finally, five prediction values can be obtained from the training data and the testing data, respectively. The predictions obtained by the training data are combined as the training data of the secondary classifier. The predictions obtained by the testing data are averaged and used as the testing data of the secondary classifier.

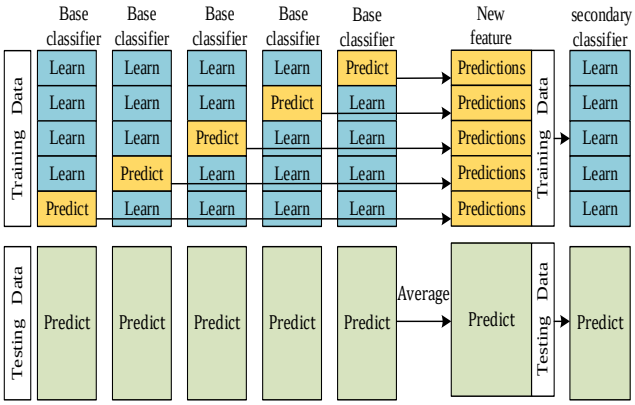

Fig. 2. The schematic of stacking

Stacking is essentially a layered structure that is very similar to the neural network. The effectiveness of stacking mainly comes from feature extraction. Different classifiers express the representation of different features through heterogeneity. Therefore, stacking means that the learning ability is not due to the effect of multi-layer stacking, but due to the learning ability of different classifiers for different features. Stacking achieves superior performance by effectively combining different classifiers. However, with the increase of the number of layers, stacking faces a serious over-fitting risk, so stacking generally only uses two layers.

IV. Results

In this study, eighty percent of the data set is used as the training set (23040×137) and the rest as the testing set (5760×137). The sample sequence of the data set has been scrambled before being split.

We set the number of hidden layers in SDAE to 2. By changing the number of first hidden layer neurons, we can get different training accuracy and testing accuracy, as shown in the figure below. The training accuracy and testing accuracy respectively represent the correct rate of the trained model to predict the training set and the testing set. Where z represents the number of second hidden layer neurons and the size of batch is 40, learning rate is 1. When the number of first hidden layer neurons is 110 and z=20, the network has the best training and testing accuracy, which are 0.9220 and 0.8158.

KNN can be obtained with different test accuracy by adjusting the k value. We choose the value from 1 to 30 for training, and the testing accuracy is best when k=20. The best testing accuracy is 0.7658.

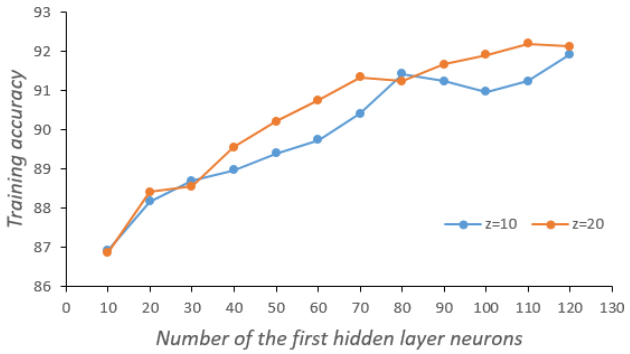

Fig. 3a. The training performance of different hidden layer neurons in SDAE

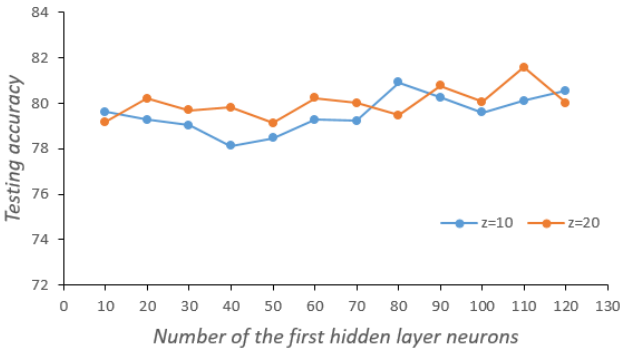

Fig. 3b. The testing performance of different hidden layer neurons in SDAE

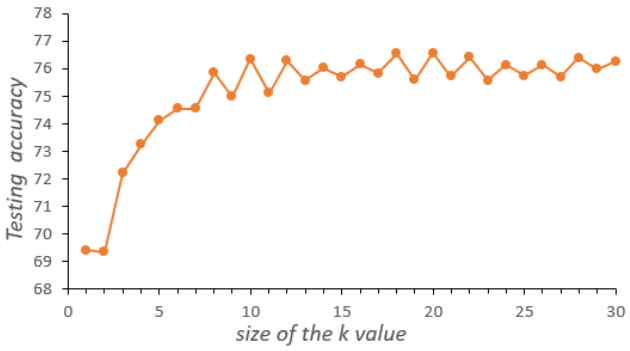

Fig. 3c. The testing performance of different k value in KNN

We select SDAE as the base classifier and KNN as the secondary classifier in stacking. Then we compare the results with SDAE and KNN, and also compare it with some mainstream classification algorithms.

TABLE I. COMPARISON OF SDAE AND KNN

	Training accuracy	**Testing accuracy**	**Training time**
SDAE	0.9220	0.8158	0.4827(s)
KNN	0.7916	0.7658	108.6870(s)
Stacking of SDAE and KNN	0.7882	0.8247	15.1286(s)

TABLE II. COMPARISON OF MAINSTREAM ALGORITHMS

	Training accuracy	**Testing accuracy**
Naive Bayes	0.6978	0.6977
Logistic regression	0.7812	0.7828
Extreme learning machine	0.8256	0.7724
Discriminant analysis classifier	0.7694	0.7677
Stacking of SDAE and KNN	0.7882	0.8247

It can be concluded from the table that some training data is lost by the stacking method, so the training accuracy is reduced, but in exchange for higher testing accuracy. In comparison with mainstream algorithms, the new model also shows relatively good training accuracy and best testing accuracy.

V. CONCLUSIONS

In this study, we proposed a new fusion model by stacking two existing classifiers. Among them, SDAE is used as the base classifier and KNN is used as the secondary classifier. In identifying EEG-based MW levels classification work, the new model demonstrates further improved testing accuracy. Stacking avoids data overfitting through k-fold cross-validation, but this results in fewer training set samples for a single training session. We also found that for classifiers like KNN that require more training time, stacking can be used to effectively reduce its time. However, the improvement in training accuracy is not obvious enough even compared to SDAE and KNN. In future work we will try more models as base classifiers and construct their predictions into multidimensional matrices to provide training to the secondary classifier.

VI. ACKNOWLEDGEMENTS

This work is sponsored by the National Natural Science Foundation of China under Grant No. 61703277 and the Shanghai Sailing Program (17YF1427000).

REFERENCES

[1] L. Habib, M.P. Pacaux-Lemoine, P. Millot, A method for designing levels of automation based on a human-machine cooperation model, IFAC-PapersOnLine. 50 (1) (2017) 1372-1377.

[2] Yue H. Yin, Andrew Y.C. Nee, S.K. Ong, Jian Y. Zhu, Pei H. Gu, Lien J. Chen, Automating design with intelligent human–machine integration, CIRP Annals. 64 (2) (2015) 655-677.

[3] Zhong Yin, Jianhua Zhang, Cross-session classification of mental workload levels using EEG and an adaptive deep learning model, Biomedical Signal Processing and Control. 33 (2017) 30-47.

[4] Christina F. RusnockBrett, J. Borghetti, Workload profiles: A continuous measure of mental workload, International Journal of Industrial Ergonomics. (2016) 1-16.

[5] R. Parasuraman, Y. Jiang, Individual differences in cognition, affect, and performance: behavioral, neuroimaging, and molecular genetic approaches, NeuroImage. 59 (1) (2012) 70–82.

[6] Ming Dong, Ke Yufeng, He Feng, Zhao Xin, Wang Chunhui, Qi Hongzhi, Jiao Xuejun, Zhang Lixin, Chen Shanguang, Psychophysiological measures based studies on mental workload assessment and adaptive automation: Review of the last 40 years and the latest developments, Journal of Electronic Measurement and Instrumentation. 29 (1) (2015) 1-13.

[7] M.S. Young, K.A. Brookhuis, C.D. Wickens, P.A. Hancock, State of science: mental workload in ergonomics, Ergonomics. 58 (1) (2015) 1-17.

[8] F.T. Eggemeier, G.F. Wilson, A.F. Kramer, D.L. Damos, Workload assessment in multi-task environments, D.L. Damos (Ed.). Multiple Task Performance. Taylor & Francis. London (1991) 207-216.

[9] Majid Fallahi, Majid Motamedzade, Rashid Heidarimoghadam, Ali Reza Soltanian, Shinji Miyakee, Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study, Applied Ergonomics. 52 (2016) 95-103.

[10] G.F. Wilson, Operator functional state assessment for adaptive automation implementation, Biomonitoring for Physiological and Cognitive Performance During Military Operations. (2005) 100-104.

[11] Kilseop Ryu, Rohae Myung, Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic, International Journal of Industrial Ergonomics. 35 (11) (2005) 991-1009.

[12] C.R.D. O'Donnell, F.T. Eggemeier, Workload assessment methodology, K. R. Boff, L. Kaufman, J. P. Thomas (Eds.), Handbook of Perception and Human Performance. Wiley. New York (1986) 42–1–42–49.

[13] T.G. Hicks, W.W. Wierwille, Comparison of five mental workload assessment procedures in a moving-base driving simulator, Human Factors: The Journal of the Human Factors and Ergonomics Society. 21 (2) (1979) 129-143.

[14] S. Makeig, Lapses in alertness: coherence of fluctuations in performance and EEG spectrum, Electroencephalogry and Clinical Neurophysiology. 86 (1993) 23-35.

[15] S. Makeig, T. Jung, Changes in alertness are a principal component of variance in the EEG spectrum, Neuroreport. 7 (1995) 213-216.

[16] Moon Kyoung Choi, Seung Min Lee, Jun Su Ha, Poong Hyun Seong, Development of an EEG-based workload measurement method in nuclear power plants, Annals of Nuclear Energy. 111 (2018) 595-607.

[17] Gianluca Borghini, Laura Astolfi, Giovanni Vecchiato, Donatella Mattia, Fabio Babiloni, Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness, Neuroscience & Biobehavioral Reviews. 44 (2014) 58-75.

[18] Ziheng Wang, Ryan M. Hope, Zuoguan Wang, Qiang Ji, Wayne D. Gray, Cross-subject workload classification with a hierarchical Bayes model, NeuroImage. 59 (1) (2012) 64-69.

[19] Yufeng Ke, Hongzhi Qi, Lixin Zhang, Shanguang Chen, Xuejun Jiao, Peng Zhou, Xin Zhao, Baikun Wan, Dong Ming, Towards an effective cross-task mental workload recognition model using electroencephalography based on feature selection and support vector machine regression, International Journal of Psychophysiology. 98 (2) (2015) 157-166.

[20] Guido Dornhege, José del R. Millán, Thilo Hinterberger, Dennis J. McFarland, Klaus-Robert Müller, Improving human performance in a real operating environment through real-time mental workload detection, Toward Brain-Computer Interfacing. (2007) 409-422.

[21] Aleksandra Vuckovic, Vlada Radivojevic, Andrew CN Chen, Dejan Popovic, Automatic recognition of alertness and drowsiness from EEG by an artificial neural network, Medical Engineering & Physics. 24 (5) (2002) 349-360.

[22] Zhihua Zhou, Machine Learning, Tsinghua University Press. (2016) 113-115.

[23] Ting Li, A deep learning method for Braille recognition, Computer and Modernization. 6 (2015) 37-40.

[24] X. Wang, W. He, X. Wang, M. Yao, Y. Qian, Capsule defects detection based on stacked denoising autoencoders, Computer Science. 43 (2) (2016) 64-67.

[25] J. Zhang, Z. Hou, Z. Wu, Y. Chen, W. Li, Research of 3D face recognition algorithm based on deep learning stacked denoising autoencoder theory, International Conference on Communication Software and Networks. (2016) 663-667.

[26] J. Zhang, Z. Yin, R. Wang. Recognition of mental workload levels under complex human–machine collaboration by using physiological features and adaptive support vector machines, IEEE Transactions on Human-Machine Systems. 45 (2) (2015) 200-214.