Using Wireless EEG Signals to Assess Memory Workload in the n-Back Task

Shouyi Wang, *Member, IEEE*, Jacek Gwizdka, and W. Art Chaovalitwongse, *Senior Member, IEEE*

***Abstract*—Assessment of mental workload using physiological measures, especially electroencephalography (EEG) signals, is an active area. Recently, a number of wireless acquisition systems to measure EEG and other physiological signals have become available. Few studies have applied such wireless systems to assess cognitive workload and evaluate their performance. This paper presents an initial step to explore the feasibility of a popular wireless system (Emotiv EPOC headset) to assess memory workload levels in a well-known n-back task. We developed a signal processing and classification framework, which integrated an automatic artifact removal algorithm, a broad spectrum of feature extraction techniques, a personalized feature scaling method, an information-theory-based feature selection approach, and a proximal-support-vector-machine-based classification model. The experimental results show that the wirelessly collected EEG signals can be used to classify different memory workload levels for nine participants. The classification accuracies between the lowest workload level (0-back) and active workload levels (1-, 2-, 3-back) were close to 100%. The best classification accuracy for 1- versus 2-back was 80%, and 1- versus 3-back was 84%. This study indicates that the wireless acquisition system and the advanced data analytics and pattern recognition techniques are promising to achieve real-time monitoring and identification of mental workload levels for humans engaged in a wide variety of cognitive activities in the modern society.**

***Index Terms*—Classification, cognitive workload, electroencephalogram (EEG), feature selection, memory load, wireless neuroimaging systems.**

I. Introduction

MENTAL workload describes the level of mental resources utilized when a person is performing a task [1]. The ability to process information, to react to the surroundings, and to make decisions is critical for people in the modern knowledge society. Mental workload assessment can be useful in monitoring and assisting people at work, as well as in evaluating and designing systems or working environments. Mental workload assessment techniques include subjective measures, performance measures, and physiological measures [2], [3]. Both subjective and performance measures are typically taken at one point after the task is completed and are thus static. Subjective measures may also potentially suffer from bias [4]. Compared with subjective and performance measures, physiological measures can provide a continuous record of workload over time and their measurement does not interfere with primary task performance. Thus, physiological measures may be more useful and suitable to assess human mental workload. Understanding human brain functions and neural mechanisms in relation to performance in everyday activities is an important area in neuroscience and neuroergonomics research [5].

Several neuroimaging techniques are available for investigating brain functions research including functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and functional near-infrared spectroscopy (fNIRS). fMRI provides the best means for localizing neural activity, but it has poor temporal resolution and participants have to lie still in a highly constrained environment instead of being in their normal working environments. fNIRS devices are portable and relatively convenient for long-term monitoring and thus have been applied to assess cognitive workload [6], [7]. However, fNIRS, similar to fMRI, has a relatively poor temporal resolution. Conversely, the temporal resolution of EEG is high and is in the order of milliseconds. This makes EEG an appropriate tool to capture fast and dynamically changing brainwave patterns in complex cognitive tasks. EEG signals have been used to detect changes in mental workload on computer-based tasks [8]–[10]. However, most studies typically use costly EEG systems that are wired and bulky, which limits the mental workload assessment in real-world applications. Developments in brain–computer interfaces targeting real-life applications include wireless EEG acquisition systems that a person can easily wear while performing everyday activities.

The use of wireless acquisition systems to assess mental workload can enable novel applications of mental workload measurement. This development supports exploring the feasibility of wireless acquisition devices in mental workload assessment [11]–[14]. For example, Anderson *et al.* [11] used a wireless EEG device to assess cognitive workload involved in processing information presented visually in data plots. The authors used an Emotiv EPOC EEG headset, which is the same system that we used in this study. They used frontal-centered Gaussian distribution and constant weighting of signal channels and calculated power changes in both alpha and theta frequency bands for trials and baseline epochs. Their results showed significant differences in cognitive load levels. However, the plot interpretation task employed was not well characterized with respect to the expected cognitive demands. Knoll *et al.* [14] also studied the feasibility of using the Emotiv EPOC device in assessing mental workload. The study showed significant

Manuscript received May 13, 2014; revised October 2, 2014, January 17, 2015, and April 28, 2015; accepted August 16, 2015. This paper was recommended by Associate Editor E. J. Bass.

S. Wang is with the Department of Industrial and Manufacturing Systems Engineering, University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: shouyiw@uta.edu).

J. Gwizdka is with the School of Information, University of Texas, Austin, TX 78701 USA (e-mail: ieee-thms-2015@gwizdka.com).

W. A. Chaovalitwongse is with the Departments of Industrial and Systems Engineering and Radiology, University of Washington, Seattle, WA 98105 USA (e-mail: artchao@uw.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2015.2476818

correlation between task difficulty levels and the spectral powers of theta, alpha, beta, and gamma frequency band signals in two frontal channels (F3 and F4). However, the employed task used in their study was not well characterized with respect to the relationship between the task difficulty levels and memory workload.

The n-back task has been widely used to investigate mental workload, in particular working memory load [15]–[19]. Working memory tasks require temporary storage and processing of information items. Specifically, n-back task is a working memory updating task. This kind of task reflects general working memory processes (keeping available a set of representations over short periods of time and retrieving them accurately) and updating working memory processes (substituting old working-memory contents by new ones) [20]. In the n-back task, a subject identifies whether the current stimulus (generally letters or numbers) matches a stimulus presented n trials before the current one (n is usually 1, 2, or 3). The load factor n can be adjusted to set the difficulty level of the task and control working memory load conditions without affecting visual input and frequency and type of motor output [15]. This elegant property has made the n-back task a widely employed tool in investigations of mental workload and cognitive performance under various conditions.

A review of research on n-back tasks and their neural correlates is provided in [21]. In their paper, Owen *et al.*, presented a metaanalysis of 24 studies that employed the n-back task and showed similarities and differences in brain regions involved in working memory. Six frontal and parietal cortical regions were found consistently activated based on the quantitative meta-analysis of brain neurological data. A number of studies have investigated memory load using scalp EEG recordings [9], [22]–[26]. Gevins *et al.* [22] reported that increased memory load was associated with increased theta band power in the frontal midline area. Similar observations were also reported in [23] and [25]. In addition to the theta-band activity changes, most studies also observed alpha band activity changes. Gevins *et al.* [22] reported that alpha signal power decreased with increased working memory load in the parieto-occipital midline areas. The same finding was also reported in [25], where the task used included the n-back task.

Several aforementioned studies that investigated classification of EEG signals distinguished discrete levels of working memory load [22], [24], [26]. Gevins *et al.* [22] achieved an accuracy of higher than 80% in binary classification of data segments associated with moderate load (2-back) versus high-(3-back) or low-(1-back) memory load data segments using a neural network-based classifier. However, that study had a few limitations: 1) a Laplacian spatial enhancement requires accurate per-subject head measurements to filter noise from the signal; 2) a manual inspection step was used to remove data segments with artifacts, which highly depended on expertise in reading EEG signals; and 3) the random-hold-out cross validation might overestimate the classification accuracy. Grimes *et al.* [24] presented a classification framework without such steps. The framework included EEG feature extraction using signal powers from the 4–13 Hz band in 1-Hz intervals, from 13–31 Hz in 2-Hz intervals, and from 32–50 Hz in 4-Hz intervals, an information gain-based feature selection method, a Naive Bayes-based classification model, and a block-based cross-validation step. The four level (0-, 1-, 2-, 3-back) classification accuracy was 88% averaged over the eight subjects. Although high classification accuracies were achieved, the model was trained on each individual subject separately. The cross-subject classification performance was reported poor due to considerable individual differences in EEG features. Brouwer *et al.* [26] used signal spectral power and event-related potentials (ERP) as EEG features, and a support vector machine (SVM) classifier was trained to predict memory load. The reported accuracies were around 80%, 75%, and 65% for 0- versus 2-back, 1- versus 2-back, and 0- versus 1-back condition, respectively. Their study also applied classification to each individual participant as memory load may affect EEG differently between individuals. However, the employed ERP feature can be impractical as it requires averaging across many trials and the stimuli and their timing may not be available in many real-life work situations.

The ability of EEG to differentiate between the n-back task levels was established in prior work with wired and clinical-grade EEG systems (e.g., [24], [26], [27]). Our work aims to establish the feasibility of using signals collected from a low-cost wireless acquisition system to assess memory workload on the n-back task. We performed a statistical and data mining analysis of wirelessly collected EEG signals for memory workload assessment. To tackle the problem of high interindividual variability in EEG-derived feature values, we used a personalized standardization method to recalculate features of different subjects into the same scale and to remove outliers. We performed feature selection from a large number of signal features to discover the most informative features for mental workload assessment. An information-theory-based feature selection technique was employed to obtain an optimal subset of the features that show strong discriminative power for different memory workload levels [28]. The SVM algorithm was employed to investigate the discriminability of brainwave patterns at different memory workload levels in an N-fold cross-validation procedure. The statistical behavior analysis also provided useful complementary information to demonstrate the usefulness of the wireless EEG signals for mental workload assessment. The findings of this work bring a promising perspective on applying inexpensive wireless acquisition systems to assess mental workload in real-world applications.

The rest of this paper is organized as follows. In Section II, the experimental methods are presented. Section III presents the results for behavioral analysis, power spectral analysis, and classification analysis. Finally, the discussion appears in Section IV.

II. Method

A. Participants

Fourteen participants (six females) were recruited from Rutgers University student body to perform n-back computer task in the lab. Due to incompatibility of the Emotiv headset with two participants' shape of head (e.g., a participant's head was too small to fit the headset), the data could not be collected. In

Fig. 1. Fourteen-channel Emotiv EPOC headset in the experimental setup.

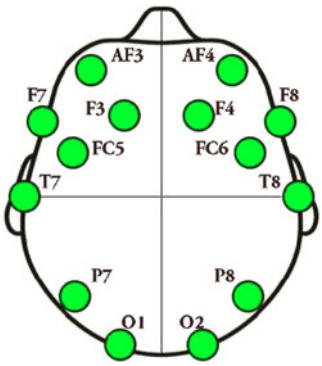

Fig. 2. Location map of the 14 signal electrodes of the Emotiv headset.

three more cases, the data collected were corrupted. Hence, we present data from nine participants (four females).

B. Apparatus

As shown in Fig. 1, we employed a wireless brain signal acquisition system manufactured by Emotiv. This device uses a 14-channel (plus references) high-resolution (2048 samples/s, downsampled internally to 128 samples/s) signal acquisition wireless neuro-headset. Fig. 2 shows the 14 recording positions, which are simplified based on the international 10–20 EEG format. The user task was presented by a program written in python and running on a PC under the Windows XP operating system. The program was based on a version of open source software "Dual N-Back Game" (available at: http://brainworkshop.sourceforge.net). We modified the software so as to support collection of the required data, including timing of trials, accuracy and timing of users responses. The Emotiv system is designed to wirelessly capture EEG signals. However, in practice, the acquired signals may be a mixture of electrical signals due to brain activity (EEG), eye movement (EOG), and other signals related to, for example, facial muscle activity. The latter two types of signals may be captured by the Emotiv system due to several prefrontal locations of electrodes (AF3 and AF4). Throughout the paper, we will use the term EEG signals, to refer to the signals acquired by the Emotiv system keeping in mind that sources of these signals may not be limited to brain activity.

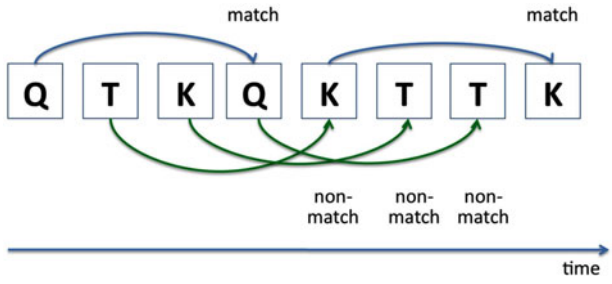

Fig. 3. n-back task principle with letter stimuli.

C. Independent and Dependent Variables

The independent variables included the level of n-back task (0–3) and the two types of stimuli (letters and position). The dependent variables included response accuracy (RA) and reaction time (RT) for the n-back task. The RTs were further divided into RT for correct responses and RT for incorrect responses. The dependent variables also included the power spectral data and other features extracted from EEG signals, and the classification performance measures.

D. User Task

Participants performed the n-back task which is well researched in cognitive psychology and can characterize clearly differentiated workload levels related to a person's working memory [21], [22], [24]. In the n-back task, participants were presented with a series of stimuli (letters or shapes), one at a time. At each stimulus presentation (trial), a participant needed to respond whether or not the current stimulus was the same as the one he/she saw "n" trials ago. Thus, a participant needed to store a sequence of previous "n" stimuli in his/her memory, match it with the current stimulus, and update the memorized sequence with the new stimulus. In this study, each participant performed two types of n-back tasks using two different set of stimuli. The first set consisted of eight letters, randomly chosen from the set of English consonants. Fig. 3 shows an example of the n-back task with letter stimuli. The second set consisted of nine spatial locations contained within a 3×3 presentation square (visually similar to tic-tac-toe game) as shown in Fig. 4.

E. Procedure

The n-back experiment lasted about 1 to 1.5 h. Each participant performed 35 sessions of n-back experiments with four difficulty levels. Each session consisted of 30 3-s trials of n-back tasks at the same difficulty level. Participants responded by pressing a "yes" key, to indicate that the current stimulus was the same as the one "n" trials ago, or "no" key, to indicate that it was different. Each participant first performed 17 practice sessions (with letter and spatial pattern stimuli) and then completed 18 testing sessions: 12 were the letter tasks and six were the spatial tasks. A standard cognitive test of memory span was administered after a participant completed all tasks.

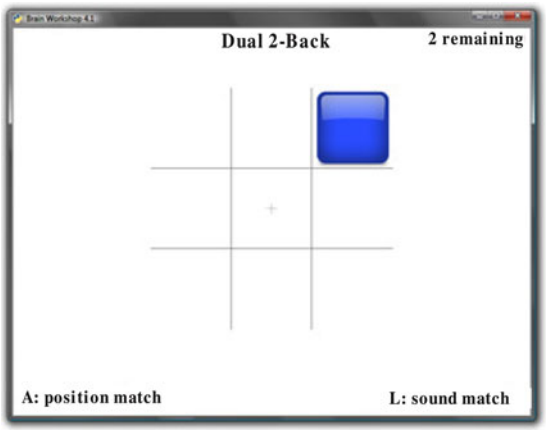

Fig. 4. n-back task with spatial pattern stimuli.

F. Experimental Design and Data Analysis

The experimental design was a 4×2 within-subject design with two factors, n-back levels (0 through 3) and types of stimulus (letters or position). In the data analysis step, we removed trials, in which participants responded by mistakenly pressing a key more than once, as in such cases, we cannot establish which response should be considered.

1) Artifact Removal: Brain signals often contain significant artifacts that lead to major problems in signal analysis, when the activity due to artifacts has a higher amplitude than the one due to neural sources. The common sources of artifacts include eye movements, muscle contractions, and electric devices interference [29]. Independent component analysis (ICA) has been successfully applied for artifacts removal in many studies. The basic idea is to decompose the brain data into independent components, determine the artifacted components using pattern and source localization analysis, and reconstruct the brain signals by excluding those artifacted components. However, linking components to artifact sources (e.g., eye blinking, muscle movements) remains largely user-dependent. We employed an automatic ICA-based algorithm ADJUST [30] for signal artifact removal. ADJUST applies stereotyped artifact-specific spatial and temporal features to identify independent components of artifacts automatically. These artifacts can be removed from the data without affecting the activity of neural sources [30]. The data analysis then uses the cleaned data after artifact removal.

2) Signal Feature Extraction: Four groups of feature extraction techniques were employed to capture signal characteristics that may be relevant to assess memory workload: signal power, statistical, morphological, and time–frequency features. For a data epoch with n channels, we first extracted features from signals at each channel and then concatenated the features of all n channels to construct the feature vector of the data epoch. Let $X = \{x_1, x_2, \ldots, x_m\}$ denote a single-channel signal with m points. The feature extraction of four groups of signal features are described as follows.

a) Signal power features: Adopting the signal features used in [24], we computed signal power for each channel in every nonoverlapping 2-Hz intervals from 4–40 Hz. The 18 power features provide finer signal power spectrum information than the commonly used brain signal frequency bands (theta, alpha, beta, and gamma bands).

b) Statistical features: The mean, variance, skewness, and kurtosis were used to characterize the distribution of signal amplitudes.

c) Morphological features: Three morphological features were extracted to describe morphological characteristics of a single-channel signal as in [31] and [32]. The morphological features are as follows:

i) *Curve length,* also known as "line length" [33], is often defined by sum of distances between successive points. Considering that EEG data may not have the exact same length, we normalized the curve length by taking into account the number of data points. Thus, the curve length employed in this study is calculated by

$$\frac{1}{m-1}\sum_{i=1}^{m-1}|x_{i+1}-x_i|. \tag{1}$$

Since curve length increases as the signal magnitude or frequency increases, it is a measure of amplitude-frequency variations of a signal. It has been used in many brain signal studies, such as epileptic seizure detection [34], and stimulation responses of the brain [35].

ii) *Number of peaks:* It is a measure of the overall frequency of a signal. The number of peaks in a signal X can be calculated by

$$\frac{1}{2}\sum_{i=1}^{m-2}\max\{0, \operatorname{sgn}(x_{i+2}-x_{i+1}) - \operatorname{sgn}(x_{i+1}-x_i)\}. \tag{2}$$

iii) *Average nonlinear energy:* The nonlinear energy [36] is sensitive to spectral changes. Thus, it is useful to capture spectral information of a signal [37]. The average nonlinear energy of the single-channel signal X is computed as

$$\frac{1}{m-2}\sum_{i=2}^{m-1}x_i^2 - x_{i-1}x_{i+1}. \tag{3}$$

The four statistical features and three morphological features were extracted from the ICA-cleaned EEG signals at four well-known frequency bands theta (4–8 Hz), alpha (8–13 Hz), beta (13–25 Hz), and low gamma (25–40 Hz). In addition, we performed feature analysis for concatenated EEG signals under three conditions: correct response, before-keystroke, after-keystroke. The abrupt epoch edge changes do not affect the four statistics of EEG amplitude distributions, while they may slightly affect the morphological feature values. To eliminate the edge artifacts caused by concatenation, we did not include the trial edge points when calculating the three morphological features in formula (1), (2), and (3), respectively.

d) Time–frequency features: Wavelet transform is a powerful tool to perform time–frequency analysis of signals [38]. Among different wavelet families, we employed Daubechies wavelet as it is frequently used in physiological signal analysis due to its orthogonality property and efficient filter implementation [39].

TABLE I
FREQUENCY RANGES AND THE CORRESPONDING BRAIN SIGNAL FREQUENCY BANDS OBTAINED BY THE FOUR-LEVEL DISCRETE WAVELET DECOMPOSITION

Decomposed Level	Frequency Range (Hz)	Approximate Band
D1	32–64	Gamma
D2	16–32	Beta
D3	8–16	Alpha
D4	4–8	Theta
A4	0–4	Delta

A four-level discrete wavelet transform (DWT) decomposition was applied to the collected signals with the sampling rate of 128 Hz. Table I lists the decomposed signals A4, D4, D3, D2, D1, which roughly corresponded to the brain signal frequency bands delta, theta, alpha, beta, and gamma, respectively.

After the four-level DWT decomposition, a set of wavelet coefficients can be obtained for each decomposed signals. To further decrease feature dimensionality, we employed a measure of wavelet coefficients called wavelet entropy (WE), which indicates the degree of multifrequency signal order/disorder in the signals [40]. To obtain WE, the first step is to calculate relative wavelet energy for each decomposition level as follows:

$$p_j = \frac{E_j}{E_{\text{tot}}} = \frac{E_j}{\sum_{j=1}^{n} E_j} \tag{4}$$

where j is the resolution level, and n is the number of decomposed signals ($n = 5$ in this study). E_j is the wavelet power, the sum of squared wavelet coefficients, of decomposed signal j. The relative wavelet energy p_j can be considered as the power density of the decomposed signal level j. Similar to Shannon entropy [41] for analyzing and comparing probability distributions, the WE is defined by

$$\text{WE} = -\sum_{j=1}^{n} p_j \times ln(p_j). \tag{5}$$

The WE offers a suitable tool for characterizing the order/disorder of signals powers in the five brain signal frequency bands (delta, theta, alpha, beta and gamma) during the n-back task.

3) Personalized Feature Standardization: A challenge for many studies that use EEG signals is high interindividual variability. Correspondingly, signal features can vary significantly across subjects. Therefore, it is often difficult to build robust models to estimate mental workload levels across subjects. In addition, due to various artifacts existing in the collected signals, there are inevitable outliers in the extracted signal features. These outlier feature values can distort model training and deteriorate model generalization performance. To tackle these problems, we applied a personalized feature standardization approach [42] to convert the extracted feature values of the subjects into the same scale.

The upper and lower limits of the distribution of a feature are determined as typically done for generating a box plot of a distribution [43]. The lower limit V_l = max(minimum feature value, lower quartile 1.5 × interquartile range), and the upper limit V_u= min(maximum feature value, upper quartile + 1.5 × interquartile range). The upper and the lower limits define an interval containing most of the extracted feature values of a subject. The feature values outside of the interval are considered to be outliers. The absolute feature values are then normalized with respect to the individual interval defined by the upper and lower limits. Assume the raw feature value is F_{raw}; then, the scaled feature value F_{scaled} is obtained by

$$F_{\text{scaled}} = \frac{F_{\text{raw}} - V_l}{V_u - V_l}. \tag{6}$$

The scaled feature value is a percentage indicating the relative position of the feature value in the feature value range $[V_l, V_u]$. Outliers are mapped to 0 or 1 depending on whether they are smaller than the lower limit or greater than the upper limit, respectively. This way, each feature of a subject was standardized into the range of [0, 1] by a personalized range of feature values. The personalized feature scaling reduces interindividual variability that may be caused by signal drift and baseline changes. It can also eliminate feature outliers associated with artifacts caused by body or muscle movements.

4) Feature Selection: For each signal channel, we extracted 47 features: 18 signal power features from 18 2-Hz frequency intervals, four statistical features, three morphological features of the filtered signal in four frequency bands (4–8, 8–13, 13–25, and 25–40 Hz), and WE. The total number of features of a 14-channel data epoch is $47 \times 14 = 658$. The high-dimensional feature space can make the classification process more complex and less reliable due to feature redundancy. It also imposes a challenge to investigate the relationship between various features and memory load levels. Thus, a feature selection step is necessary to select the most informative features to assess memory workload levels. To achieve a reliable feature selection, we employed the minimum redundancy maximum relevance (mRMR) approach [28]. The basic idea of mRMR is to select the most relevant features with respect to class labels while minimizing redundancy amongst the selected features. The mRMR algorithm uses mutual information as the distance measure to compute feature-to-feature and feature-to-class-label nonlinear similarities. For two features X and Y, $p(X)$ and $p(Y)$ are marginal probability functions, and $p(X, Y)$ is the connected probability distribution while $I(X, Y)$ is the amount of mutual information of a and b:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\, p(y)} \right). \tag{7}$$

The mRMR method aims to minimize redundancy (Rd) while maximizing relevance (Re) among the features. An optimal subset of features can be obtained by minimizing the following objective function:

$$\phi(Rd, Re) = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j) - \frac{1}{|S|} \sum_{i \in S} I(h,j) \tag{8}$$

where S is the set of features, h is the vector of target class labels, and $I(i, j)$ is the mutual information between features i and j.

5) Classification Method: With collected signals from four mental workload levels (0-, 1-, 2-, 3-back), SVM was employed to investigate the separability between different mental workload levels. For two mental workload levels A and B, each sample is represented by a vector of features selected by mRMR. Assume there are n data samples in total, denoted by denoted by $x_1, x_2, \ldots, x_n$. In addition, let l denote the class label with $l = 1$ for workload level A and $l = -1$ for workload level B. An SVM classifier's goal is to find a hyperplane that simultaneously minimizes the empirical classification error and maximizes the margin (model generalization) to separate the samples of the two classes. Since the standard SVM classifiers usually require a large amount of computation time for training, the proximal SVM (PSVM) algorithm [44] was introduced as a fast and robust alternative to the standard SVM formulation. We employed a balanced PSVM model which weighs the classes depending on the number of samples in each class and balances them in the training error term, as the sample size of each memory load level can be very unbalanced.

6) Training and Evaluation: The N-fold cross-validation is an attractive model evaluation method when the sample size is small. It is capable of providing an almost unbiased estimate of the generalization ability of a classification model [45]. For the nine subjects, the total number of data samples (sessions) for 0-, 1-, 2-, and 3-back are 17, 35, 63, and 63, respectively. To explore the separability of different workload levels, we categorized the memory workload levels into ten comparison groups. They are 0- versus 1-, 2-, 3-back, 1- versus 2-, 3-back, 1-, 2- versus 3-back, 0-, 1- versus 2-, 3-back, 0- versus 1-back, 0- versus 2-back, 0- versus 3-back, 1- versus 2-back, 1- versus 3-back, and 2- versus 3-back. For each comparison group, we employed a tenfold cross-validation procedure to train and evaluate the PSVM classifier. In particular, for each comparison of two memory load levels, we divided the data samples into ten nonoverlap subsets. Each time we left one subset out and performed feature selection and PSVM model training using the remaining nine subsets. The samples in the left-out subset were used to evaluate the performance of the trained PSVM classifier. All data samples were tested once after repeating this procedure for all ten subsets. Sensitivity and specificity were used to evaluate binary classification performance of each pair of compared workload levels. Given two workload levels A and B, the sensitivity and specificity can be defined as follows:

$$\text{sensitivity} = \frac{\text{\# correctly classified level A samples}}{\text{Total number of level A samples}} \quad (9)$$

$$\text{specificity} = \frac{\text{\# correctly classified level B samples}}{\text{Total number of level B samples}}. \quad (10)$$

The average of sensitivity and specificity was employed to evaluate the classification performance of the two memory load level pairs in the n-back task.

TABLE II
MEAN REACTION TIMES AND ACCURACY FOR THE THREE LEVELS OF n-BACK TASK

n-Back level	Reaction time (ms)	Response accuracy (%)
1-back	667	94
2-back	854	86
3-back	905	79

III. RESULTS

A. Behavioral Results

Results are presented using a priori significance levels of $\alpha = 0.05$, and for trends, $0.05 < \alpha \leq 0.1$. We first performed 3×2 analysis of variance with n-back and stimulus type as independent factors, and RT and RA as dependent variables. The analysis was performed for n-back levels 1, 2, and 3 because RTs could not be obtained for n-back level 0. There were no significant effects of stimulus type in any of the performed analyses. Average RTs differed significantly between the three n-back levels for all responses ($F(2, 149) = 7.3, p = 0.001$) as well as for only correct responses ($F(2, 149) = 6.93, p = 0.001$). Examining post-hoc tests (Fisher's least significant difference), we found significant differences in the expected direction (i.e., shortest RT for 1-back and longest for 3-back; see Table II) between 1- versus 2-back and 1- versus 3-back. The differences in RTs became larger after we removed variability due to individual participants (for all responses $F(2, 149) = 23.54, p < 0.001$). Posthoc tests showed the differences between 1- versus 2- and 1- versus 3-back ($p < 0.001$), and the difference between 2- and 3-back indicated a trend ($p = 0.072$). The accuracy also differed significantly between the three levels of the n-back task ($F(2, 149) = 19.4, p < 0.001$ and $F(2, 149) = 39.54, p < 0.001$ after variability due to individual participants was removed). Posthoc tests showed significant difference in accuracy in the expected direction (i.e., most accurate for 1-back and least for 3-back; see Table II) between all three task levels (for all $p \leq 0.001$). These results confirmed the expected mental workload differences between the difficulty levels of the n-back task and thus provided a validation of the n-back task.

B. Signal Power Spectral Analysis

We computed power spectrum of each 3-s trial of each subject at the four workload levels. For each pair of n-back levels, a two-sample t-test with a significant level of 0.05 was employed to determine whether the signal powers of the 3-s trials of the two n-back levels were significantly different at each channel and at each 1-Hz interval from 0 to 40 Hz. We made 9 subjects $\times$ 4 frequencies $\times$ 14 channels $\times$ 6 comparison pairs, that is, 3024 hypothesis tests. To compensate for a large number of pairwise comparisons, we adopted a false discovery rate (FDR) procedure to control the alpha inflation problem [46], [47]. With a large number of hypothesis tests, the FDR controlling procedure has been shown effective to make a balanced control of both Type-I and Type-II errors in neuroimaging data analysis literature [48]. Taking into account the joint distribution of the p-values

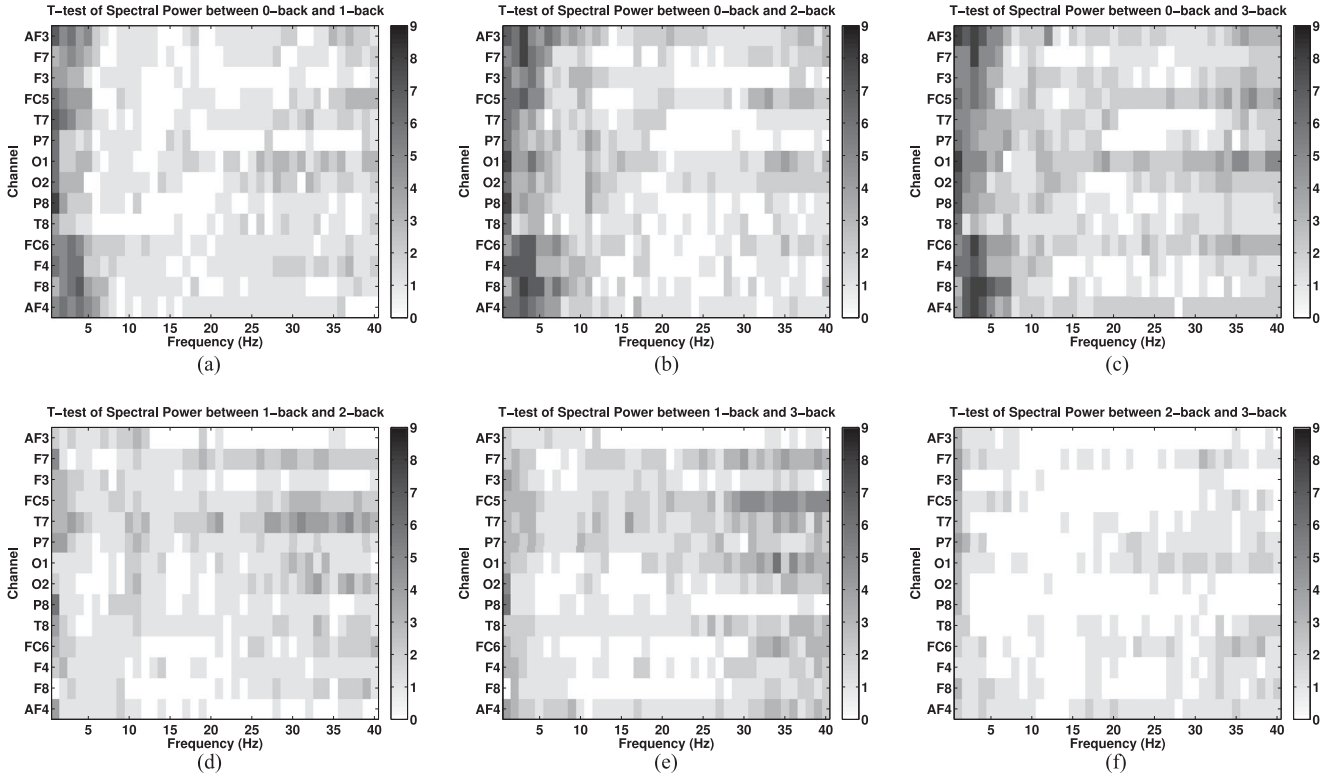

Fig. 5. Number of subjects (indicated by gray level) that show a significant difference in signal power under six comparison conditions [(a) 0- versus 1-back, (b) 0- versus 2-back, (c) 0- versus 3-back, (d) 1- versus 2-back, (e) 1- versus 3-back, (f) 2- versus 3-back) using the FDR corrected p-value threshold at FDR rate of 0.05. For each pair of n-back levels, the t-tests were performed to compare signals powers at each channel and each 1-Hz frequency interval from 0–40 Hz for each subject. Signal power that distinguishes between 0-back and 1-back conditions was present in Delta and low Theta band (0–6 Hz) at frontal channels AF3, AF4, F7, and F8. For 1- versus 2-back and 1- versus 3-back, more subjects showed different signal power at low gamma band (30–40 Hz) at frontal central channels FC5 and F7 and temporal site channel T7, while the signal power differences between 2-back and 3-back were not significant for most subjects.

across channels, frequency bands, and conditions levels, the FDR-corrected critical alpha $= 6.91\times10^{-4}$ given an FDR rate of 0.05. Fig. 5 shows the number of participants that exhibited a significant signal power difference between two workload levels at each channel and each 1-Hz interval. The six subplots represent the results of six pairwise workload level comparisons: 0-back versus 1-back, 0-back versus 2-back, 0-back versus 3-back, 1-back versus 2-back, 1-back versus 2-back, and 2-back versus 3-back. As shown in the subplots (a)–(c), the signal powers that distinguish between 0- and 1-back, 0- and 2-back, and 0- and 3-back are present mainly in 0–6 Hz (approximately delta and low theta bands) at four frontal channels AF3, AF4, F7, and F8. For 0- versus 2-back and 0- versus 3-back, the signal power differences can also be observed in the alpha band (10–13 Hz) at two occipital sites (O1 and O2) and the inferior parietal region (P8 and P7). For 1- versus 2-back and 1- versus 3-back, the most significant areas were presented around 32–40 Hz (low gamma band) at left-fronto-central channel FC5, the left temporal site T7, and the occipital site O1. Finally, for 2- versus 3-back, the significant areas were considerably weaker than other comparison groups. The signal power differences between 2-back and 3-back were not significant for most subjects.

Fig. 6 shows the grand average of the power spectrum over the nine subjects for the four memory workload levels at two frontal locations (FC5, F3) and two back locations (O1, P8). The power spectra at these channels show that alpha power decreases with memory load, while the high beta and low gamma power (20–40 Hz) increases with memory load. In addition, theta powers of 1-, 2-, 3-back are higher than that of 0-back. Although differences were observed in the averaged power spectrum, the signal power spectrum varied greatly across the subjects; this is in accordance with the findings in previous studies [24], [26]. Thus, it would be difficult to accurately differentiate working memory load levels only using power spectral features.

C. Classification Results

We investigated the classifiability of the collected data on ten pairs of memory workload levels based on three datasets: 1) using the entire EEG segments of each session; 2) using the concatenated before-keystroke (before user response) data of each trial to represent each session; and 3) using the concatenated after-keystroke data of each trial to represent each session. For each dataset, we performed automatic artifacts removal, feature extraction, personalized feature standardization, feature selection using mRMR, and classification using PSVM. Using ten top-ranked features selected by mRMR, Table III summarizes the classification performance of the ten memory workload comparison groups based on the tenfold cross-validation. The classification

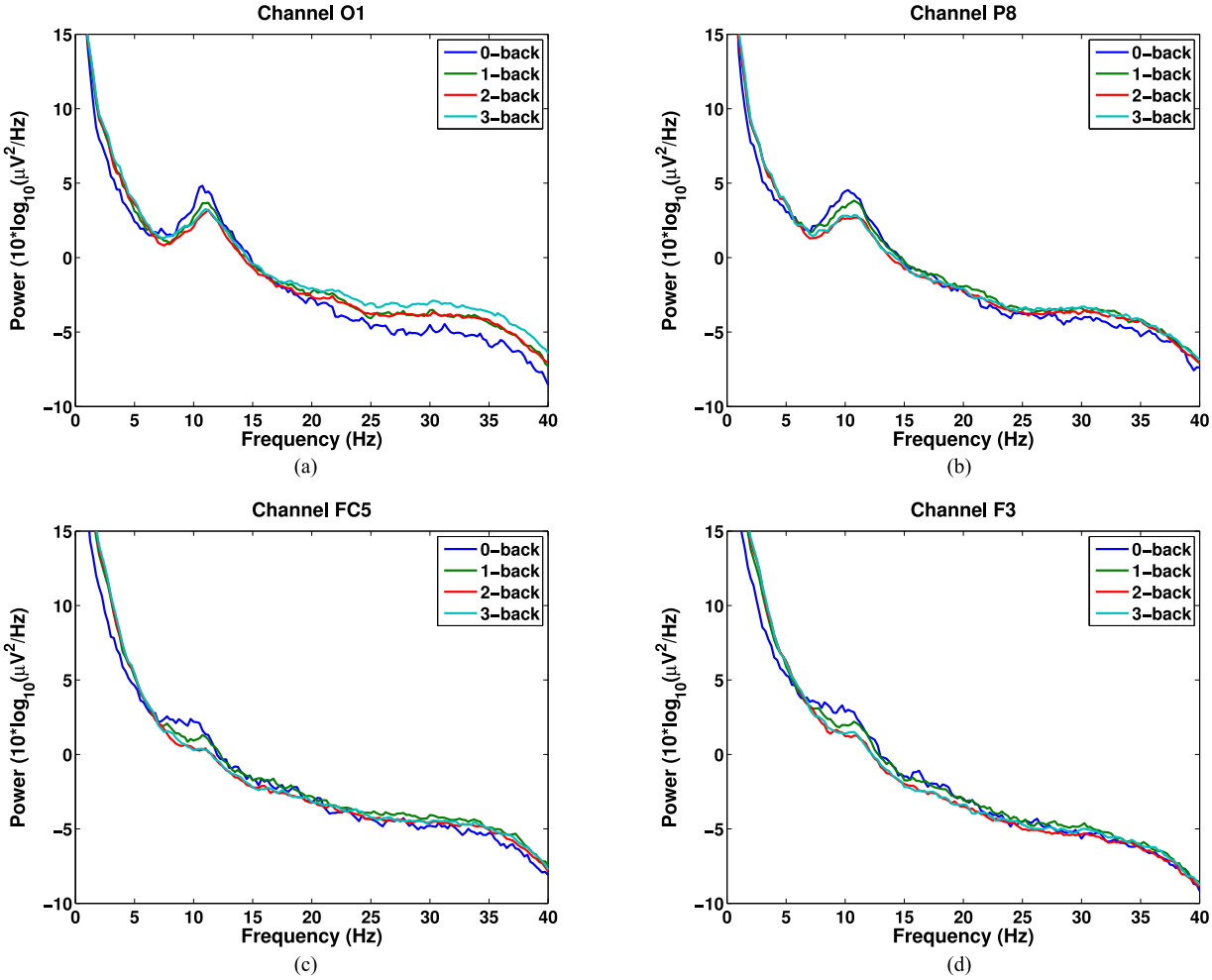

Fig. 6. Grand average of power spectra over the nine subjects for the four memory load levels at two front locations (FC5, F3) and two back locations (P7, P8). The power spectra at these channels show that alpha power decreases with memory load, while the high beta and low gamma power (20–40 Hz) increases with memory load. Also, theta powers of 1-, 2-, 3-back are higher than that of 0-back in the sample distribution. (a) Power spectra of channel O1. (b) Power spectra of channel P8. (c) Power spectra of channel FC5. (d) Power spectra of channel F3.

TABLE III
CLASSIFICATION RESULTS OF WORKING MEMORY LOAD LEVELS USING PSVM AND TEN FEATURES SELECTED BY MRMR ON THREE DATASETS: 1) ENTIRE EEG DATA IN EACH SESSION, 2) CONCATENATED BEFORE-KEYSTROKE DATA IN EACH SESSION, AND 3) CONCATENATED AFTER-KEYSTROKE DATA IN EACH SESSION

	Classification Performance Using All Trials in the n-back Task					
	Entire Session Data		Concatenated Before-Keystroke Data		Concatenated After-Keystroke Data	
Conditions	accuracy	std.	accuracy	std.	accuracy	std.
0-back versus 1-, 2-, 3-back	0.81	0.17	1.00	0.00	1.00	0.01
1-back versus 2-, 3-back	0.59	0.09	0.68	0.12	0.76	0.09
1-, 2-back versus 3-back	0.58	0.09	0.61	0.10	0.69	0.09
0-, 1-back versus 2-, 3-back	0.66	0.12	0.60	0.14	0.82	0.09
0-back versus 1-back	0.63	0.24	1.00	0.00	0.90	0.18
0-back versus 2-back	0.70	0.11	1.00	0.00	1.00	0.00
0-back versus 3-back	0.81	0.16	1.00	0.00	1.00	0.00
1-back versus 2-back	0.60	0.09	0.70	0.11	0.77	0.12
1-back versus 3-back	0.59	0.10	0.72	0.15	0.74	0.08
2-back versus 3-back	0.65	0.11	0.57	0.08	0.57	0.13

TABLE IV
CLASSIFICATION RESULTS OF WORKING MEMORY LOAD LEVELS USING PSVM AND TEN FEATURES SELECTED BY MRMR ON THREE DATASETS THAT ONLY INCLUDE THE TRIALS WITH CORRECT RESPONSES: 1) CONCATENATED EEG TRIALS WITH CORRECT KEYSTROKES, 2) CONCATENATED BEFORE-KEYSTROKE DATA OF CORRECT TRIALS IN EACH SESSION, AND 3) CONCATENATED AFTER-KEYSTROKE DATA OF CORRECT TRIALS IN EACH SESSION

	Classification Performance Using Only Trials with Correct Responses					
	Entire Session Data		Concatenated Before-Keystroke Data		Concatenated After-Keystroke Data	
Conditions	accuracy	std.	accuracy	std.	accuracy	std.
0-back versus 1-, 2-, 3-back	0.80	0.11	1.00	0.00	0.98	0.03
1-back versus 2-, 3-back	0.73	0.16	0.57	0.09	0.84	0.07
1-, 2-back versus 3-back	0.66	0.07	0.53	0.08	0.71	0.10
0-, 1-back versus 2-, 3-back	0.74	0.10	0.61	0.05	0.88	0.04
0-back versus 1-back	0.65	0.19	1.00	0.00	0.78	0.05
0-back versus 2-back	0.71	0.16	0.99	0.03	1.00	0.00
0-back versus 3-back	0.82	0.15	1.00	0.00	1.00	0.00
1-back versus 2-back	0.71	0.09	0.61	0.11	0.80	0.07
1-back versus 3-back	0.71	0.15	0.51	0.11	0.84	0.08
2-back versus 3-back	0.67	0.12	0.60	0.13	0.68	0.07

performances using the before-keystroke and after-keystroke data were mostly better than those using the entire-session data. The classification results using before-keystroke and after-keystroke data both show very strong discrimination between 0-back and 1-, 2-, 3-backs with an accuracy of 100%. The results indicate that the EEG patterns with the lowest memory workload (0-back) were different from the patterns with active memory workload (1, 2, or 3-back). The best discrimination performance between 1- and 2-back was achieved using the after-keystroke data with an accuracy of 77%; the best classification performance between 1- and 3-back was also achieved using the after-keystroke data with an accuracy of 74%; and the best classification accuracy between 2- and 3-back was about 65%. The decrease of the classification accuracies from 1- versus 2-back (77%) and 1- versus 3-back (74%) to 2- versus 3-back (65%) may indicate that the brainwave patterns of the high workload levels (2- and 3-back) are more complicated and thus harder to capture by the current classification model. Another reason might be the patterns of high workload levels have higher inter-individual variability across subjects. Thus, it is more difficult to discriminate the patterns by a single classification model across the subjects. A personalized pattern recognition techniques could be useful to improve the assessment accuracy for high workload levels.

In addition, we performed the classification on the concatenated datasets only using the trials with correct responses. We studied signals from correct trials because we observed that the subjects had varying accuracies in the n-back tasks and exhibited different behaviors when they made incorrect keystroke actions, especially during the 2-back and 3-back tasks. Some subjects might give up or respond randomly for several trials, refresh their memory, and start over again to establish a new memory queue to catch up the pace of the ongoing n-back sequences. Thus, the brainwave patterns in the trials with incorrect responses may be influenced by complicated cognitive activities other than working memory. Table IV summarizes the classification results using the concatenated data from trials with correct responses. Comparing with the results using both correct and incorrect trials, the classification accuracies all increased: the accuracy for 1-, versus 2-, 3-back increased to 84% from 76%; for 1- versus 3-back to 84% from 74%; and for 2- versus 3-back to 68% from 65% with a reduced standard error. Such performance improvements indicate that eliminating the distracted cognitive activities evoked by response errors can be useful to capture brainwave patterns associated with working memory load.

To show the effectiveness of the proposed feature set, Table V presents the classification results using only the power spectral features as in [22], [24], and [26]. The power spectral features used in this study were the EEG signal powers in every nonoverlap 2-Hz intervals from 4 to 40 Hz. The overall classification performances were considerably lower than those using the proposed EEG feature set including statistical features, pattern morphological features, and time–frequency features in addition to the power spectral features. In addition, to show the discrimination between the four memory load levels, Fig. 7 shows the plots of the samples of 0- versus 1-, 1- versus 2-, and 2- versus 3-back using the three highest ranked EEG features selected by the mRMR approach. In particular, feature 1 is the number of peaks in alpha band (8–13 Hz) of EEG channel P7; feature 2 is the WE of EEG channel FC6; and feature 3 is the signal power of the 32–34 Hz interval of EEG channel O1. In the 3-D feature space, the samples of 0-back can be clearly discriminated from 1-back samples, the distributions of 1-back and 2-back samples are also largely separated though having some overlaps, and the 2-back and 3-back samples are heavily overlapped but still show some trends of distinction in the two sample distributions.

IV. DISCUSSION AND CONCLUDING REMARKS

This paper has investigated the feasibility of using wirelessly acquired EEG signals to assess memory workload in a well-controlled n-back task using a wireless EEG system with 14 signal channels. The behavioral measures (increased RTs and decreased RAs) confirmed that different memory workload

TABLE V
CLASSIFICATION RESULTS OF WORKING MEMORY LOAD LEVELS USING PSVM AND TEN FEATURES SELECTED FROM THE POWER SPECTRAL FEATURES ON THE DATASET OF CORRECT TRIALS

	Classification Performance Using Only Power Spectral Features on Data of Correct Trials					
	Entire Session Data		Concatenated Before-Keystroke Data		Concatenated After-Keystroke Data	
Conditions	accuracy	std.	accuracy	std.	accuracy	std.
0-back versus 1-, 2-, 3-back	0.66	0.15	0.71	0.16	0.74	0.15
1-back versus 2-, 3-back	0.66	0.13	0.61	0.07	0.58	0.11
1-, 2-back versus 3-back	0.61	0.08	0.63	0.08	0.66	0.06
0-, 1-back versus 2-, 3-back	0.60	0.12	0.59	0.09	0.62	0.08
0-back versus 1-back	0.70	0.16	0.83	0.14	0.73	0.20
0-back versus 2-back	0.55	0.15	0.71	0.09	0.74	0.10
0-back versus 3-back	0.71	0.11	0.80	0.10	0.72	0.11
1-back versus 2-back	0.69	0.09	0.61	0.09	0.56	0.06
1-back versus 3-back	0.62	0.06	0.63	0.06	0.77	0.11
2-back versus 3-back	0.63	0.04	0.54	0.06	0.68	0.04

The power spectral features are signal powers in every nonoverlapping 2-Hz intervals from 4–40 Hz. The overall classification performances were considerably lower than those using the full proposed EEG feature set including power spectral features, statistical features, pattern morphological features, and time–frequency features.

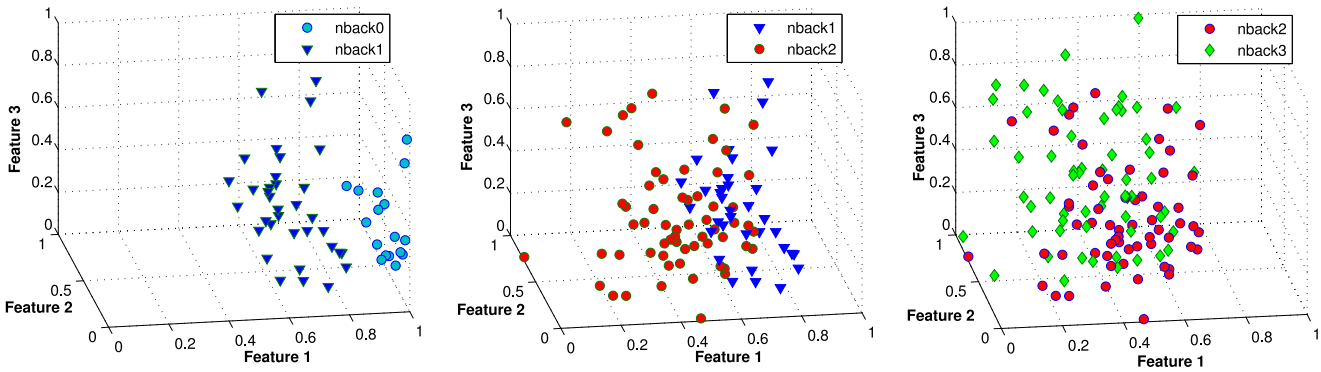

Fig. 7. Plots of the 1-, 2-, and 3-back sessions of the nine subjects using the three highest ranked EEG features selected by the mRMR approach. In particular, feature 1: the number of peaks in alpha band (8–13 Hz) of EEG channel P7; feature 2: the WE of EEG channel FC6; and feature 3: the signal power of channel O1 at 32–34 Hz. In the 3-D feature space, the samples of 0-back can be clearly discriminated from 1-back samples; the distributions of 1-back and 2-back samples are also largely separated although having some overlaps; the 2-back and 3-back samples are heavily overlapped but still show some trends of distinction in sample distribution.

levels were experienced corresponding to different n-back levels. Different memory workload evoked associated EEG signal patterns that made it possible to classify the corresponding memory load levels. The change in signal power in the theta band (4–8 Hz) at frontal channels was found to be significant for distinguishing the lowest workload level (0-back) from the higher workload levels. The change in alpha band (9–13 Hz) and the low gamma band (30–40 Hz) were found to be useful for distinguishing memory workload levels between 1-, 2-, and 3-back levels. In this study, we presented a computational EEG data analysis framework, which integrated recent advances in automated artifact removal, four sets of feature extraction techniques, a personalized feature scaling approach, an information-theory-based feature selection, and a balanced PSVM classification model. The proposed data analysis methodology achieved high classification accuracies for the four memory load levels. The accuracies of two-class classification of the lowest level (0-back) and all the higher memory load levels (1-, 2-, 3-back) were close to 100%, and the accuracies between 1-back and 2- or 3-back were all greater than 80%. Compared with other recent studies, such as [24] and [26], that performed in-subject classification, the classifications presented in this paper were crosstask (letter and position) and cross-subject. In this study, we observed that the averaged power spectra over the nine subjects showed the trends such that with the increased memory load, the alpha power decreased and the theta power increased. The high beta and low gamma (20–40 Hz) power increased in front locations (such as FC5, F3) and back locations (such as P7, P8) with increased memory load. However, similarly to previous studies [24], [26], there were high variations of power spectra between memory load levels among subjects. Some studies employed event related synchronization/de-synchronization (ERS/ERD) as EEG features in the n-back task [49], but only statistical significance testing was performed rather than a classification study. In this study, we also tried ERS/ERD features, but did not get better classification performance. The usefulness of these features may need further investigations in future work.

In this study, we showed that the performance of classification which used the data from correct trials only was considerably

improved compared to classification which used the data that include incorrect trials. The performance improvement may be due to the patterns of memory load being more prominent in the trials with correct responses than in those with incorrect responses. The data from incorrect trials may be contaminated by complicated cognitive activities associated with response errors. We also found that making a distinction between the before-keystroke and after-keystroke data can be useful in further improvement of classification performance of the memory load levels. In particular, using the after-keystroke dataset we achieved the best classification performance. Due to the sequential property of the n-back task, once a subject completes an action (keystroke), he/she will actively prepare for the next cycle by refreshing the current items in working memory, storing a new stimuli (currently observed) into the memory, thus reinforcing updates to working memory. When the next stimuli appears, the subject will compare the stimuli with the ones in the memory, and perform an action accordingly. From this perspective, the working memory activity may be more dominant in the time period between keystroke and a new-stimuli onset than between the stimuli onset and the keystroke; in the before-keystroke period, the brainwave patterns can be more complicated due to the patterns evoked by a new visual stimuli, the stimuli matching in memory, and the action activity. Our experimental results showed that the after-keystroke period is often twice as long or longer than the before-keystroke period in a trial. For the 3 s trails, the averaged before-keystroke periods were 667 ms, 854 ms, and 905 ms long for the 1-back, 2-back, and 3-back trials, respectively. This observation indicates that the after-keystroke period takes a major portion of a trial with working memory as the dominant brain activity. This plausibly explains why the best classification performance was achieved using the after-keystroke dataset.

In summary, the wirelessly collected EEG signals appear to be useful in measuring brain activity changes associated with different working memory levels. The outcome of this study suggests that wireless acquisition systems are promising in monitoring and assessment of mental workload. The wireless acquisition systems offer an excellent opportunity to develop new BCI technologies for mental workload assessment in many real-world applications, as such systems do not require dealing with cables and allow a convenient and portable data collection. Example application domains where such a low-cost workload assessment system would be useful, include interactive information retrieval systems [3] and online learning environments [27]. In the latter domain, Gerjets *et al.* [27] have recently demonstrated a feasibility of cross-task classification developed from data collected in well-controlled working memory tasks (including n-back task) and applied to classification of realistic learning tasks. We plan to use a similar approach and perform cross-task classification between n-back tasks and information retrieval tasks. Although our work is based on the n-back task, the developed method is a systematic computational analysis framework that is purely data-driven from raw data to decisions. Such computational framework can be useful in many other applications that involve pattern recognition or abnormality detection in multivariate EEG signals or brainwave signals which show high interindividual variability. This study can also be further expanded to develop a personalized online monitoring and pattern recognition framework that enables assessment of mental workload in real working environments.

REFERENCES

[1] N. Moray, *Mental Workload: Its Theory and Measurement*. New York, NY, USA: Plenum, 1979.

[2] J. Cegarra and A. Chevalier, "The use of tholos software for combining measures of mental workload: Toward theoretical and methodological improvements," *Behavior Res. Methods*, vol. 40, no. 4, pp. 988–1000, 2008.

[3] J. Gwizdka, "Distribution of cognitive load in web search," *J. Am. Soc. Inform. Sci. Technol.*, vol. 61, no. 11, pp. 2167–2187, 2010.

[4] S. Hart and L. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," in *Human Mental Workload*, P. Hancock and N. Meshkati, Eds. Amsterdam, The Netherlands: North Holland, 1988, pp. 239–250.

[5] R. Parasuraman, "Neuroergonomics brain, cognition, and performance at work," *Current Directions Psychol. Sci.*, vol. 20, no. 3, pp. 181–186, 2011.

[6] J. Harrison, K. Izzetoglu, H. Ayaz, B. Willems, S. Hah, U. Ahlstrom, H. Woo, P. Shewokis, S. Bunce, and B. Onaral, "Cognitive workload and learning assessment during the implementation of a next-generation air traffic control technology using functional near-infrared spectroscopy," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 4, pp. 429–440, Aug. 2014.

[7] L. Hirshfield, K. Chauncey, R. Gulotta, A. Girouard, E. Solovey, R. Jacob, A. Sassaroli, and S. Fantini, "Combining electroencephalograph and functional near infrared spectroscopy to explore users' mental workload," in *Proc. 5th Int. Conf. Found. Augmented Cognition. Neuroergonomics Oper. Neurosci.: Held as Part of HCI Int. 2009*, 2009, pp. 239–247.

[8] P. Antonenko, F. Paas, R. Grabner, and T. van Gog, "Using electroencephalography to measure cognitive load," *Educ. Psychol. Rev.*, vol. 22, no. 4, pp. 425–438, 2010.

[9] A. Gevins and M. Smith, "Neurophysiological measures of cognitive workload during human–computer interaction," *Theor. Issues Ergonom. Sci.*, vol. 4, nos. 1/2, pp. 113–131, 2003.

[10] M. Smith, A. Gevins, H. Brown, A. Karnik, and R. Du, "Monitoring task loading with multivariate EEG measures during complex forms of human-computer interaction," *Human Factors*, vol. 43, no. 3, pp. 366–380, 2001.

[11] E. Anderson, K. Potter, L. Matzen, J. Shepherd, G. Preston, and C. Silva, "A user study of visualization effectiveness using EEG and cognitive load," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 791–800, 2011.

[12] H. Ayaz, B. Onaral, K. Izzetoglu, P. Shewokis, R. McKendrick, and R. Parasuraman, "Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: Empirical examples and a technological development," *Frontiers Human Neurosci.*, vol. 7, pp. 1–13, 2013.

[13] C. Berka, D. Levendowski, M. Cvetinovic, M. Petrovic, G. Davis, M. Lumicao, V. Zivkovic, M. Popovic, and R. Olmstead, "Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset," *Int. J. Human-Comput. Interaction*, vol. 17, no. 2, pp. 151–170, 2004.

[14] A. Knoll, Y. Wang, F. Chen, J. Xu, N. Ruiz, J. Epps, and P. Zarjam, "Measuring cognitive workload with low-cost electroencephalograph," in *Proc. 13th IFIP TC 13 Int. Conf. Human-Comput. Interaction*, 2011, pp. 568–571.

[15] T. S. Braver, J. D. Cohen, L. E. Nystrom, J. Jonides, E. E. Smith, and D. C. Noll, "A parametric study of prefrontal cortex involvement in human working memory," *Neuroimage*, vol. 5, pp. 49–62, 1997.

[16] A. Conway, M. Kane, M. Bunting, D. Hambrick, O. Wilhelm, and R. Engle, "Working memory span tasks: A methodological review and users guide," *Psychonomic Bull. Rev.*, vol. 12, pp. 769–786, 2005.

[17] J. M. Jansma, N. F. Ramsey, R. Coppola, and R. Kahn, "Specific versus nonspecific brain activity in a parametric n-back task," *Neuroimage*, vol. 12, pp. 688–697, 2000.

[18] J. Ragland, B. Turetsky, R. Gur, F. Gunning-Dixon, T. Turner, L. Schroeder, R. Chan, and R. Gur, "Working memory for complex figures: An fmri comparison of letter and fractal n-back tasks," *Neuropsychology*, vol. 16, pp. 370–379, 2002.

[19] S. Ravizza, M. Behrmann, and J. Fiez, "Right parietal contributions to verbal working memory: Spatial or executive?" *Neuropsychologia*, vol. 43, pp. 2057–2067, 2005.

[20] O. Wilhelm, A. Hildebrandt, and K. Oberauer, "What is working memory capacity, and how can we measure it? Frontiers in personality science and individual differences," *Frontiers Psychol.*, vol. 433, pp. 1–22, 2001.
[21] A. Owen, K. McMillan, A. Laird, and E. Bullmore, "N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies," *Human Brain Mapping*, vol. 25, no. 1, pp. 46–59, 2005.
[22] A. Gevins, M. Smith, H. Leong, L. McEvoy, S. Whitfield, R. Du, and G. Rush, "Monitoring working memory load during computer-based tasks with EEG pattern recognition methods," *Human Factors*, vol. 40, no. 1, pp. 79–91, 1998.
[23] A. Gevins and M. Smith, "Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style," *Cerebral Cortex*, vol. 10, no. 9, pp. 829–839, 2000.
[24] D. Grimes, D. Tan, S. Hudson, P. Shenoy, and R. Rao, "Feasibility and pragmatics of classifying working memory load with an electroencephalograph," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2008, pp. 835–844.
[25] S. Lei and M. Roetting, "Influence of task combination on EEG spectrum modulation for driver workload estimation," *Human Factors*, vol. 53, pp. 168–179, 2011.
[26] A. Brouwer, M. Hogervorst, J. Van Erp, T. Heffelaar, P. Zimmerman, and R. Oostenveld, "Estimating workload using EEG spectral power and ERPs in the n-back task," *J. Neural Eng.*, vol. 9, no. 4, p. 045008, 2012.
[27] P. Gerjets, C. Walter, W. Rosenstiel, M. Bogdan, and T. Zander, "Cognitive state monitoring and the design of adaptive instruction in digital environments: Lessons learned from cognitive workload assessment using a passive brain-computer interface approach," *Frontiers Neurosci.*, vol. 8, p. 385, 2014.
[28] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[29] R. Croft and R. Barry, "Removal of ocular artifact from the EEG: A review," *Clinical Neurophysiol.*, vol. 30, pp. 5–19, 2000.
[30] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "Adjust: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, no. 2, pp. 229–240, 2011.
[31] S. Wang, C. Lin, C. Wu, and W. Chaovalitwongse, "Early detection of numerical typing errors using data mining techniques," *IEEE Trans. Syst. Man Cybern. A, Syst. Humans*, vol. 41, no. 6, pp. 1199–1212, Nov. 2011.
[32] S. Wong, G. Baltuch, J. Jaggi, and S. Danish, "Functional localization and visualization of the subthalamic nucleus from microelectrode recordings acquired during DBS surgery with unsupervised machine learning," *J. Neural Eng.*, vol. 6, p. 026006, 2009.
[33] D. Olsen, R. Lesser, J. Harris, R.Webber, and J. Cristion, "Automatic detection of seizures using electroencephalographic signals," U.S. Patent 5 311 876, 1994.
[34] R. Esteller, J. Echauz, T. Cheng, B. Litt, and B. Pless, "Line length: An efficient feature for seizure onset detection," in *Proc. 23rd Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2001, vol. 2, pp. 1707–1710.
[35] R. Esteller, J. Echauz, and T. Tcheng, "Comparison of line length feature before and after brain electrical stimulation in epileptic patients," *Proc. 26th Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2004, pp. 4710–4713.
[36] J. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Proc. Int. Conf. Acoustis, Speech, Signal Process.*, 1990, vol. 1, pp. 381–384.
[37] R. Agarwal and J. Gotman, "Adaptive segmentation of electroencephalographic data using a nonlinear energy operator," in *Proc. IEEE Int. Symp. Circuits Syst.*, 1999, vol. 4, pp. 199–202.
[38] N. Addison, *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine, and Finance*. New York, NY, USA: Taylor & Francis, 2002.
[39] A. Subasi, "Eeg signal classification using wavelet feature extraction and a mixture of expert model," *Expert Syst. Appl.*, vol. 32, no. 4, pp. 1084–1093, 2007.
[40] O. Rosso, S. Blanco, J. Yordanova, V. Kolev, A. Figliola, M. Schürmann, and E. Başar, "Wavelet entropy: A new tool for analysis of short duration brain electrical signals," *J. Neurosci. Methods*, vol. 105, no. 1, pp. 65–75, 2001.
[41] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
[42] G. Buscher, A. Dengel, R. Biedert, and L. V. Elst, "Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond," *ACM Trans. Interactive Intell. Syst.*, vol. 1, no. 2, pp. 9:1–9:30, 2012.
[43] R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, 2nd ed. New York, NY, USA: Academic, 2005.
[44] O. Mangasarian and E. Wild, "Proximal support vector machine classifiers," in *Proc. Knowl. Discovery Data Mining*, 2001, pp. 77–86.
[45] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Roy. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 36, no. 2, pp. 111–147, 1974.
[46] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc. Ser. B*, vol. 57, pp. 289–300, 1995.
[47] Y. Benjamini and Y. Hochberg, "The control of the false discovery rate in multiple testing under dependency," *Ann. Statist.*, vol. 29, no. 4, pp. 1165–1188, 2001.
[48] C. R. Genovese, N. A. Lazar, and T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *Neuroimage*, vol. 15, no. 4, pp. 870–878, 2002.
[49] C. M. Krause, L. Sillanmäki, M. Koivisto, C. Saarela, A. Häggqvist, M. Laine, and H. Hämäläinen, "The effects of memory load on event-related EEG desynchronization and synchronization," *Clin. Neurophysiol.*, vol. 111, no. 11, pp. 2071–2078, 2000.

Shouyi Wang (M'09) received the M.S. degree in systems and control engineering from the Delft University of Technology, Delft, The Netherlands, in 2005, and the Ph.D. degree in industrial and systems engineering from Rutgers University, New Brunswick, NJ, USA, in 2012.

He is currently an Assistant Professor with the Department of Industrial, Manufacturing, and Systems Engineering and the Center on Stochastic Modeling, Optimization, and Statistics, University of Texas at Arlington, Arlington, TX, USA. His research interests include data mining, pattern recognition, statistical learning, big data analytics, multivariate time series monitoring and prediction, and applied operation research.

Jacek Gwizdka received the M.Eng. degree in electrical engineering from the Technical University of Lodz, Lodz, Poland, and the M.A.Sc. and Ph.D. degrees in industrial engineering from the University of Toronto, Toronto, ON, Canada, in 1998 and 2004, respectively.

He is currently an Assistant Professor and Information eXperience Lab Co-Director with the School of Information at University of Texas, Austin,TX, USA. His research interests include the intersection of interactive-information retrieval and human–computer interaction. He applies neuro-physiological methods to the study of human information interaction and is one of the pioneers of Neuro-Information Science.

W. Art Chaovalitwongse (M'05–SM'11) received the M.S. and Ph.D. degrees in industrial and systems engineering from the University of Florida, Gainesville, FL, USA, in 2000 and 2003, respectively.

He is currently a Professor with the Departments of Industrial and Systems Engineering, Radiology (joint), and Bioengineering (adjunct), University of Washington, Seattle (UW), WA, USA. He also serves as an Associate Director of the Integrated Brain Imaging Center, UW Medical Center. His research group conducts basic computational science, applied, and translational research at the interface of engineering, medicine, and other emerging disciplines.