Detection of lapses in responsiveness from the EEG

2011 J. Neural Eng. 8 016003

(http://iopscience.iop.org/1741-2552/8/1/016003)

J. Neural Eng. **8** (2011) 016003 (15pp) doi:10.1088/1741-2560/8/1/016003

Detection of lapses in responsiveness from the EEG

Malik T R Peiris[1,2], Paul R Davidson[1,2,3], Philip J Bones[1,2] and Richard D Jones[1,2,3,4,5,6]

[1] Van der Veer Institute for Parkinson's and Brain Research, Christchurch, New Zealand
[2] Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand
[3] Department of Medical Physics and Bioengineering, Christchurch Hospital, Christchurch, New Zealand
[4] Department of Medicine, University of Otago, Christchurch, New Zealand
[5] Department of Psychology, University of Canterbury, Christchurch, New Zealand

E-mail: richard.jones@vanderveer.org.nz

Received 26 May 2010
Accepted for publication 22 November 2010
Published 19 January 2011
Online at stacks.iop.org/JNE/8/016003

Abstract
A system capable of reliably detecting lapses in responsiveness ('lapses') has the potential to increase safety in many occupations. We have developed an approach for detecting the state of lapsing with second-scale temporal resolution using data from 15 subjects performing a one-dimensional (1D) visuomotor tracking task for two 1 h sessions while their electroencephalogram (EEG), facial video, and tracking performances were recorded. Lapses identified using a combination of facial video and tracking behaviour were used to train the classification models. Linear discriminant analysis was used to form detection models based on individual subject data and stacked generalization was utilized to combine the outputs of multiple classifiers to obtain the final prediction. The performance of detectors estimating the lapse/not-lapse state at 1 Hz based on power spectral features, approximate entropy, fractal dimension, and Lempel-Ziv complexity of the EEG was compared. Best lapse state estimation performance was achieved using the detector model created using power spectral features with an area under the curve from receiver operating characteristic analysis of 0.86 ± 0.03 (mean$\pm$SE) and an area under the precision-recall curve of 0.43 ± 0.09. A novel technique was developed to provide improved estimation of accuracy of detection of variable-duration events. Via this approach, we were able to show that the detection of lapse events from spectral power features was of moderate accuracy (sensitivity = 73.5%, selectivity = 25.5%).

1. Introduction

Human operators continue to have a critical hands-on role in many transportation sectors. The human operator can become fatigued, lose motivation and become considerably less effective, especially during long-term monotonous activities such as driving (Bittner *et al* 2000). In many cases, the operator can experience brief instances of complete loss of responsiveness—'lapses'. These can occur from a complex interaction of factors such as boredom, physical and mental exhaustion, lack of sleep or reduced quality of sleep, and the influence of circadian rhythms (Freund *et al* 1995). Lapses generally manifest as either lapses of sustained attention (Weissman *et al* 2006) or behavioural microsleeps (Peiris *et al* 2006). The term 'lapse' has also been used by others to describe delayed responses to target stimuli (Dorrian *et al* 2005, Weissman *et al* 2006) and response errors (Padilla *et al* 2006).

Such lapses are an occupational hazard for professional transport operators, such as coach and truck drivers, train drivers, air traffic controllers, and long-haul flight crew, who are expected to maintain schedules, and work shifts performing monotonous tasks for extended periods of time, despite the level of physical or mental fatigue they may be feeling. A

[6] Author to whom any correspondence should be addressed.

complete loss of responsiveness, even for a few seconds, while engaged in a critical task such as driving a vehicle or landing an aircraft can have consequences ranging from minor injuries to multiple fatalities.

A device able to detect the onset of lapses in real time using physiological cues from an individual would be especially beneficial to workers in the transport sector and would help minimize accidents caused by lapsing while performing tasks such as driving. In contrast to a substantial number of drowsiness estimation approaches in the literature (Belyavin and Wright 1987, Jung *et al* 1997, Grace 2001, Grace *et al* 1998, Lal and Craig 2002, 2005, Lin *et al* 2005a, 2005b, 2006, Makeig and Inlow 1993, Makeig and Jung 1995, Matousek and Petersen 1983, Papadelis *et al* 2007, Van Orden *et al* 2000, Arjunan *et al* 2009, Davidson *et al* 2007, Golz *et al* 2007, Sommer *et al* 2009, Johns 2003, Jap *et al* 2009, 2010, Eoh *et al* 2005, Schleicher *et al* 2008), there have been few serious attempts to detect lapses. These include EEG-based approaches utilizing ANNs (Sommer *et al* 2002, 2009, Davidson *et al* 2007, Golz *et al* 2007), video- or EOG-based eye-closure-based systems utilizing PERCLOS (Wierwille and Ellsworth 1994), and fusion of multiple behavioural and physiological features (Golz *et al* 2007).

Several drowsiness/microsleep detection systems based upon eye closure, eye gaze, and/or head pose have advanced to the level of becoming available on the market. These generally use a video-based approach (SeeingMachines—www.seeingmachines.com; SmartEye—www.smarteye.se) or an infra-red reflectrometry approach (Optalert—www.optalert.com). These systems appear sensitive in the detection of longer microsleeps and, hence, are an important advance towards reducing accidents in the transport sector. However, they are only able to reliably detect microsleeps several seconds after their onset. Only EEG has the intrinsic potential to allow detection of lapses close to, and even before, their onset (Davidson *et al* 2007).

EEG-based alertness/drowsiness detection systems to date have used larger time scales to smooth the performance metric, resulting in a time resolution of 1 min or more (Makeig *et al* 1996, Lin *et al* 2005a, 2005b, 2006, Jung *et al* 1997, Sommer *et al* 2009, Arjunan *et al* 2009). Furthermore, several drowsiness/alertness estimation methods also required training a model for each individual, to predict their performance in subsequent sessions (Lin *et al* 2005a, 2005b, 2006, Jung *et al* 1997).

Our previous paper (Peiris *et al* 2006) demonstrated that lapses are a common phenomenon, even in non-sleep-deprived subjects performing a monotonous task during normal work hours, with subjects lapsing frequently (39.3 ± 12.9/h) (mean $\pm$ SE). Analysis also revealed that these lapses are relatively brief (3.4 ± 0.5 s). The current paper presents an approach used for the detection of lapses with high temporal resolution using changes in the power spectrum and nonlinear features of the EEG.

Several researchers have used EEG spectral power to detect changes in the level of alertness and arousal (Huang *et al* 2001, Jung *et al* 1997, Makeig and Inlow 1993, Makeig and Jung 1995, 1996). Therefore, it seemed appropriate to begin the search for a reliable lapse detector by determining if there are EEG power spectral changes associated with lapses and, if so, determining the efficacy of a spectral-based lapse detection system.

Normal EEG is understood to have both linear and nonlinear dynamic properties, leading to EEG patterns with different degrees of complexity (Natarajan *et al* 2004). Thus it was hypothesized that nonlinear dynamical analysis techniques might prove a better approach to detect lapses than traditional linear methods (such as power spectral analysis) as they make better use of nonlinearities and dynamics in the EEG.

In recent years, progress in nonlinear dynamics theory has contributed new tools, useful in the analysis of the EEG (Elbert *et al* 1994). For example, nonlinear analytical techniques have been used to investigate the EEG associated with various physiological and pathological states, such as during meditation (Aftanas and Golocheikine 2002), sleep and slow-wave sleep (Kobayashi *et al* 2001, Ferri *et al* 1996), epilepsy (Elger *et al* 2000, Lehnertz 1999), and for assessing the depth of sedation (Klonowski *et al* 2006). There is evidence to suggest that nonlinear methods can be used to detect changes in the EEG that are not visible via visual observation or FFT (Le Van Quyen *et al* 2001).

Thus, in addition to spectral analysis, three nonlinear methods—fractal dimension (FD), approximate entropy (ApEn), and Lempel-Ziv complexity (LZ)—were investigated to determine their efficacy in EEG-based detection of lapses.

2. Methods

2.1. Subjects

As previously reported (Peiris *et al* 2006), 15 normal healthy male volunteers aged 18–36 years (mean = 26.5) performed a visuomotor tracking task while EEG, facial video, and tracking behaviour was recorded. The age range and gender were restricted to limit the sources of variation in the data. Based upon self-report, no subject had a current or previous neurological or sleep disorder and all had visual acuities of 6/9 or better in each eye. All subjects considered that they had slept normally the previous night (mean = 7.8 h, SD = 1.2 h, min = 5.1 h) and were considered non-sleep-deprived. Ethical approval for the study was obtained from the Canterbury Ethics Committee.

2.2. Visuomotor tracking task

Subjects were asked to perform a one-dimensional (1D) visuomotor tracking task with a continuous random preview target (Jones and Donaldson 1986; Jones 2006). The task, developed in-house, had a steering wheel (395 mm diameter, wheel-to-screen gain = 1.075 mm deg^{-1}) to control an arrow-shaped cursor located near the bottom of the screen. The eye-to-screen distance was 136 cm. Subjects were provided with an 8 s preview of a pseudo-random target (bandwidth 0.164 Hz, period 128 s) which scrolled down the screen at a rate of 21.8 mm s^{-1}. The task required smooth movements over a

175° range of the steering wheel and measured a subject's ability to keep the point of the arrow on the moving target. The position of the steering wheel was sampled at 64 Hz using a potentiometer mounted on the shaft of the wheel.

2.3. Video

Head and facial features were recorded from an analogue video camera (Sony Handycam) positioned 1 m in front of the subject using a frame rate of 25 Hz. The video was time-stamped. The time-synchronized video provided an independent measure of alertness and enabled us to confirm the presence of behavioural microsleeps.

2.4. EEG

EEG was recorded from electrodes at 16 scalp locations and digitized at 256 Hz (bandwidth 0.1–100 Hz) with a 16-bit A–D converter. Electrodes were placed according to the international 10–20 system (Klem *et al* 1999). The following standard bipolar derivations were used in the feature calculations: Fp1-F7, F7-T3, T3-T5, T5-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F4, F4-C4, C4-P4, and P4-O2. Bipolar channels were preferred over referential channels as they reject common-mode noise better.

2.5. Procedure

Each subject attended two sessions at least one week apart (mean 17 days, range 7–50 days) and held, following lunch, between 12.30 pm and 5.00 pm. They were instructed to stay alert and perform the task to the best of their ability for a full continuous hour in both sessions. Participants were instructed to refrain from taking any alertness-altering drug or medication (e.g. stimulants—coffee, amphetamine, tea; depressants—alcohol) 4 h prior to all test sessions.

3. Analysis

3.1. Lapse index

One of the first challenges in the data analysis phase was to develop an independent measure of whether or not the subjects were lapsing, i.e. a *lapse index* (LI). We used two independent measures—tracking task performance and facial video—to determine and categorize lapses in a subject.

Lapses in tracking performance are most obvious when the response cursor simply stops moving for an extended period while the target is moving or when the tracking response is non-coherent with the target. Only the first category—*flat spots*—was included in an intentionally conservative analysis, as lapses in the second category are difficult to identify with confidence. Flat spots occurring when the target velocity was approximately 0 (at turning points) were not counted, as at these times the subject could track adequately without moving the response cursor.

The video recording of each session was conservatively rated (by MP) (Peiris *et al* 2006), without knowledge of the corresponding tracking performance. Being 'blind' to tracking performance ensured that an independent measure of alertness was obtained from the video data. The video was rated on a 6-level scale: 1 = alert, 2 = distracted, 3 = forced eye closure while alert, 4 = light drowsy, 5 = deep drowsy, and 6 = behavioural microsleep (BM). Criteria similar to Wierwille and Ellsworth (1994) were used to define the video rating scale. *Video BMs* (i.e. sleep events) were identified by prolonged eye-lid closure, sometimes accompanied by rolling upward or sideways movements of the eyes, head-nodding, and often terminated by waking head jerks. Transitions in the video recording had a time resolution of 1.0 s.

Intervals in which flat spots and video BMs overlapped in time were defined as *definite BMs*. EEG data of subjects who had at least one definite BM over the two sessions were selected for lapse detector design ($N = 8$). Lapses, as defined by the presence of either a video BM and/or a flat spot, were selected as the events to be detected by an EEG-feature-based lapse detector. The LI was generated at a frequency of 1 Hz.

3.2. EEG feature extraction—an overview

Each EEG channel was processed by rejecting epochs contaminated with artefacts, as preliminary work had shown that removal of artefacts from the EEG improved detector performance. Noise introduced by EEG artefacts may be counterproductive during model formation and classifier performance evaluation and, hence, reduce classifier effectiveness. As the first step of the artefact removal stage, the EEG was pre-processed using independent components analysis (ICA) to remove eye blink artefacts (Delorme and Makeig 2004, Jung *et al* 2000). The eye-blink artefact-free signal from each derivation was then filtered to remove 50 Hz mains activity using an IIR notch filter with a Q-factor of 35.

The mean and standard deviation of the first 2 min (baseline) of the signal were calculated. The signal was then transformed into z-scores relative to the baseline of the signal, thus enabling comparisons to be made between subjects and sessions. Epochs of 2 s containing samples with an absolute z-score >3.0 were rejected as artefacts and excluded from analysis in the signal processing algorithms.

A *feature* is defined here as an arbitrary time series extracted from a single EEG derivation using a given signal processing algorithm. For example, if power spectral analysis is used to process the EEG, an extracted feature is the power in the alpha band over a set of consecutive epochs.

An epoch length of 2.0 s and an overlap of 1.0 s (50%) between successive epochs were used for all signal processing algorithms. The sliding process generated feature samples at a rate of 1 Hz, resulting in a 3600-element feature vector for 1 h recording. The 2.0 s epoch length was chosen to obtain a reasonable degree of spectral resolution (where appropriate) and the overlap of 1.0 s was chosen to ensure reasonable temporal resolution (an estimate every second) for the features. This was important since a key requirement of the desired lapse detection system was its ability to detect short lapses (1–2 s).

3.3. Power spectral analysis

Data in each 2 s epoch were first detrended to remove any linear trends (i.e. DC shifts) and the spectrum then estimated using a

Table 1. Spectral features calculated from each EEG derivation.

Feature	Frequency band
Mean spectral power[a]	
Delta (δ)	1.0–4.5 Hz
Theta (θ)	4.5–8.0 Hz
Alpha 1 (α_1)	8.0–10.5 Hz
Alpha 2 (α_2)	10.5–12.5 Hz
Alpha (α)	8.0–12.5 Hz
Beta 1 (β_1)	12.5–15.0 Hz
Beta 2 (β_2)	15.0–25.0 Hz
Beta (β)	12.5–25.0 Hz
Gamma 1 (γ_1)	25.0–35.0 Hz
Gamma 2 (γ_2)	35.0–45.0 Hz
Gamma (γ)	25.0–45.0 Hz
High	>45.0 Hz
Overall	0.1–100 Hz
Spectral power ratios[b]	
$\theta/\beta, \theta/\alpha, \alpha/\beta, \delta/\theta, \alpha/\delta, \beta/\delta, \beta_2/\alpha, \beta_1/\beta_2$	–

[a] Absolute values and normalized values.

[b] Absolute values only.

40th-order Burg model (Naidu 1996). This parametric model method was selected to estimate power spectra due to its ability to provide a high degree of frequency resolution for short data records (Subasi 2005). A high model order was found necessary to obtain adequate separation of the spectral bands of interest as lower-order Burg models 'blurred' the spectrum, hindering the separation of spectral peaks in adjacent bands.

The spectral features listed in table 1 were calculated. For a given epoch, the *spectral power* in each EEG band was calculated by finding the mean power across the band. Next, the *normalized power* was calculated for each band by dividing the spectral power in that band by the overall mean power across the entire spectrum. In addition, power ratios between bands were also calculated (table 1). Power spectral analysis produced 13 spectral power (SP), 12 normalized spectral power (NSP), and 9 power ratio (PR) features per EEG derivation, giving a total of 34 features per derivation and $34 \times 16 = 544$ spectral features over the 16 derivations.

3.4. Fractal dimension

Fractal dimension (FD) provides an estimate of the complexity of a signal. It has the advantage that it can be calculated directly in the time domain without reconstruction and, hence, provides a direct link between EEG variations and complexity changes (Accardo *et al* 1997).

The brain has been interpreted as a nonlinear dynamical system whose state can be described by self-similar curves (Lutzenberger *et al* 1995). EEG signals are an example of such curves and their complexity, as estimated by FD, has been shown to correspond to different physiopathological conditions (Accardo *et al* 1997). The FD of any signal varies between 1 and 2: the more complex a waveform, the higher is its FD. FD has been shown to be effective as a means of comparing differences in the complexity of EEG signals recorded from patients with bipolar mood disorder and controls (Bahrami *et al* 2005) and in the analysis of epileptic ictal events (Bullmore *et al* 1994).

Higuchi's algorithm (Higuchi 1988) was used to estimate the fractal dimension (FD) of each EEG derivation because it is computationally efficient and also provides a stable estimate of FD using a lower number of samples of data ($N \geqslant 125$) compared to other FD algorithm implementations (Accardo *et al* 1997). This allowed the FD to be estimated with the same temporal resolution as other features.

The parameters suggested by Accardo *et al* (1997) were used for estimating the FD of the EEG ($k_{\max} = 6$, in which k is the scale size, corresponding to successive samples apart). The FD was estimated for each EEG derivation, resulting in 16 FD feature vectors per session with a value of FD every 1.0 s.

3.5. Approximate entropy

Entropy is a concept that addresses system randomness and predictability (Grassberger and Procaccia 1983). It quantifies the predictability of the amplitude values of a signal, based on knowledge of amplitudes of previous samples (Bruhn *et al* 2000). Approximate entropy (ApEn) is non-negative, with a larger number indicating more irregularity, unpredictability, and randomness of the raw signal (Zhang and Roy 2001). A perfectly regular data series, in which knowledge of prior values enables the subsequent value to be predicted perfectly, has an associated ApEn measure of 0. However, with increasing irregularity, the prediction of a subsequent value becomes increasingly worse, leading to an increased ApEn value.

ApEn has been used in a variety of contexts including human respiratory variability (Burioka *et al* 2003), estimation of depth of anaesthesia (Zhang and Roy 2001, Bruhn *et al* 2000) and differentiating between sleep stages (Burioka *et al* 2005).

The two parameters in the ApEn algorithm are the embedding dimension m and tolerance of the noise filter r. The embedding dimension m specifies the number of previous values used for the prediction of the subsequent value and the noise filter value r is expressed as a proportion of the standard deviation of the amplitude values of the n samples in the data sequence (Bruhn *et al* 2000). Values suggested in the literature for m and r of 2 and 0.2, respectively, were used in the ApEn algorithm (Bruhn *et al* 2000, Pincus 1995, Pincus *et al* 1991, Zhang and Roy 2001). ApEn was estimated for each EEG derivation, resulting in 16 ApEn feature vectors per session with an ApEn value calculated every 1.0 s.

3.6. Lempel-Ziv complexity

LZ complexity (Lempel and Ziv 1976) provides a non-parametric measure of complexity of a 1D signal, such as the EEG. Its advantages are that it is simple to compute, does not require long data segments to be effective, and is more effective for real-time EEG processing (Zhang *et al* 1999, Zhang and Roy 1999, Radhakrishnan and Gangadhar 1998) compared to other complexity measures such as correlation dimension (Yaylali *et al* 1996) and neural complexity (Tononi *et al* 1994). It has been useful in quantifying the depth of anaesthesia (Zhang and Roy 2001), predicting epileptic

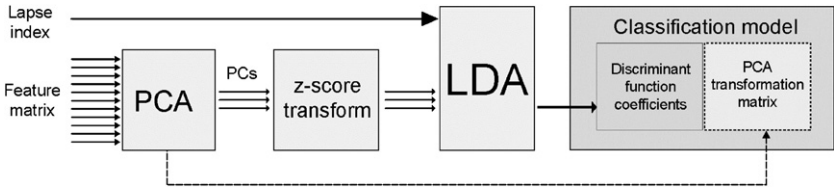

Figure 1. A block diagram depicting the creation of a lapse classification model.

seizures (Radhakrishnan and Gangadhar 1998), and analysing the dynamical behaviour of the background EEG of patients with Alzheimer's disease (Abasolo *et al* 2006).

Lempel and Ziv proposed that the complexity of a finite sequence could be evaluated from the point of view of a 'simple self-delimiting learning machine which, as it scans a given digit sequence $S = s_1, s_2, \ldots, s_n$ from left to right, added a new word to its memory every time it discovered a sub-string of consecutive digits not previously encountered'. The complexity counter $c(n)$ is increased by one unit each time a new sub-string of characters is encountered along S (Lempel and Ziv 1976, Radhakrishnan and Gangadhar 1998). Only two operations are permitted in the construction of a string: copying old patterns and inserting new ones (Zhang and Roy 2001).

LZ complexity features were estimated for each EEG derivation, resulting in 16 LZ complexity feature vectors per session with a value calculated every 1.0 s.

3.7. Feature matrix assembly

A *feature matrix* for a session's EEG data was created by grouping various combinations of EEG features calculated using the signal processing algorithms described earlier. This was achieved by placing m feature vectors (each of length n) as row vectors in the feature matrix.

For algorithms which produced one feature per EEG derivation (such as FD, ApEn and LZ), the generated EEG feature matrix was of size 16 by n. For example, the EEG feature matrix based on FD measures had 16 FD feature vectors as row vectors in the matrix. However, for measures such as power spectra, which had multiple features per EEG derivation, the size of the EEG feature matrix was much larger. These were arranged in the EEG feature matrix as rows, in order of EEG derivation. That is, all the spectral features for the first derivation were listed in the first 34 rows of the feature matrix, followed by the features of the second derivation from rows 35 to 69, etc until all features from all derivations were entered into the feature matrix. This resulted in 544 (34×16) feature vectors in the matrix for power spectra.

3.8. Classification models to detect lapses from EEG features

Principal component analysis (PCA) was used to transform the feature vectors into orthogonal components, so as to reduce the redundancy within the original features and aid the formation of the classification models.

The principal components (PCs) were ranked in order of descending importance in terms of the amount of variance explained. It was then possible to reduce the dimensionality of the data by using the first p of the total m PCs ($p < m$) without significant loss of information by choosing p appropriately.

The next step in the design of the lapse detection system was to train a *classification model* capable of detecting lapses in new subjects, using data from their feature matrices (figure 1). The process involved forming a classification model, based upon linear discriminant analysis (LDA) (Fisher 1936) and using PCs extracted from the feature matrix as predictive variables and LI as the grouping variable. The lapse indices and EEG feature matrices from both sessions of each subject were concatenated to form a single feature matrix ($m \times 7200$) and LI (1×7200) per subject. These data were then used to form a classification model for each of the eight subjects.

Firstly, the mean over the entire length of the record was calculated for each vector of the feature matrix. The means were then subtracted from the feature vectors to produce zero-mean vectors in the feature matrix. This was a necessary prerequisite for PCA. Following this, PCA was performed on the mean-subtracted $m \times n$ feature matrix to derive m PCs, each of length n. The PCs were then converted to z-scores by subtracting the overall means and dividing by the standard deviations. The z-score transformed PCs and the LI were used to form a linear discriminant classification model for each subject, via MATLAB® discriminant analysis toolbox (Kiefte 1999).

3.9. Combining multiple classification models to form an overall detection model

Combining the output of several models generally increases predictive performance over a single model (Witten and Frank 2000). Stacked generalization (or simply 'stacking') (Wolpert 1992) was chosen to combine the outputs of the multiple lapse classifiers in this work. Stacking aims to determine how to best combine the base models via an additional *meta-learner* algorithm.

The outputs of the base models (also known as level-0 models) were fed as the inputs to the meta-learner (level-1 model). During the classification phase of the stacked learner, new cases were fed into the level-0 models, each producing a classification value at their output. These level-0 predictions were then fed into the level-1 model which combined them linearly by scaling the output of each model by its weight, summing the scaled model outputs, and applying a threshold to the summed output to obtain an overall prediction (figure 2).

It has been suggested that some of the test data be held back and used to train the level-1 model, with the level-0

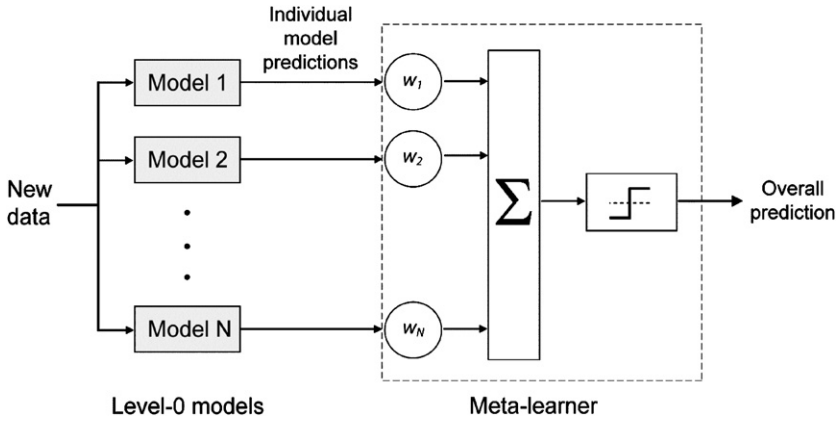

Figure 2. Diagram showing the internal structure of the meta-learner used in the stacked generalization approach. The model weights are depicted by w_1, w_2, ..., w_N and these are used by the meta-learner to scale the level-0 model outputs before combining them to form an overall prediction.

models being trained on the remaining data (Witten and Frank 2000). Once the level-0 models are trained, the holdout data are classified using the level-0 models, which then form the training data for the level-1 model. Since the holdout data were not used to train the level-0 models, their predictions are unbiased and, therefore, the level-1 training data accurately reflect the true performance of the level-0 models. However, the downside of the holdout method is that it deprives the level-1 model of some of the training data. This problem was overcome by applying eightfold cross-validation which ensured that all of the training data were used to train the level-1 model. Each instance of the training data was used in one test-fold of the cross-validation and the predictions from the models built from the corresponding training fold were used to build the level-1 training set. This generated a level-1 training set for each level-0 training set.

3.10. Overall detection model validation

The following steps were followed to validate the overall lapse detection model.

(1) Reserve one of the eight subjects as the *validation subject* and put his/her data aside.
(2) Create classification models using data from the seven remaining subjects.
(3) Select one subject from the seven subjects (*test subject*) in () and feed their features into the six level-0 models (excluding their own) for classification. This yields six level-0 outputs which are stored in a matrix. Note that the 'raw' output of the classifiers (i.e. continuous values between 0.0 and 1.0 indicating the probability of a given sample being a lapse) was used in the steps that follow.
(4) Determine the meta-learner weights for the six level-0 model outputs by linearly combining them to estimate the LI of the test subject. Constrained least-squares fitting (coefficients restricted to >0) was used to combine the output of the six models. This approach minimized the least-squares error between the combined output (i.e. meta-learner output) and LI. It produced a set of positive regression coefficients for the six models, with larger coefficients associated with models contributing a greater degree towards the meta-learner output. These coefficients were stored in a matrix.
(5) Determine the optimal threshold value required to be applied to the meta-learner output to obtain a binary classification (i.e. lapse/not lapse) by selecting the threshold that yields the maximum phi correlation between the meta-learner output and LI.
(6) Repeat steps (3) to (5) until all seven subjects are used as test subjects.
(7) Calculate mean *meta-learner weights* and mean *meta-learner output threshold* by averaging over the seven test subjects.
(8) Feed the validation subject's data to all seven level-0 models in the stacked generalization system and obtain the final prediction from the meta-learner output. The meta-learner scales the individual predictions of the level-0 models by the weights calculated in (7), sums predictions of all level-0 models, and finally applies the output threshold also determined in (7) to provide a final (binary) prediction of lapse (1) or no-lapse (0).
(9) Calculate the correlation between the validation subject's LI and the meta-learner output after applying the mean output threshold. The correlation measure used was the phi correlation coefficient (φ) (Sheskin 1997) with each validation subject's phi coefficient denoted by φ_v.
(10) Repeat steps (1) to (9) and obtain φ_v for each of the eight subjects.
(11) Calculate the mean across all eight values of φ_v to give the overall detector performance.

Performance of the lapse detector was evaluated using several metrics. The primary performance metric was the mean phi correlation, as described above. In addition, two other performance measures (which are independent of operating point) were also calculated: (a) area under the receiver-operator characteristic curve (AUC-ROC) and (b) area under the precision-recall curve (AUC-PR). These calculations were performed using the ROCR package (Sing *et al* 2005). The performance of the classification models is quoted using all three measures.

The effect of using uniform meta-learner weights on the performance of the overall lapse detector was also investigated. This process is equivalent to removing the stacked generalization section from the system described above.

3.11. Contribution of features to discrimination

After determining the performance of the overall detection model via cross-validation, it was possible to determine the amount each feature contributed to the overall discrimination ability of the model. Firstly, the discriminant coefficients of each classification model were converted to standardized discriminant coefficients. These coefficients were then normalized and used to form a *feedback weight vector*. This contained m elements, corresponding to the PC features used to construct the classifier. That is, the feedback weight vector's elements indicated the relative contribution towards lapse classification of each of the PCs. However, as each PC was a linear combination of all input features, it was possible to translate the feedback weight vector back to feature space to determine the relative contribution of each of the original features to the classification model. This was achieved by multiplying the feedback weight vector by the inverse of the PCA transformation matrix (P^{-1}) calculated during PCA at the model formation phase. This procedure provided the *relative contribution* of each feature towards the classification power of a particular level-0 model. The procedure was repeated to calculate the relative contribution of features in all level-0 models. However, the generalization performances of the level-0 models were not equal and hence the relative contributions of each level-0 model were adjusted according to the model weights. These scaled contributions were then summed to obtain the contribution of features towards the overall detection model.

3.12. Detection of lapse events

In addition to estimating performance of detection of the lapse state (in 1.0 s epochs), we also wished to determine the detector's ability to detect discrete *lapse events*. The following novel procedure was used to determine lapse event detection performance.

(1) An *event signal* was created. This was the same length as the LI with a sampling frequency of 1 Hz.
(2) The event signal was initialized to a default value of 0, which was defined as a true negative (TN) event.
(3) A predetermined optimum threshold was applied to the overall lapse detector output to obtain binary *detector lapse events*. A detector output of 1 was defined to correspond to a *detector lapse event*, and a detector output of 0 to correspond to the responsive state.
(4) The gold standard (i.e. LI) was traversed until a *gold standard lapse event* was encountered.
(5) The detector output was checked to see if it equalled 1 during any portion of the gold standard lapse event. If yes, the entire portion of the event signal corresponding with the gold standard lapse event was marked as a true positive (TP) event. Furthermore, the TP event was extended at either end to include any overlapping detector lapse event. If the detector output was 0 during the entirety of the gold lapse event, the corresponding region of the event signal was marked as a false negative (FN) event.
(6) The detector output was traversed until a detector lapse event was encountered.
(7) The region of the event signal corresponding to the detector lapse event was checked. If marked as a TN, the region in the event signal corresponding to the detector lapse event was re-marked as a false positive (FP) event.
(8) The number of TPs, FPs, TNs, and FNs in the event signal were counted and the following performance parameters calculated.

- Sensitivity [*true positive rate* or *hit rate*] = TP/(TP + FN)
- Selectivity [*positive predictive value* or *precision*] = TP/(TP + FP)
- Specificity = TN/(TN + FP)
- Negative predictive value = TN/(TN + FN)
- Accuracy = (TP + TN)/(TP + TN + FP + FN)

An example of how the event signal was generated from gold standard lapse events and detector lapse events is shown in figure 3.

4. Results

The performance of a lapse detector can be measured both in terms of ability to detect the *lapse state* (in 1 s epochs) and ability to detect actual *lapse events*. Most of the following results are with respect to lapse state. The relative contributions of EEG features and derivations to the best overall lapse detection model are also presented.

Detector performance reached a plateau after approximately 50 PCs but adding additional PCs to the model did not cause over-fitting nor reduce overall cross-validation performance. Therefore, it was decided to include all PCs in the construction of subsequent lapse detector models as this avoided having to determine the number of PCs to be used for model formation.

4.1. Detector performance—spectral measures

Table 2 provides a summary of system performance for a lapse detector based on spectral power (SP), normalized spectral power (NSP), power ratios (PR) and combinations thereof. Detector performances are shown for the two weighting schemes (uniform and least-squares) used to scale the classifier outputs to obtain the overall prediction. Detectors based solely on SP, or incorporating SP with NSP, and utilizing least-squares weights for combining the classifier outputs (to obtain the overall prediction) provided the best generalization performance ($\varphi = 0.39$). This was confirmed by observing that SP and SP+NSP based detectors had the largest AUC-ROC and AUC-PR values, as shown in table 3. As mentioned earlier, AUC-ROC and AUC-PR being threshold-independent measures emphasizes that SP and SP+NSP indeed give the best performance out of the seven spectral-feature-based detectors.

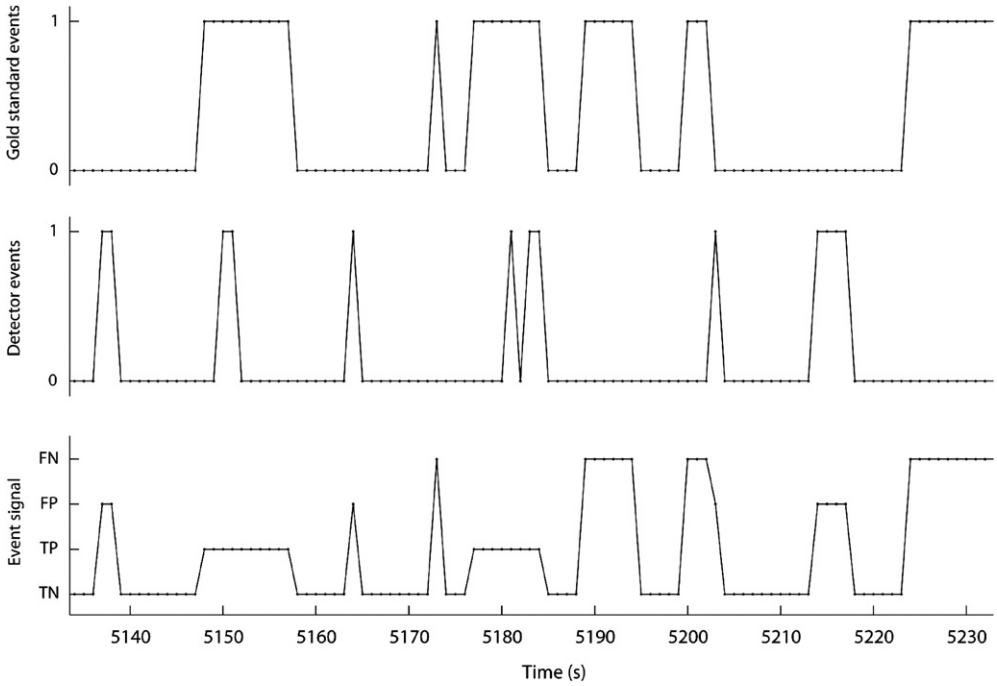

Figure 3. An example illustrating how the event signal (bottom plot) was derived from gold-standard events (top) and detector events (middle). Note that in the top two plots, values of 1 and 0 indicate lapse events and responsive states, respectively.

Table 2. The mean detector performances (φ) for systems trained to detect lapses using spectral power, normalized spectral power and power ratio measures. The detector performances for uniform and constrained least-squares weighting regimes are shown.

	Detector performance (φ)	
Detector features	Uniform weights (mean ± SE (min, max))	Constrained least-squares weights (mean ± SE (min, max))
Spectral power (SP)	0.38 ± 0.06 (0.06, 0.59)	0.39 ± 0.06 (0.13, 0.62)
Normalized spectral power (NSP)	0.32 ± 0.05 (0.12, 0.49)	0.33 ± 0.05 (0.11, 0.57)
Power ratios (PR)	0.34 ± 0.05 (0.12, 0.47)	0.33 ± 0.05 (0.10, 0.52)
SP+NSP	0.37 ± 0.06 (0.11, 0.56)	0.39 ± 0.06 (0.12, 0.62)
SP+PR	0.37 ± 0.06 (0.09, 0.57)	0.37 ± 0.06 (0.12, 0.57)
NSP+PR	0.32 ± 0.05 (0.10, 0.50)	0.32 ± 0.05 (0.09, 0.49)
SP+NSP+PR	0.36 ± 0.06 (0.10, 0.56)	0.36 ± 0.06 (0.11, 0.59)
Spectral asymmetry	0.18 ± 0.05 (0.02, 0.36)	0.17 ± 0.05 (0.02, 0.36)
Spectral coherence	0.15 ± 0.03 (0.00, 0.30)	0.15 ± 0.03 (0.00, 0.26)

Table 3. AUC-ROC and AUC-PR curves for spectral detectors used to detect lapses for both uniform and constrained least-squares weighting regimes. These curves indicate detector performance independent of meta-learner output threshold (cf table 2).

	Detector performance			
	Uniform weights		Constrained least-squares weights	
Detector features	AUC-ROC (mean±SE)	AUC-PR (mean±SE)	AUC-ROC (mean±SE)	AUC-PR (mean±SE)
Spectral power (SP)	0.86 ± 0.03	0.41 ± 0.09	0.86 ± 0.03	0.43 ± 0.09
Normalized spectral power (NSP)	0.82 ± 0.04	0.40 ± 0.09	0.82 ± 0.04	0.38 ± 0.09
Power ratios (PR)	0.83 ± 0.03	0.38 ± 0.09	0.83 ± 0.03	0.39 ± 0.09
SP+NSP	0.86 ± 0.03	0.42 ± 0.09	0.86 ± 0.03	0.44 ± 0.10
SP+PR	0.85 ± 0.03	0.42 ± 0.10	0.85 ± 0.03	0.43 ± 0.10
NSP+PR	0.81 ± 0.04	0.39 ± 0.09	0.81 ± 0.04	0.39 ± 0.09
SP+NSP+PR	0.85 ± 0.03	0.42 ± 0.10	0.85 ± 0.03	0.43 ± 0.10

Figure 4 shows the mean ROC and PR curve for the spectral power (SP)-based detector. Overall, there was no difference in performance between detectors trained using uniform weights versus constrained least-squares weights.

4.2. Simple SP detector model (non-stacked)

The performance of a single detection model (cf stacked model consisting of seven level-0 models followed by a level-1 meta-learner) created by lumping data of seven subjects together and

Table 4. Mean detector performances (φ) for systems trained to detect lapses using FD, ApEn, and LZ complexity measures, and using best spectral features (SP, SP+NSP) combined with LZ. The detector performances for uniform and constrained-LS weighting regimes are shown.

	Detector performance (φ)	
Detector features	Uniform weights (mean ± SE (min, max))	Constrained least-squares weights (mean ± SE (min, max))
Fractal dimension (FD)	0.21 ± 0.04 (0.08,0.40)	0.20 ± 0.03 (0.10, 0.38)
Approximate entropy (ApEn)	0.24 ± 0.06 (0.01,0.42)	0.22 ± 0.04 (0.05, 0.38)
Lempel-Ziv complexity (LZ)	0.28 ± 0.06 (0.04,0.46)	0.26 ± 0.05 (0.07, 0.49)
SP+LZ	0.38 ± 0.06 (0.08, 0.59)	0.39 ± 0.06 (0.12, 0.63)
SP+NSP+LZ	0.36 ± 0.06 (0.09, 0.58)	0.38 ± 0.06 (0.12, 0.62)

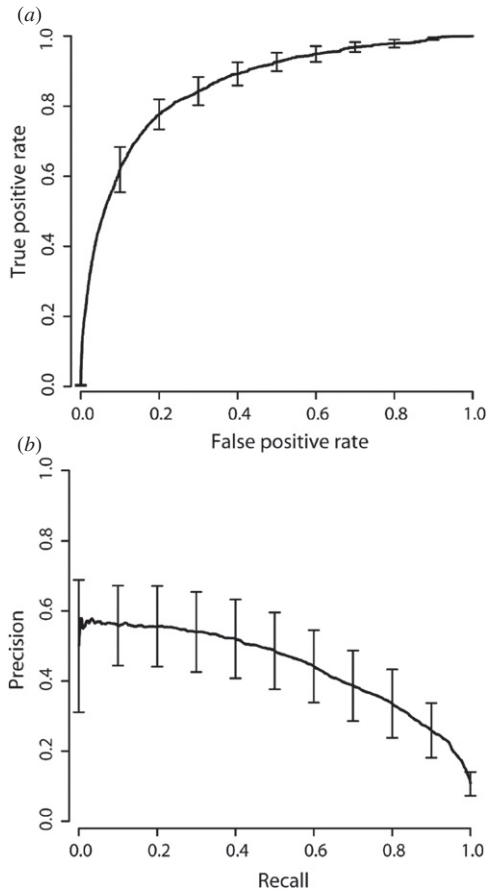

Figure 4. (a) Mean ROC and (b) PR curve for the lapse detector based on spectral power (SP) features. The vertical bars indicate standard error on both plots.

validated using the remaining subject's data was investigated to compare its performance with the stacked approach.

Firstly, one subject was left out for validation. The data of the remaining seven subjects were concatenated and a single classification model created. The lumped data were then fed through the model to determine the optimal output threshold. Finally, the validation subject's data was fed through the model and the phi correlation between the classification model output and the subject's LI calculated. This procedure was repeated until all eight subjects had been used for validation. The mean performance of the simple detector model was the mean over the eight validation runs.

The performance for the simple detector model using SP features was $\varphi = 0.31 \pm 0.07$ (0.06, 0.58). In comparison, when using the same spectral features, the stacked approach with LS constraints yields $\varphi = 0.39 \pm 0.06$ (0.13, 0.62) (table 2).

4.3. Detector performance—complexity measures

Tables 4 and 5 provide a summary of system performance for a lapse detector using measures of complexity of the EEG. The detector based on the LZ complexity measure provided the largest detector performance, as shown by the mean φ correlation. Another interesting observation is that a detector using uniform classifier outputs to generate the overall prediction performed better than a detector applying constrained least-squares weights to the classification models to arrive at the final prediction.

4.4. Detector performance—spectral and complexity features combined

As LZ complexity yielded the best detector performance of the three complexity measures, an investigation was undertaken to determine if detector performance would be improved by adding LZ to the spectral features. However, as tables 4 and 5 show, no performance improvement was seen by adding the LZ feature to the spectral-power-based detector.

4.5. Contribution of features to discrimination in a best lapse detector

The detector model based on spectral power was selected for analysis as it displayed the highest performance level ($\varphi \approx 0.39$). The proportion of contribution of each spectral feature towards the overall lapse detection model is shown in figure 5. Each proportion was determined by summing the contributions of the selected feature across all EEG derivations in all level-0 models. Likewise, the proportion of contribution to the overall detection model by each EEG derivation was calculated by summing the contributions of all spectral features of each EEG derivation across all level-0 models. The proportion of contribution by each EEG derivation is shown in figure 6. Generally, no strong spatial patterns are visible across derivations (apart from T6-O2 which has the largest contribution), indicating that each derivation contributes approximately equally to the overall model. In terms of power spectral features, changes in spectral power in the alpha band seem to be the largest contributor to the detection model (figure 5), although all frequency bands contributed to detection performance.

Table 5. AUC-ROC and AUC-PR curves for lapse detectors based on estimates of FD, ApEn, LZ complexity, and on best spectral features (SP, SP+NSP) combined with LZ. These curves indicate detector performance independent of meta-learner output threshold.

	Detector performance			
	Uniform weights		Constrained least-squares weights	
Detector features	AUC-ROC (mean ± SE)	AUC-PR (mean ± SE)	AUC-ROC (mean ± SE)	AUC-PR (mean ± SE)
Fractal dimension (FD)	0.77 ± 0.03	0.28 ± 0.07	0.75 ± 0.03	0.22 ± 0.05
Approximate entropy (ApEn)	0.77 ± 0.04	0.29 ± 0.07	0.74 ± 0.04	0.23 ± 0.05
Lempel-Ziv complexity (LZ)	0.80 ± 0.04	0.34 ± 0.08	0.78 ± 0.04	0.30 ± 0.08
SP+LZ	0.85 ± 0.03	0.42 ± 0.10	0.86 ± 0.03	0.44 ± 0.10
SP+NSP+LZ	0.85 ± 0.03	0.42 ± 0.10	0.86 ± 0.03	0.44 ± 0.10

Table 6. Event detection performance of the spectral-power-based lapse detector in terms of TPs, TNs, FPs, FNs, and sensitivity, specificity, selectivity, NPV and accuracy percentages.

		Event detector performance								
Subject	Total lapses	TP	TN	FP	FN	Sen.	Spec.	Sel.	NPV	Accy
1	235	118	328	102	117	50.2	76.3	53.6	73.7	67.1
2	77	59	289	221	18	76.6	56.7	21.1	94.1	59.3
3	12	9	299	290	3	75.0	50.8	3.0	99.0	51.2
4	55	52	483	433	3	94.5	52.7	10.7	99.4	55.1
5	46	45	292	246	1	97.8	54.3	15.4	99.7	57.7
6	109	70	327	223	39	64.2	59.4	23.9	89.3	60.2
7	194	155	453	278	39	79.9	62.0	35.8	92.1	65.7
8	189	166	352	179	23	87.8	66.3	48.1	93.9	72.0
Overall	917	674	2823	1972	243	73.5	58.9	25.5	92.1	61.2

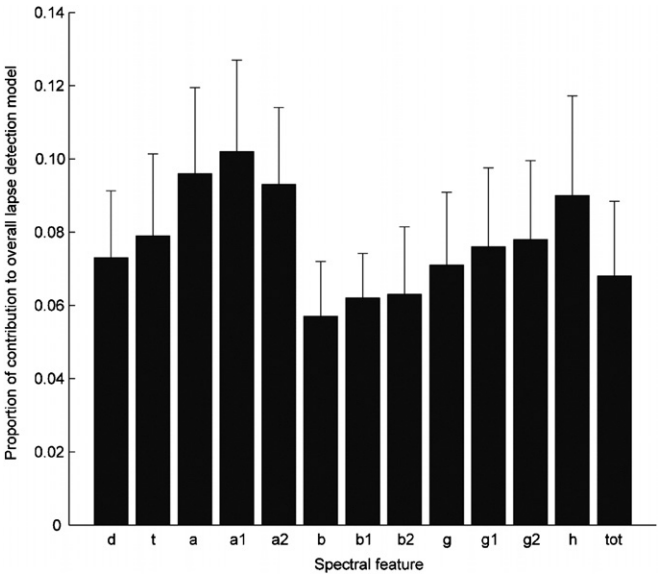

Figure 5. Mean proportion of contribution by each spectral feature to the overall lapse detection model. The contribution of each spectral feature was found by summing the contributions of the selected feature across all EEG derivations. The spectral features were delta (d), theta (t), alpha (a), alpha 1 (a1), alpha 2 (a2), beta (b), beta 1 (b1), beta 2 (b2), gamma (g), gamma 1 (g1), gamma 2 (g2), high (h) and total (tot).

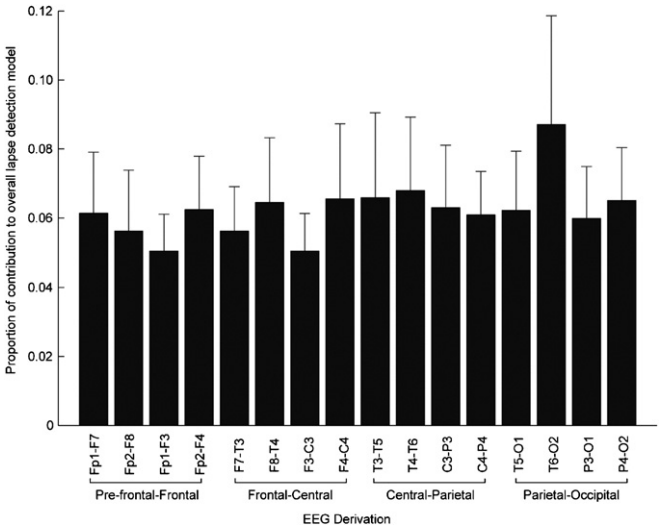

Figure 6. Mean proportion of contribution by each EEG derivation to the overall lapse detection model (based on power spectral features). The contribution of each derivation was found by summing the individual contributions of all spectral features within each derivation.

4.6. Detection of lapse events

The performance of the spectral-power-based lapse detector in terms of its ability to detect lapse events is summarized in table 6. Overall event detection performance was calculated by concatenating the data from all eight subjects. This yielded an overall sensitivity of 73.5%, selectivity of 25.5% and accuracy of 61.2%.

4.7. Effect of lapse duration on detection

An analysis was conducted to determine the relationship between the duration of a lapse and the likelihood of it being detected. As before, the best-performing lapse detector (i.e. the spectral-power-based detector) was used.

The system successfully detected 362 lapses and missed 621 lapses over the eight subjects. The median durations of detected and missed lapses in the pooled data from all eight subjects were 4.0 and 3.0 s, respectively (note that the lapse detection resolution is 1.0 s), the difference being marginal (Wilcoxon: $p = 0.059$). There were 424 false detections over the eight subjects.

An improvement in detector sensitivity and selectivity with increasing lapse duration was observed. All lapses greater than 20 s were detected by the system. There was only one false detection beyond 10 s (duration = 37 s).

5. Discussion

We have developed and evaluated procedures for detection of lapses of responsiveness based on linear (spectral), nonlinear, and a combination of both linear and nonlinear features of the EEG. The best detector performance was achieved with a detector model created using spectral power (SP) features and a meta-learner based upon stacked generalization and constrained least-squares weights. Classification models created using normalized spectral power (NSP) or power ratios (PR) features had lower mean performances than the SP-based detector. The performance of NSP/PR detectors showed a marginal increase when SP features were added to the NSP/PR features.

Of the three nonlinear feature-based detectors, the LZ complexity feature-based detector showed the highest performance, followed by the detector based on approximate entropy. Interestingly, the use of uniform meta-learner weights showed a marginally higher performance than using constrained least-squares meta-learner weights (cf spectral detectors) for all three nonlinear detectors in terms of φ, AUC-ROC, and AUC-PR.

The performance of a lapse detector created by adding the best performing nonlinear features (LZ complexity) to the best linear features (SP) was no greater than the detector based on SP alone. This is similar to the finding of Golz *et al* (2007) and suggests that nonlinear features contribute no additional information to the detector and that they effectively contain information similar to SP features. Note that LZ complexity generated 1 feature per channel to give 16 features per subject (1 feature/channel $\times$ 16 channels), whereas SP contained 208 features (13 features/channel $\times$ 16 channels). One might therefore propose that the LZ detector had a disadvantage over SP in that it had a much smaller set of features compared to SP and, hence, a lower detection performance due to this. However, analysis showed that an SP detector limited to using the first 16 PCs still performed better than the LZ detector ($\varphi = 0.36 \pm 0.05$ versus 0.28 ± 0.06).

We have also developed a novel procedure for estimating measures of accuracy (sensitivity, specificity, etc) of detection/classification of variable-length behavioural events, particularly as it applies to events occurring over an extended continuous recording. Using this approach, we were able to show that the performance of the spectral-power-based system in terms of lapse events was, at best, moderate, i.e. an average detection sensitivity of 73% and a substantial number of false detections. This low selectivity (25%) could have been improved by increasing the output threshold of the overall detector, but at the expense of decreased sensitivity.

It was also shown that the lapse detection system is more likely to detect longer lapses indicated by increasing values of sensitivity and selectivity with lapse duration. This is presumed to be primarily due to longer lapses having more pronounced EEG spectral changes related to microsleeps than shorter lapses. Note, however, that an increase in detection sensitivity with increased duration of lapse could also occur simply by chance (i.e. even if the detector output had no relationship with the occurrence of lapses), although this would also tend to be offset by a concomitant increase in longer-duration false detections.

EEG epochs with z-scores >3.0 were excluded as artefacts and not used for training and testing the lapse detection models. This may have biased the detector towards a higher level of performance. However, the excluded proportion of epochs was relatively low (8.5%) and, hence, it is unlikely that their elimination caused the system to display a substantially higher level of performance than what would have been obtained if 'contaminated' data were used to test the models.

Overall, the levels of performance and reliability demonstrated by the lapse detection models are considered too low to be of substantial value in real-world lapse/microsleep detection applications. However, they are encouraging as this task involved the substantial challenge of aiming to detect lapses at a temporal resolution of 1.0 s. This contrasts with studies that used a larger time scale to smooth performance metrics resulting in an estimate of alertness/drowsiness on a scale of 1 min or more (Makeig *et al* 1996, Lin *et al* 2005a, 2005b, 2006, Jung *et al* 1997, Sommer *et al* 2009, Arjunan *et al* 2009).

It must be emphasized that the approach proposed in this paper could not be used in real-time lapse detection, as means of the entire length of EEG feature matrix vectors were used during the normalization process. It is possible that this could be overcome by using mean feature vector values from previous sessions, but the efficacy of this still has to be demonstrated.

One of the main considerations in developing the current lapse detection system was the ability to generalize well to new subjects. This is another important distinction between the approach in this paper and that used in estimation of drowsiness (Lin *et al* 2005a, 2005b, 2006, Jung *et al* 1997) which required training a model for each individual, to predict their performance in subsequent sessions. Since a model tuned for each subject takes subtle differences in the individual's EEG into consideration, it is likely to yield superior performance. However, a major disadvantage is that each user requires substantial training prior to use of the device. Our expectation is that the overall detector performance would increase if the size of the training set was increased. Similarly, the small number of eight subjects in the cross-validation is a limitation of the current study. A larger number of subjects for cross-validation would have provided us with greater confidence in the generalization ability of the lapse detection model.

We have previously used normalized EEG log-power spectrum inputs to train a long short-term memory

recurrent neural network (RNN)-based lapse detector, which demonstrated a mean performance of $\varphi = 0.38 \pm 0.05$, AUC-ROC $= 0.84 \pm 0.02$, AUC-PR $= 0.41 \pm 0.08$ (Davidson *et al* 2007). The RNN approach was applied to the same dataset described in this paper. Both the RNN approach and the current linear approach used a time-resolution criterion of 1.0 s for detection of the lapse state. The results presented in this paper demonstrate that a relatively simple linear approach based upon spectral power is capable of achieving a very similar level of performance to that of an RNN approach. The linear approach has the added advantage of being computationally less intensive than the RNN to train. The comparable level of performance between the RNN and the linear spectral power approach described in this work is somewhat surprising and suggests that the linear detector has superior parsimony. It also suggests that lapses involve, at most, only a mild nonlinearity as otherwise one would expect neural-network-based detectors and linear-nonlinear-feature detectors to achieve higher performances if the solution was substantially nonlinear (Bishop 1995).

The only other group to have made substantial inroads into the detection of behavioural microsleeps from EEG is that of Golz *et al* (Golz *et al* 2007, Sommer *et al* 2002, 2009). They used a combination of EEG, EOG and video-based eye measures to detect behavioural microsleeps from spectral and nonlinear features in subjects during overnight sessions on a driving simulator. Using a support vector machine, they were able to classify correctly 98% of definitive microsleeps from an equivalent set of definitive non-microsleeps (Sommer *et al* 2009). This is an impressive achievement but needs to be moderated by the classification only having been applied to 15% of the overall dataset. That is, their system has a demonstrable high sensitivity for clear microsleeps but with an undetermined specificity. Notwithstanding, they were able to show both a high sensitivity and specificity for the microsleep state during the entire dataset when rated at 30 s intervals (Sommer *et al* 2009). They also used a continuous version of their classifier to estimate the binary microsleep/non-microsleep state at a rate of 10 Hz. However, in contrast to the techniques presented in the current paper and that of Davidson *et al* (2007), they were unable to validate their classifier in the detection of either the microsleep state (e.g. at 0.1 s or 1.0 s intervals) or individual microsleep/lapse events. This aside, by averaging the output of their continuous microsleep-state classifier over 4 min epochs, they were able to estimate more tonic levels of alertness/drowsiness (Sommer *et al* 2009).

In our detector, the advantage of creating multiple models and the resulting increase in detector performance was demonstrated. Using eight level-0 models and a meta-learner resulted in a mean phi correlation of 0.39 for the SP detector, whereas lumping all the features from seven subjects to create a single model and validating the lumped model on the eighth subject resulted in a mean phi of 0.31. The lumping of data prior to model formation may result in the model being biased towards certain subjects, such as those with the most lapses, resulting in a loss of generalizability. Using a stacked approach yields better performance as the level-0 models are adjusted by the meta-learner according to how well they generalize over the training set. It is expected that the mean detector performance would increase if the size of the training set was increased. However, this was not possible in this work as the dataset was limited to eight subjects.

Our finding that constrained-LS weights gave no clear improvement in detection performance over uniform weights for the meta-learner was unexpected. Stacked generalization was used to combine the model outputs because this was expected to be the best method of combining the level-0 models by determining the optimal weights for the level-0 models using the training data. However, only a slight trend in increased performance (in terms of mean phi) was observed in lapse detectors based on SP features. In fact, nonlinear feature-based lapse detectors with uniform weights showed a trend towards slightly outperforming the constrained-LS weighted meta-learner. A possible reason for the lack of substantial improvement in detector performance with the use of the stacked approach may be due to the level-0 models being very similar to each other. It has also been suggested that one must use dissimilar predictors to obtain the most improvement in performance when using a stacked system (Breiman 1996).

Investigation of the features and channels that contributed most to the SP-based lapse detector (highest performer of all the detectors) revealed that the alpha band contributed the most to the overall detection model. This is consistent with previous research which has demonstrated correlations (albeit relatively low) between amplitude changes in the alpha range during or immediately after auditory lapses (Makeig and Jung 1996) and visual lapses (Cajochen *et al* 1999). A decrease in alpha power, together with an increase in theta power, has also been reported during EEG microsleeps (Harrison and Horne 1996, Valley and Broughton 1983). It is likely that a change in alpha was 'selected' by our lapse detector as a useful cue to determine the occurrence of lapses. Torsvall and Åkerstedt (1988) noted that alpha power peaked during a ~22 s period preceding the onset of a 'dozing off' event during a visual tracking task. However, the use of alpha band power as a feature in a lapse detection system in a real-life task such as driving is likely to be confounded by the fact that alpha rhythm is suppressed during body movement (Salmelin and Hari 1994). Eoh *et al* (2005) found a relationship between EEG parameters β and $(\alpha + \theta)/\beta$ and mental alertness level. Jap *et al* (2011, 2009) also found evidence to support the $(\alpha + \theta)/\beta$ ratio providing a reliable estimate of fatigue as it yielded larger amplitude differences than θ/β, $\theta/(\alpha + \beta)$ and $(\alpha + \theta)/(\alpha + \beta)$. The use of $(\alpha + \theta)/\beta$ may have increased the performance of our lapse detection system. Coherence in the α-band may be another useful parameter to include in a future spectral-based lapse detection system (Jap *et al* 2010).

As theta power was shown to have one of the highest (albeit small) correlations with lapses in our previous paper (Peiris *et al* 2006), and has also been associated with reduced auditory alertness (Huang *et al* 2001, Jung *et al* 1997), driver fatigue (Lal and Craig 2005) and EEG-microsleeps (Harrison and Horne 1996, Priest *et al* 2001, Valley and Broughton 1983), it was surprising to find that theta power did not contribute substantially to the overall spectral-power-based lapse detector. This suggests that changes in theta power

did not feature prominently in our dataset. Of all the features, beta power contributed the least to the SP detector, contrasting with Belyavin and Wright (1987), who found beta power to be the most useful discriminator of worsening vigilance in a visual vigilance and letter discrimination task. This apparent discrepancy is probably due to beta power being correlated with depth of drowsiness rather than behavioural microsleeps.

Cajochen *et al* (1999) reported that frontal delta+theta EEG activity (1–7 Hz) increased with deteriorating performance in sleep-deprived subjects, whereas Santamaria and Chiappa (1987) reported an increase in centrofrontal alpha, often concurrent with a decrease in occipital alpha, as drowsiness increased. This contrasts with the current findings in which there was an approximately equal contribution towards the overall detection model from all EEG derivations, indicating that there is no strong spatial pattern that could be used to detect lapses (apart from derivation T6-O2 which showed a marginally higher contribution). Again, the apparent differences may be due to the focus of the aforementioned studies being on depth of drowsiness rather than lapses.

A probable substantial contributor to the relatively low detector performance is that the majority of lapses, being of only a few seconds duration, were too brief for substantial changes to develop in the EEG. This is consistent with the EEGer being generally unable to detect visual changes in the raw EEG during lapses on a psychomotor vigilance task (Peiris *et al* 2005).

The tracking task in the current study was not one of driving simulation, let alone on-road driving. Hence, it is not possible to extrapolate the finding of a high rate of lapses (Peiris *et al* 2006) nor the detection of such to real-life driving. This is because a real driving task tends to be more stimulating to the subject and, hence, is more likely to keep their attention focused on the task. Furthermore, perhaps the most important difference between off-road lab-based tasks, such as tracking and driving simulation, and on-road driving is the disparity of the consequences for lapsing. This aspect of lapsing could be investigated further using an on-road driving test, such as done by Papadelis *et al* (2007).

Future work will look more closely at video lapses which do not contain flats spots, to determine whether there is any substantial deterioration in performance, such as incoherent tracking, or if a subject is able to track the target despite appearing video-wise to be having a behavioural microsleep. Conversely, instances where the subject appears to be alert video-wise but shows poor or no tracking performance (i.e. reflected either as flat spots or erratic tracking) need to be investigated as lapses of sustained attention or of task-focused attention to rule out cases of distraction.

Future studies will also include a larger and wider cross-section of the population. Detection models formed from a wider demographic base should generalize better to the population. It will also be desirable to recruit a larger number of subjects to increase the quantity and diversity of data for (i) improved training of a more generalized detector, (ii) improved estimation of the accuracy of detection and (iii) determination of the effects of age, sex, sleep-deprivation etc on lapsing.

Acknowledgments

This work was supported by the New Zealand Foundation for Research Science and Technology, the Christchurch Neurotechnology Research Programme and the University of Canterbury. The authors thank Cathy Lai for her customization of the tracking software.

References

Abasolo D, Hornero R, Gomez C, Garcia M and Lopez M 2006 Analysis of EEG background activity in Alzheimer's disease patients with Lempel-Ziv complexity and central tendency measure *Med. Eng. Phys.* **28** 315–22

Accardo A, Affinito M, Carrozzi M and Bouquet F 1997 Use of the fractal dimension for the analysis of electroencephalographic time series *Biol. Cybern.* **77** 339–50

Aftanas L I and Golocheikine S A 2002 Non-linear dynamic complexity of the human EEG during meditation *Neurosci. Lett.* **330** 143–6

Arjunan S P, Kumar D K and Jung T-P 2009 Changes in decibel scale wavelength properties of EEG with alertness levels while performing sustained attention tasks *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.* vol 31 pp 6288–91

Bahrami B, Seyedsadjadi R, Babadi B and Noroozian M 2005 Brain complexity increases in mania *Neuroreport* **16** 187–91

Belyavin A and Wright N A 1987 Changes in electrical activity of the brain with vigilance *Electroencephalogr. Clin. Neurophysiol.* **66** 137–44

Bishop C M 1995 *Neural Networks for Pattern Recognition* (Oxford: Oxford University Press)

Bittner R, Smrcka P, Vysoký P, Haná K, Pousek L and Scheib P 2000 *Detecting of Fatigue States of a Car Driver* (Berlin: Springer)

Breiman L 1996 Stacked regressions *Mach. Learn.* **24** 49–64

Bruhn J, Ropcke H and Hoeft A 2000 Approximate entropy as an electroencephalographic measure of anesthetic drug effect during desflurane anesthesia *Anesthesiology* **92** 715–26

Bullmore E T, Brammer M J, Bourlon P, Alarcon G, Polkey C E, Elwes R and Binnie C D 1994 Fractal analysis of electroencephalographic signals intracerebrally recorded during 35 epileptic seizures: evaluation of a new method for synoptic visualisation of ictal events *Electroencephalogr. Clin. Neurophysiol.* **91** 337–45

Burioka N, Cornelissen G, Halberg F, Kaplan D T, Suyama H, Sako T and Shimizu E 2003 Approximate entropy of human respiratory movement during eye-closed waking and different sleep stages *Chest* **123** 80–6

Burioka N *et al* 2005 Approximate entropy in the electroencephalogram during wake and sleep *Clin. EEG Neurosci.* **36** 21–4

Cajochen C, Khalsa S B, Wyatt J K, Czeisler C A and Dijk D J 1999 EEG and ocular correlates of circadian melatonin phase and human performance decrements during sleep loss *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **277** 640–9

Davidson P R, Jones R D and Peiris M T R 2007 EEG-based lapse detection with high temporal resolution *IEEE Trans. Biomed. Eng.* **54** 832–9

Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Meth.* **134** 9–21

Dorrian J, Rogers N L and Dinges D F 2005 Psychomotor vigilance performance: neurocognitive assay sensitive to sleep loss *Sleep Deprivation—Clinical Issues, Pharmacology, and Sleep Loss Effects* ed C A Kushida (New York: Dekker)

Elbert T, Ray W J, Kowalik Z J, Skinner J E, Graf K E and Birbaumer N 1994 Chaos and physiology: deterministic chaos in excitable cell assemblies *Physiol. Rev.* **74** 1–47

Elger C E, Widman G, Andrzejak R, Arnhold J, David P and Lehnertz K 2000 Nonlinear EEG analysis and its potential role in epileptology *Epilepsia* **41** Suppl 3 S34–8

Eoh H J, Chung M K and Kim S-H 2005 Electroencephalographic study of drowsiness in simulated driving with sleep deprivation *Int. J. Ind. Ergon.* **35** 307–20

Ferri R, Alicata F, Del Gracco S, Elia M, Musumeci S A and Stefanini M C 1996 Chaotic behavior of EEG slow-wave activity during sleep *Electroencephalogr. Clin. Neurophysiol.* **99** 539–43

Fisher R A 1936 The use of multiple measurements in taxonomic problems *Ann. Eugen.* **7** 179–88

Freund D M, Knipling R R, Landsburg A C, Simmons R R and Thomas G R 1995 A holistic approach to operator alertness research *Proc. Transport. Res. Board 74th Annu. Meeting.* (Washington, DC)

Golz M, Sommer D, Chen M, Trutschel U and Mandic D 2007 Feature fusion for the detection of microsleep events *J. VLSI Signal Proc.* **49** 329–42

Grace R 2001 Drowsy driver monitor and warning system *Proc. Int. Driving Symp. Hum. Factors Driver Assess. Training Vehicle Des.* vol 2 pp 64–9

Grace R, Byrne V E, Bierman D M, Legrand J M, Gricourt D and Davis B K 1998 A drowsy driver detection system for heavy vehicles *Proc. AIAA/IEEE/SAE Dig. Avionics Systems Conf.*vol 17 pp I36/1–8

Grassberger P and Procaccia I 1983 Estimation of the Kolmogorov entropy from a chaotic signal *Phys. Rev.* A **28** 2591–3

Harrison Y and Horne J A 1996 Occurrence of 'microsleeps' during daytime sleep onset in normal subjects *Electroencephalogr. Clin. Neurophysiol.* **98** 411–6

Higuchi T 1988 Approach to an irregular time series on the basis of the fractal theory *Physica* D **31** 277–83

Huang R, Tsai L and Kuo C J 2001 Selection of valid and reliable EEG features for predicting auditory and visual alertness levels *Proc. Nat. Sci. Council* **25** 17–25

Jap B T, Lal S and Fischer P 2010 Inter-hemispheric electroencephalography coherence analysis: assessing brain activity during monotonous driving *Int. J. Psychophysiol.* **76** 169–73

Jap B T, Lal S and Fischer P 2011 Comparing combinations of EEG activity in train drivers during monotonous driving *Expert Sys. Appl.* **38** 996–1003

Jap B T, Lal S, Fischer P and Bekiaris E 2009 Using EEG spectral components to assess algorithms for detecting fatigue *Expert Sys. Appl.* **36** 2352–9

Johns M W 2003 The amplitude-velocity ratio of blinks: a new method for monitoring drowsiness *Sleep* **26** (Suppl.) A51–2

Jones R D 2006 Measurement of sensory-motor control performance capacities: tracking tasks *The Biomedical Engineering Handbook—Biomedical Engineering Fundamentals* ed J D Bronzino (Boca Raton, FL: CRC Press)

Jones R D and Donaldson I M 1986 Measurement of sensory-motor integrated function in neurological disorders: three computerised tracking tasks *Med. Biol. Eng. Comput.* **24** 536–40

Jung T, Makeig S, Stensmo M and Sejnowski T 1997 Estimating alertness from the EEG power spectrum *IEEE Trans. Biomed. Eng.* **44** 60–9

Jung T-P, Makeig S, Humphries C, Lee T-W, Mckeown M J, Iragui V and Sejnowski T J 2000 Removing electroencephalographic artifacts by blind source separation *Psychophysiology* **37** 163–78

Kiefte M 1999 Discriminant analysis toolbox MATLAB ver. 0.3 (Natwick, MA: Mathworks)

Klem G H, Luders H O, Jasper H H and Elger C 1999 The ten-twenty electrode system of the international federation *Electroencephalogr. Clin. Neurophysiol. Suppl.* **52** 3–6

Klonowski W, Olejarczyk E, Stepien R, Jalowiecki P and Rudner R 2006 Monitoring the depth of anaesthesia using fractal complexity method *Complexus Mundi. Emergent Patterns in Nature* ed M N Novak (Hackensack, NJ: World Scientific) pp 333–42

Kobayashi T, Madokoro S, Wada Y, Misaki K and Nakagawa H 2001 Human sleep EEG analysis using the correlation dimension *Clin. Electroencephalogr.* **32** 112–8

Lal S K L and Craig A 2002 Driver fatigue: Electroencephalography and psychological assessment *Psychophysiology* **39** 313–21

Lal S K L and Craig A 2005 Reproducibility of the spectral components of the electroencephalogram during driver fatigue *Int. J. Psychophysiol.* **55** 137–43

Le Van Quyen M, Martinerie J, Navarro V, Boon P, D'have M, Adam C, Renault B, Varela F and Baulac M 2001 Anticipation of epileptic seizures from standard EEG recordings *Lancet* **357** 183–8

Lehnertz K 1999 Non-linear time series analysis of intracranial EEG recordings in patients with epilepsy—an overview *Int. J. Psychophysiol.* **34** 45–52

Lempel A and Ziv J 1976 On the complexity of finite sequences *IEEE Trans. Inform. Theory* **22** 75–81

Lin C-T, Ko L-W, Chung I-F, Huang T-Y, Chen Y-C, Jung T-P and Liang S-F 2006 Adaptive EEG-based alertness estimation system by using ICA-based fuzzy neural networks *IEEE Trans. Circuits Syst.* **53** 2469–76

Lin C-T, Wu R-C, Jung T-P, Liang S-F and Huang T-Y 2005a Estimating driving performance based on EEG spectrum analysis *EURASIP J. Appl. Sig. Process.* **19** 3165–74

Lin C-T, Wu R-C, Liang S-F, Chao W-H, Chen Y-J and Jung T-P 2005b EEG-based drowsiness estimation for safety driving using independent component analysis *IEEE Trans. Circuits Syst.* **52** 2726–38

Lutzenberger W, Preissl H and Pulvermuller F 1995 Fractal dimension of electroencephalographic time series and underlying brain processes *Biol. Cybern.* **73** 477–82

Makeig S and Inlow M 1993 Lapses in alertness: coherence of fluctuations in performance and EEG spectrum *Electroencephalogr. Clin. Neurophysiol.* **86** 23–35

Makeig S and Jung T 1995 Changes in alertness are a principal component of variance in the EEG spectrum *NeuroRep.* **7** 213–6

Makeig S and Jung T-P 1996 Tonic, phasic, and transient EEG correlates of auditory awareness in drowsiness *Brain Res. Cogn. Brain Res.* **4** 15–25

Makeig S, Jung T P and Sejnowski T 1996 Using feedforward neural networks to monitor alertness from changes in EEG correlation and coherence *Advances in Neural Information Processing Systems* ed M Touretzky, M Mozer and M Hasselmo (Cambridge, MA: MIT Press)

Matousek M and Petersen I 1983 A method for assessing alertness fluctuations from EEG spectra *Electroencephalogr. Clin. Neurophysiol.* **55** 108–13

Naidu P S 1996 *Modern Spectral Analysis of Time Series* (Boca Raton, FL: CRC Press)

Natarajan K, Acharya R, Alias F, Tiboleng T and Puthusserypady K 2004 Nonlinear analysis of EEG signals at different mental states *Biomed. Eng. Online* **3** 1–11

Padilla M L, Wood R A, Hale L A and Knight R T 2006 Lapses in a prefrontal-extrastriate preparatory attention network predict mistakes *J. Cogn. Neurosci.* **18** 1477–87

Papadelis C, Chen Z, Kourtidou-Papadeli C, Bamidis P D, Chouvarda I and Bekiaris E 2007 Monitoring sleepiness with on-board electrophysiological recordings for preventing sleep-deprived traffic accidents *Clin. Neurophysiol.* **118** 1906–22

Peiris M T R, Jones R D, Davidson P R, Carroll G J and Bones P J 2006 Frequent lapses of responsiveness during an extended visuomotor tracking task in non-sleep-deprived subjects *J. Sleep Res.* **15** 291–300

Peiris M T R, Jones R D, Davidson P R, Carroll G J, Signal T L, Parkin P J, Van Den Berg M and Bones P J 2005 Identification of vigilance lapses using EEG/EOG by expert human raters *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.* vol 27 pp 5736–8

Pincus S 1995 Approximate entropy (ApEn) as a complexity measure *Chaos* **5** 110–7

Pincus S M, Gladstone I M and Ehrenkranz R A 1991 A regularity statistic for medical data analysis *J. Clin. Monit.* **7** 335–45

Priest B, Brichard C, Aubert G, Liistro G and Rodenstein D O 2001 Microsleep during a simplified maintenance of wakefulness test: a validation study of the OSLER test *Am. J. Resp. Crit. Care Med.* **163** 1619–25

Radhakrishnan N and Gangadhar B N 1998 Estimating regularity in epileptic seizure time-series data. A complexity measure approach *IEEE Eng. Med. Biol.* **17** 89–94

Salmelin R and Hari R 1994 Spatiotemporal characteristics of sensorimotor neuromagnetic rhythms related to thumb movement *Neuroscience* **60** 537–50

Santamaria J and Chiappa K H 1987 The EEG of drowsiness in normal adults *J. Clin. Neurophysiol.* **4** 327–82

Schleicher R, Galley N, Briest S and Galley L 2008 Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics* **51** 982–1010

Sheskin D 1997 *Handbook of Parametric and Nonparametric Statistical Procedures* (Boca Raton, FL: CRC Press)

Sing T, Sander O, Beerenwinkel N and Lengauer T 2005 ROCR: visualizing classifier performance in R *Bioinformatics* **21** 3940–1

Sommer D, Golz M, Schnupp T, Krajewski J, Trutschel U and Edwards D 2009 A measure of strong driver fatigue *Proc. Int. Driving Symp. Hum. Factors Driver Assess. Training Vehicle Des.* vol 5 pp 9–15

Sommer D, Hink T and Golz M 2002 Application of learning vector quantization to detect drivers dozing-off *Eur. Symp. Intelligent Technologies Hybrid Systems and Implementation on Smart Adaptive Systems* vol 2 pp 275–9

Subasi A 2005 Application of classical and model-based spectral methods to describe the state of alertness in EEG *J. Med. Syst.* **29** 473–86

Tononi G, Sporns O and Edelman G M 1994 A measure for brain complexity: relating functional segregation and integration in the nervous system *Proc. Natl Acad. Sci. USA* **91** 5033–7

Torsvall L and Åkerstedt T 1988 Extreme sleepiness: quantification of EOG and spectral EEG parameters *Int. J. Neurosci.* **38** 435–41

Valley V and Broughton R 1983 The physiological (EEG) nature of drowsiness and its relation to performance deficits in narcoleptics *Electroencephalogr. Clin. Neurophysiol.* **55** 243–51

Van Orden K F, Jung T-P and Makeig S 2000 Combined eye activity measures accurately estimate changes in sustained visual task performance *Biol. Psychol.* **52** 221–40

Weissman D H, Roberts K C, Visscher K M and Woldorff M G 2006 The neural bases of momentary lapses in attention *Nat. Neurosci.* **9** 971–8

Wierwille W W and Ellsworth L A 1994 Evaluation of driver drowsiness by trained raters *Accid. Anal. Prev.* **26** 571–81

Witten I H and Frank E 2000 *Data Mining—Practical Machine Learning Tools and Techniques with Java Implementations* (San Francisco, CA: Morgan Kaufmann Publishers)

Wolpert D 1992 Stacked generalization *Neural Netw.* **5** 241–59

Yaylali I, Kocak H and Jayakar P 1996 Detection of seizures from small samples using nonlinear dynamic system theory *IEEE Trans. Biomed. Eng.* **43** 743–51

Zhang X and Roy R J 2001 Derived fuzzy knowledge model for estimating the depth of anesthesia *IEEE Trans. Biomed. Eng.* **48** 312–23

Zhang X S and Roy R J 1999 Predicting movement during anaesthesia by complexity analysis of electroencephalograms *Med. Biol. Eng. Comput.* **37** 327–34

Zhang X S, Zhu Y S, Thakor N V and Wang Z Z 1999 Detecting ventricular tachycardia and fibrillation by complexity measure *IEEE Trans. Biomed. Eng.* **46** 548–55