3.6 Engagement Assessment Using EEG signals

Engagement Assessment Using EEG signals

Feng Li, Jiang Li, and Frederic McKenzie
Flixx003@odu.edu;jli@odu.edu;rdmckenz@odu.edu

Guangfan Zhang, Wei Wang, Aaron Pepe, and Roger Xu
Intelligent Automation, Inc.
gzhang@i-a-i.com; jleddo@i-a-i.com; hgxu@i-a-i.com

Tom Schnell, Nick Anderson, Dean Heitkamp
University of Iowa
richeyc@ccad.uiowa.edu; thomas-schnell@uiowa.edu;

Abstract. In this paper, we present methods to analyze and improve an EEG-based engagement assessment approach, consisting of data preprocessing, feature extraction and engagement state classification. During data preprocessing, spikes, baseline drift and saturation caused by recording devices in EEG signals are identified and eliminated, and a wavelet based method is utilized to remove ocular and muscular artifacts in the EEG recordings. In feature extraction, power spectrum densities with 1 Hz bin are calculated as features, and these features are analyzed using the Fisher score and the one way ANOVA method. In the classification step, a committee classifier is trained based on the extracted features to assess engagement status. Finally, experiment results showed that there exist significant differences in the extracted features among different subjects, and we have implemented a feature normalization procedure to mitigate the differences and significantly improved the engagement assessment performance.

1.0 INTRODUCTION

Operator Functional State (OFS) assessment is important in aviation to ensure mission success and improve mission performance. According to North Atlantic Treaty Organization (NATO) [1], OFS is defined as the multidimensional pattern of human psychophysiological condition that mediates performance in relation to physiological and psychological costs. The application of EEG signals for human mental states assessment, such as mental fatigue, operator engagement, and workload, has been widely studied.

In this paper, we present methods to analyze and improve an EEG-based engagement assessment approach. The developed methods were evaluated with EEG signals collected from four pilots. More specifically, the raw EEG recordings were first cleaned up by removing spikes, baseline drift and other artifacts in a data preprocessing step. We then computed 1-Hz bin PSDs from 1 Hz to 40 Hz as features. Those extracted features were subsequently fed into a committee machine classifier to identify if the pilots were engaged. To train the committee classifier, engagement ground truth for each subject was identified using the methods described in the accompanying paper [11].

We also presented feature analysis results using the Fisher score measure and the one-way ANOVA method. Fisher score is a measure describing how good a feature is for differentiating one category from another, i.e., engaged vs. disengaged, while the ANOVA method performs hypothesis tests to verify if features coming from different categories are statistically different. We found that features identified for each subject by Fisher score can effectively discriminate different engagement states for the subject. However, significant differences in EEG signal features among different subjects were found by the ANOVA method.

To mitigate the feature differences among subjects, we performed a normalization procedure on the extracted features for each subject to be tested. We assume that there was a small set of EEG recordings available for the subject, i.e., from a baseline experiment [2]. The means and standard deviations were then computed for features from the small available dataset

and were subsequently used to normalize all remaining data for that subject. Finally, we evaluated the proposed methods for engagement assessment on the data of four subjects using a 5-fold cross validation method.

The remainder of the paper is organized as follows: section 2 presents our methods to assess the pilot's engagement state. Section 3 illustrates the achieved results and section 4 concludes this paper.

2.0 METHODS

2.1 Data collection

In order to study engagement, experiments in a fully equipped Boeing 737 simulator were conducted [11]. Participants involved pilots with commercial/private/ATP (Airline Transport Pilot) licenses. Each experiment simulated a flight from Seattle Tacoma International Airport to Chicago O'Hare International Airport. Video, audio and physiological information, including EEG, ECG and eye tracking were recorded during the whole simulated flight.

2.2 Engagement ground truth finding

Although there was no sensor to provide online engagement ground truth, the ground truth can be assessed by incorporating pilots' self-evaluation and behavioral measurements etc. as we described in the accompanying paper [11]. For example, in the phase of taking off, landing or a simulated failure, a pilot was more likely to be engaged, while in the phase of flight or the pilot was napping, it is likely that he was disengaged. By using the method described in [11], we identified several data segments for the "engaged" and "disengaged" states as listed in Table 1 in this study to validate the proposed methods.

2.3 EEG artifact removal

Although it has been found that EEG signals from certain sensor locations and EEG features in different frequency bands are highly correlated to OFS [1]-[4], EEG recordings are known to be often contaminated by physiological artifacts from various sources, such as eye blinking/movement, heart beating and movement of other muscle groups [5]. Artifacts are often mixed together with brain signals, making interpretation of EEG signals difficult [6]. To perform OFS assessment using EEG recordings, it is critical to exclude EEG artifacts contained in the signals. Many methods, such as Principle Component Analysis (PCA) [7] and Independent Component Analysis (ICA) [8]-[9] have been developed to remove the EEG artifacts. Wavelet-based methods have also received significant attentions [10] for EEG artifact removal.

Table 1: Pilots' engagement states (1: disengaged; 2: engaged)

Subject	Time duration	State	comments
1	19:08 ~ 19:18	2	taking off
	21:08 ~ 21:17	1	flat flight
2	19:52 ~ 20:03	2	taking off
	21:19 ~ 21:29	1	flat flight
3	19:13 ~ 19:23	2	taking off
	21:54 ~ 22:04	1	flat flight
4	20:58 ~ 21:08	2	taking off
	23:25 ~ 23:35	1	flat flight

In this paper, the proposed procedure for artifact removal is shown in Fig. 1. We started with the removal of environmental artifacts by applying a 60-Hz notch filter followed by the removal of the baseline drift by utilizing a high-pass filter with a cutoff frequency of 0.5-Hz. Due to subject movements in the experiments, some EEG sensors failed because of lose connections between the sensors and scalp for a specific time duration. To remove signals from those sensors, we checked the standard deviations of each EEG channels, and those with zero or values much larger than the average standard deviation, computed from other valid channels, were

discarded. According to the criteria used in the ABM's model [2], EEG segments were identified as spikes or baseline drifts if the EEG amplitude changed abruptly (100uv over a short duration of 30 ms in this paper).

Once the environmental and baseline drifts were removed, the remaining EEG signals were segmented to epochs using a 3-second window. Our EEG signals were sampled at the frequency of 200 Hz and each epoch contains 600 data-points. In our study, two consecutive epochs had 2 seconds or 400 data points of overlapping. To remove ocular and muscular artifacts, each epoch was decomposed using a six level's stationary wavelet transformation, yielding a set of wavelet bands: 0-1.56, 1.56-3.13, 3.13-6.25, 6.25-12.5, 12.5-25, 25-50, and 50-100 Hz. For each wavelet band, the mean and standard deviation of the coefficients were calculated. Coefficients in the band were set to its mean if the absolute difference between the coefficient and the mean was larger than 1.5 times of the standard deviation in that band. Finally, the EEG signals were reconstructed from the modified coefficients.

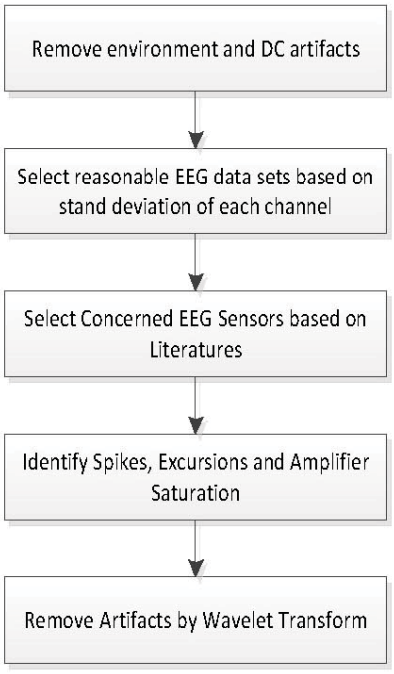

Figure 1: EEG signal preprocessing

2.4 Feature extraction

For each decontaminated epoch, 1-Hz bins of Power Spectral Densities (PSD) from 1 Hz to 40 Hz were calculated as features. In our study, we utilized eight channels of EEG recordings, which yielded 312 (8*39) features for each epoch.

2.5 Feature analysis

We employed two methods, Fisher score and one-way ANOVA, to analyze the extracted features for the subjects. Fisher score was calculated based on the following formula: $\frac{(m_1-m_2)^2}{{\sigma_1}^2+{\sigma_2}^2}$ where $m_{1,2}$ and $\sigma_{1,2}$ are the means and standard deviations of the data points belonging to state 1 or 2 (engagement or disengagement), respectively. The Fisher score is the normalized distance between data points belonging to different states. The larger the fisher score, the more powerful the feature is. By sorting features based on the Fisher score, we can identify the most effective ones that can differentiate the two states. On the other hand, the one-way ANOVA analysis provides an intuitive way to compare data points belonging to difference groups/states. This method can help us verify the effectiveness of those features ranked by the Fisher score for each individual or across subjects.

2.6 Committee machine classifier

We utilized a committee machine as the classification model, which was developed previously [12]. A committee machine is an ensemble classifier consisting of multiple classifiers whose responses are combined as a single response. Fig. 2 shows the main procedures:

- Use the bootstrapping technique to 'disturb' the training data, resulting different sets of training data,
- Train a Multi-layer perceptron (MLP) on each set of training data. The trained model is regarded as a base classification model/committee member.

To , to make each of the committee members diversified, we apply an advanced feature selection algorithm, Piecewise Linear Orthogonal Floating Search (PLOFS) [13],

- Delete the committee members having high biases (accuracy < 50%),
- Utilize the majority vote scheme to fuse decisions from committee members.

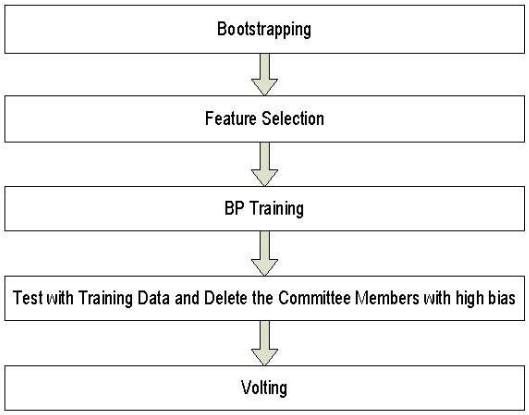

Figure 2: Diagram of the committee machine

2.7 Baseline normalization

It has been observed that there were significant differences in the PSD features among different subjects. To address the individual variation, we assume that a small set of data samples from the test subject are available (i.e., from baseline experiments), and normalize the features based on the small dataset from the test subject. Mean and standard deviation of the subject were computed from his/her available data samples and the remaining data from the subject was normalized by the computed mean and standard deviation.

2.8 Diagram of the proposed approach

Combining the components described above, we present an integrated approach as shown in Fig. 3, where we assume that our subjects are available for the assessment. The purpose of the system is to adapt the model trained on subjects 1, 2 and 3 to subject 4.

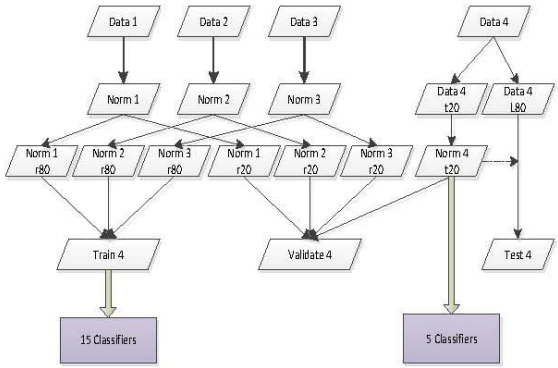

Figure 3: System diagram for the committee machine based engagement assessment

1. Data from subjects 1, 2 and 3 are normalized by their own mean and standard deviations,
2. For the data set from subject 4, top 20% data samples are extracted and normalized by their own mean and standard deviation, forming dataset 'Norm4_t20'. The mean and standard deviation are then utilized to normalize the remaining 80% of the data points,
3. From the three normalized data sets from subjects 1, 2 and 3, 80% of them are extracted and combined as the training dataset (train4); the other 20% are combined with the top 20% of data samples from subject 4 as the validation dataset (validate 4),
4. 15 classifiers are trained based on 'Train4'; 5 other classifiers are trained based on 'Norm4_t20',
5. 20 classifiers are applied to 'Validate 4' and only those classifiers with accuracy over 50% will be kept for testing,
6. The kept classifiers are then utilized to form the final classification results by majority voting.

3.0 RESULTS

3.1 Data preprocessing

Fig. 4 (a) shows a raw segment recording in which baseline drift and other artifacts are presented. Fig. 4 (b) shows the signal after the drift being removed and Fig. 4 (c) shows the "clean" data after the ocular artifact

being eliminated. It is observed that the proposed methods can effectively remove

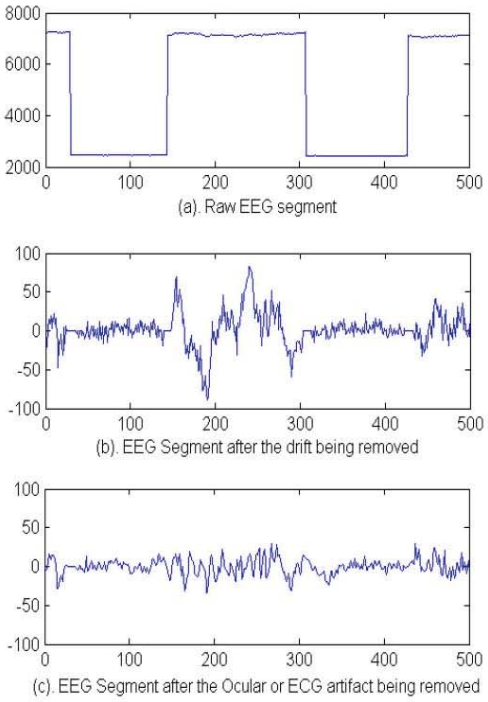

Figure 4: Artifacts removal

common artifacts contained in raw EEG recordings.

3.2 Feature analysis by Fisher score

To find the most valuable features, we firstly calculated the Fisher score for each feature of each subject. Then features were sorted and finally we selected the common features from the top 150 features for the four subjects. In our study, we found 42 common features as shown in Table 2. Four bi-polar sites, Cz-Oz, P7-Oz, P8-Oz, and Pz-Oz were involved and most of the features were in the frequency range of 25~40 Hz.

Table 3 shows top 10 common features and it is clear that the 39 Hz PSD bin from Cz-Oz is the most valuable feature. Figures 5-8 show one common feature among the four subjects, where the feature is large in magnitude for the 'engaged' state and small for the 'disengaged' state. It is also observed that features from different subjects are not in the same scale, implying that the data from different subjects cannot be combined directly and necessary adaptation techniques are mandatory. In our study, we normalized each feature from a subject by its own mean and standard deviation computed using a small dataset from the subject.

Table 2. Distribution of high rank features

	1~4 Hz	5~7 Hz	8~13 Hz	14~24 Hz	25~40 Hz
Cz-Oz	0	0	0	0	8
P7-Oz	4	1	0	9	11
P8-Oz	0	0	0	1	7
Pz-Oz	0	0	0	0	1

Table 3: Fisher scores for the top 10 common features

Subject 1	Subject 2	Subject 3	Subject 4	Sensor	band
1.54	1.61	0.39	4.32	Cz-Oz	39
0.31	2.27	0.57	1.41	P7-Oz	35
0.22	1.78	0.69	1.60	P7-Oz	32
0.29	1.67	0.60	1.59	P7-Oz	29
1.19	1.50	0.35	0.94	Cz-Oz	38
0.30	1.81	0.56	1.29	P7-Oz	34
0.25	1.59	0.65	1.45	P7-Oz	33
0.77	1.90	0.26	0.99	Pz-Oz	33
0.32	1.61	0.52	1.45	P7-Oz	31
0.29	1.60	0.68	1.34	P7-Oz	28

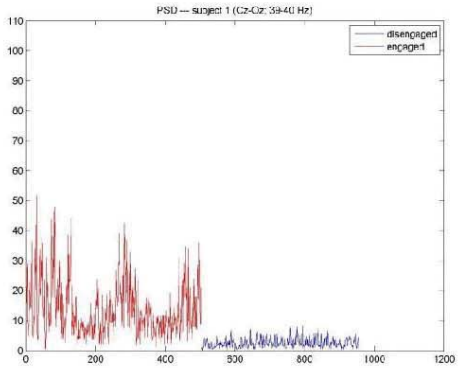

Figure 5: Feature Cz-Oz from 39-40 Hz for subject 1

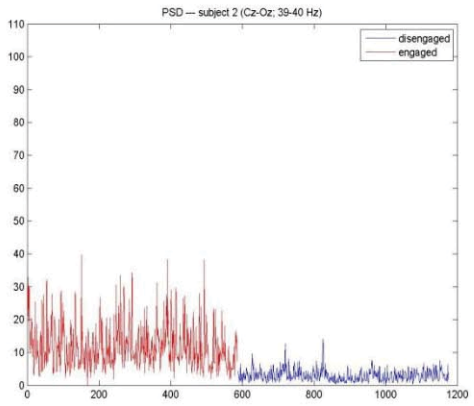

Figure 6: Feature Cz-Oz from 39-40 Hz for subject 2

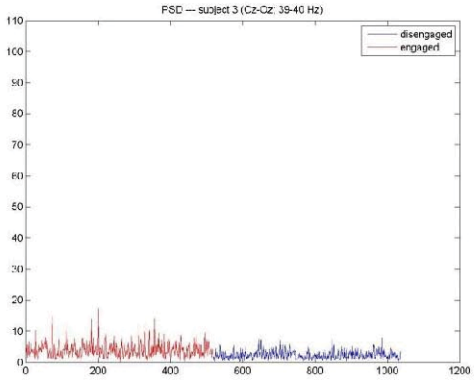

Figure 7: Feature Cz-Oz from 39-40 Hz for subject 3

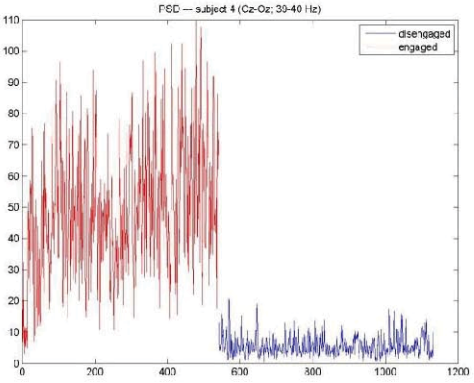

Figure 8: Feature Cz-Oz from 39-40 Hz for subject 4

3.3 Feature analysis by ANOVA

We also analyzed the features using the one-way ANOVA tool and results confirmed that for the feature ranked first by the Fisher score, its PSD for the engaged state is significantly larger than the PSD for the disengaged state. Fig. 9 shows the result of ANOVA analysis for subject 1 and results for other subjects are similar. Fig. 10 shows the ANOVA analysis for the four subjects all together, where the features for each subject were paired. It is clear that features from different subjects are significantly different even if they were all engaged.

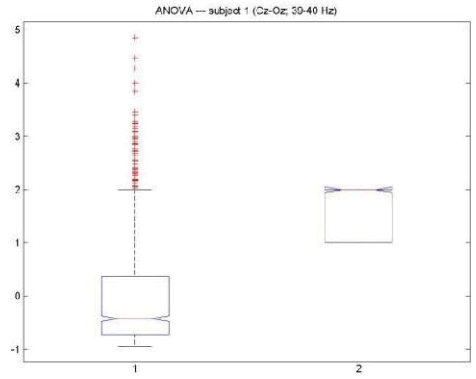

Figure 9: ANOVA analysis for subject 1

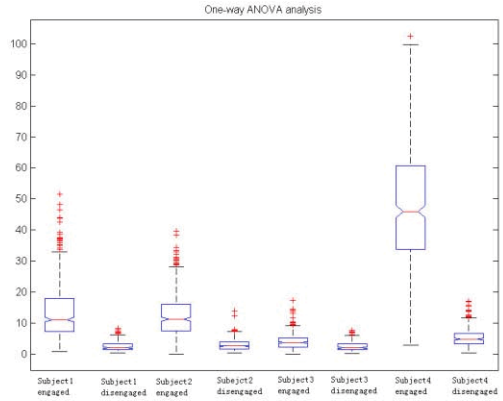

Figure 10: ANOVA analysis for four subjects

3.4 Classification results

We studied two scenarios of the individual assessment model, in which training and testing of the committee machine was performed for each subject independently and results are shown in Table 4. In scenario 1, we used a 5-fold cross-validation scheme to evaluate the engagement assessment model and the average accuracy is 98.55%. In scenario 2, we trained the model by the top 20% data samples from each subject and tested the model on the remaining data. The average accuracy is 94.01%. Both of them prove the effectiveness of the extracted features.

In addition, two scenarios had been studied for evaluating the proposed system if data samples from a subject are not available or very limited. For each of the four subjects, we first trained an average model by using the data from the other three subjects. In scenario 1, we assumed that data samples from the subject to be tested were not available and we normalized the testing data by the mean and standard deviation computed from other subjects, producing poor accuracies as shown in Table 5. In scenario 2, we assumed that a small set of data samples from the subject to be tested were available (the top 20% of the testing data) and used their mean and standard deviation to normalize the remaining testing data. It can be observed in Table 5 that the results are significantly improved.

Table 4: Classification accuracy (%) of the individual models

Scenario	subject 1	subject 2	subject 3	subject 4	Average
1	98.76	97.18	98.43	99.82	98.55
2	90.35	95.32	90.49	99.89	94.01

Table 5: Classification accuracy (%) of the average models

Scenario	Subject 1	Subject 2	Subject 3	Subject 4	Average
1	52.66	78.55	53.95	97.26	70.60
2	93.21	91.66	89.11	69.88	85.96

4.0 CONCLUSIONS

In this research, we proposed methods for engagement assessment based on EEG signals. The methods include EEG artifacts removal, feature analysis and ranking, and a feature normalization procedure. Experimental results illustrated that the artifact removal methods eliminated most of the artifacts in the EEG recordings. Feature analysis showed that engagement can be effectively assessed using the identified features but there existed large differences in features from different subjects. Finally, we demonstrated that the feature normalization procedure significantly mitigated feature variations across subjects.

5.0 ACKNOWLEDGMENTS

The OFS assessment model development is being funded by the NASA (Contract No: NNX10CB27C). We thank Dr. Alan T. Pope, Mr. Chad L. Stephens, and Dr. Kara Latorella, for their comments and suggestions as we performed this research.

6.0 REFERENCES

[1]. G. Robert J. Hockey, Robert Hockey, "Operator Functional State: The Assessment and Prediction of Human Performance Degradation in Complex Tasks," *NATO ASI SERIES* vol. 355.

[2]. Chris Berka, Daniel J. Levendowski, Michelle N. Lumicao, Alan Yau, Gene Davis, Vladimir T. Zivkovic, Richard E. Olmstead, Patrice D. Tremoulet, Patrick L. Craven, "EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks," *Aviation, Space and Environmental Medicine*, vol. 78, No. 5, Section II, May 2007.

[3]. Leonard J. Trejo, Kevin Knuth, Raquel Prado, Roman Rosipal, Karla Kubitz, Rebekah Kochavi, Bryan Matthews, Yuzhen Zhang, "EEG-Based Estimation of Mental Fatigue: Convergent Evidence for a Three-State Model," *HCI International 2007 and Augmented Cognition International Conference*, Beijing, China, July 22-27. In D.D. Schmorrow, L.M. Reeves(Eds.):

Augmented Cognition, HCII 2007, LNAI 4565, pp. 201-211, New York: Springer LNCS, 2007.

[4]. Alan T. Pope, Edward H. Bogart, Debbie S. Bartolome, "Biocybernetic system evaluates indices of operator engagement in automated task," *Biological Psychology*, vol. 40, pp. 187 – 195, 1995.

[5]. T. Jung, S. Makeig, C. Humphries, et. al., "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, pp. 163-178, 2000.

[6]. E. Urretarazu, J. Iriarte, M. Alegre, M. Valencia, C. Vireri, and J. Artieda, "Independent component analysis removing artifacts inictal recordings," *Epilepsia*, vol. 45, pp. 1071-1078, 2004.

[7]. P. Berg and M. Scherg, "A multiple source approach to the correction of eye artifacts," *Electroencephalography and clinical Neurophysiology*, vol. 90, no. 3, pp. 229-241, 1994.

[8]. P. Common, "Independent Component Analysis, A new concept?," *Signal Processing*, vol. 36, pp. 287-314, 1994.

[9]. A. Delorme, S. Makeig and T. Sejnowski, "Automatic artifact rejection for EEG data using high-order statistics and independent component analysis," *Proceedings of the Third International ICA Conference*, Dec 9-12, San Diego, 2001.

[10]. V. Krishnaveni, S. Jayaraman, L. Anitha and K. Ramadoss, "Removal of ocular artifacts from EEG using adaptive thresholding of wavelet coefficients," *Journal of Neural Engineering*, vol. 3, pp. 338-346, 2006.

[11]. Guangfan Zhang, Wei Wang, Aaron Pepe, Roger Xu, Tom Schnell, Nick Anderson, Dean Heitkamp, Jiang Li, Feng Li and Frederic McKenzie, "A Systematic Approach for Real-time Operator Functional State Assessment," *ModSim 2011*.

[12]. Feng Li, Frederick McKenzie, Jiang Li, Guangfan Zhang, Roger Xu, Carl Richey, Tom Schnell, "Imbalanced Learning for Functional State Assessment," *MODSIM 2010*.

[13]. Jiang Li et al., "Feature Selection Using a Piecewise Linear Network," *IEEE Trans. Neural Network*, vol. 17, no. 5, pp. 1101-1115, 2006.