Prediction of Microsleeps Using Pairwise Joint Entropy and Mutual Information between EEG channels

Abdul Baseer, *Student Member, IEEE*, Stephen J. Weddell, *Senior Member, IEEE*, Richard D. Jones, *Fellow, IEEE*

Abstract— **Microsleeps are involuntary and brief instances of complete loss of responsiveness, typically of 0.5–15 s duration. They adversely affect performance in extended attention-driven jobs and can be fatal. Our aim was to predict microsleeps from 16 channel EEG signals. Two information theoretic concepts – pairwise joint entropy and mutual information – were independently used to continuously extract features from EEG signals. k-nearest neighbor (kNN) with k = 3 was used to calculate both joint entropy and mutual information. Highly correlated features were discarded and the rest were ranked using Fisher score followed by an average of 3-fold cross-validation area under the curve of the receiver operating characteristic (AUC_{ROC}). Leave-one-out method (LOOM) was performed to test the performance of microsleep prediction system on independent data. The best prediction for 0.25 s ahead was AUC_{ROC}, sensitivity, precision, geometric mean (GM), and ϕ of 0.93, 0.68, 0.33, 0.75, and 0.38 respectively with joint entropy using single linear discriminant analysis (LDA) classifier.**

I. Introduction

Microsleeps are brief instances of complete and unintentional loss of responsiveness that typically last between 0.5–15 s in which even a non-sleep-deprived person momentarily falls asleep [1, 2]. They are generally cued from eye closure, droopy eyes, eye blinks, head nodding, and absent visuomotor performance [1, 3]. They adversely affect human performance in different sectors and in some cases, can be fatal on extended-attention monotonous activities such as driving, piloting, and air traffic control.

Traditionally, power spectral features have been used for EEG-based microsleep detection [3-5]. Davidson et al. [3] achieved an AUC_{ROC} of 0.81 and ϕ of 0.38 with log-power spectral features, principle component analysis (PCA) for dimensionality reduction, and long-short-term-memory (LSTM) recurrent neural networks (RNN) for classification. Peiris et al. [4] achieved an AUC_{ROC} of 0.86 and ϕ of 0.39 with the same feature extraction and reduction techniques but with stacking of 6 LDA classifiers. Ayyagari et al. [5] used the same approach but with stacked echo state networks (ESN) with leaky neurons to achieve ϕ of 0.51, sensitivity of 0.85, and specificity of 0.94. An epoch length of 2 s with 50% overlap was used in these studies.

Shoorangiz et al. [6] rigorously revised the gold standard by taking account of tracking velocity, the transition between states, and uncertain segments. They also extracted log-power spectral features from different EEG frequency bands but used mutual information-based greedy forward feature selection algorithm to overcome the curse of dimensionality. They used synthetic minority oversampling (SMOTE) and adaptive synthetic sampling (ADASYN) to reduce the effect of data imbalance and compared their performance on a single LDA classifier with no resampling. Their best prediction for 0.25 s ahead was an AUC_{ROC} of 0.90 and ϕ of 0.33 with resampling and SMOTE, whereas the best GM of 0.74 was achieved with SMOTE and the best precision of 0.37 was achieved with no resampling. These studies used independently log-power spectral features extracted from individual EEG channels.

We were motivated by the fact that EEG signals are generally multivariate and synchronously recorded, there is information in the inter-channel relationships that may be advantageous in the prediction of brain states over independently extracted features from individual channels. Neural interactions are transient and inherently non-stationary [7]. Joint entropy is a measure of uncertainty associated with two random variables. It is the sum of log of joint probabilities of random variables and regarded as their joint information. Mutual information estimates linear and nonlinear dependencies between two random variables [8]. It is the sum of marginal entropies discounted by joint entropy or the sum of log ratio of joint probabilities of random variables to their marginal probabilities. Mutual information is also interpreted as the average number of bits of one variable, X, that can be predicted by another variable Y and vice versa. Bonita et al. [9] reported that, with EEG signals, mutual information at group level gave statistically significant separation between the two behavioural states of eyes open and eyes closed. They found that, even with small data lengths, typically of 1 s (1000 data points), mutual information was more robust to noise than correlations in time, such as Pearson, Spearman, and Kendall, using balanced data (number of both states the same). However, Quiroga et al. [10] compared the performance of several synchronization measures on rat EEG and concluded that with 1000 data points (5 s) mutual information could not produce robust estimates of synchronization in all three cases. Except for mutual information, phase synchronization, cross correlation, and coherence function produced

* This work was supported by the University of Canterbury and the New Zealand Brain Research Institute.

Abdul Baseer is with Department of Electrical and Computer Engineering at University of Canterbury, Christchurch Neurotechnology Research Programme, and New Zealand Brain Research Institute, Christchurch, New Zealand, (e-mail: abdul.buriro@nzbri.org).

Stephen Weddell is with the Department of Electrical and Computer Engineering at University of Canterbury, and Christchurch Neurotechnology Research Programme, Christchurch, New Zealand, (e-mail: steve.weddell@canterbury.ac.nz).

Richard Jones is with the Department of Electrical and Computer Engineering at University of Canterbury, Christchurch Neurotechnology Research Programme, and New Zealand Brain Research Institute, Christchurch, New Zealand, (e-mail: richard.jones@nzbri.org).

978-1-5090-2809-2/17/$31.00 ©2017 IEEE

qualitatively similar results [10]. Blinowska [11] considered that mutual information works well in ideal conditions. However, the practicality of extracting mutual information from experimental data is limited by systematic errors and requires a large amount of data [7].

To the best of our knowledge, both joint entropy and mutual information have not been used at epoch level to extract features from EEG to predict microsleeps.

The aim of this study was to use two information theoretic concepts, i.e., joint entropy and mutual information between EEG channels, to improve the accuracy of microsleep prediction.

II. Methods

A. Data

Fifteen non-sleep-deprived healthy subjects aged 18–36 years (mean 26.5) were recruited. None of the subjects reported any neurological or sleep disorder. All subjects slept normally and the average duration of sleep for the previous night was 7.8±1.2 h [2].

Each subject performed a 1-D preview tracking task in two 1-hour sessions, one week apart. The task was to keep the tracking cursor as close as possible to the pseudorandom target. During the task, EEG, facial video, and tracking error were recorded. EEG was sampled at 256 Hz from 16 channels placed per international 10–20 system, namely Fp1, Fp2, F3, F4, F7, F8, C3, C4, O1, O2, P3, P4, T3, T4, T5, and T6. Tracking performance was recorded at 64 Hz and facial video at 25 fps.

B. Gold Standard

Tracking task performance and video ratings were used to develop a gold standard. A responsive state corresponds to coherent tracking with the target regardless of video ratings. A microsleep was defined as mean absolute error > 3 cm lasting for more than 1 s or a drop in tracking velocity up to 10th percentile of target velocity in combination with a video rating of deep drowsy or lapse [6].

The gold standard was used in both the feature selection and classification stage as shown in Fig. 1, and in the evaluation of prediction performance.

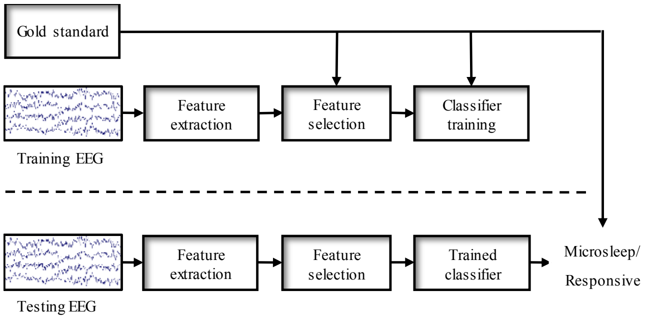

Fig. 1. Schematic of microsleep detection and prediction system.

A. Joint Entropy

For continuous-time bivariate random variables (*X*, *Y*), joint Shannon entropy *H* is defined as

$$H(X,Y) = -\int\int p(x,y)\ln p(x,y)dxdy, \qquad (1)$$

where *H* is the entropy, $p(x,y)$ is the joint probability density function (pdf) between *X* and *Y*. For discrete data, the integrals reduce to summation and probabilities can be estimated by counting samples. However, for continuous data, pdfs are estimated. Singh et at. [12] estimated entropy based on the k^{th} nearest neighbour distance between *N* sample points as

$$\widehat{H}(X) = \ln(N) - \psi(k) + \ln(v) + \frac{d}{N}\sum_{i=1}^{N}\log\epsilon_i, \qquad (2)$$

where, N shows the number of sample points, ψ is digamma function, k is the k^{th} nearest neighbour, d is the dimension of a random variable, v is the volume of d-dimensional unit ball, and ϵ_i is the distance between x_i and its k^{th} neighbourhood. The volume of the d-dimensional unit ball is calculated as

$$v = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}, \qquad (3)$$

where, Γ is the gamma function.

A. Mutual Information

Mutual information is denoted by *I* and defined as

$$I(X,Y) = -\int\int p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}dxdy, \qquad (4)$$

where, $p(x)$ and $p(y)$ is marginal pdf of *X* and *Y* respectively. In term of Shannon's entropies, mutual information can be expressed as

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \qquad (5)$$

where, $H(X)$ is the marginal entropy of *X* and $H(Y)$ is the marginal entropies of *Y*, and $H(X,Y)$ is their joint entropy.

A. Feature Extraction

Artefact subspace reconstruction (ASR) following band-pass filtering of 1–45 Hz was used to remove artefacts from EEG [6]. The preprocessed EEG signals were then decomposed into overlapping EEG subbands, i.e., delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz) following common average reference of channels. All EEG signals were decimated to 128 Hz to reduce processing time. EEG signals from each subband were segmented to 5 s epochs and steps of 0.25 s to account for shortest microsleep of 250 ms. Among different estimators of MI, Khan et al. [8] reported that kNN outperforms across all noise levels and for a small number of data points. We selected k = 3 to best account for the issue of bias-variance trade-off. Joint entropy and mutual information between each pair of EEG channel were calculated for each epoch and every subband. EEG with 16 channels results in 120 distinct channel pairs per subband. Thus, for a single epoch, there were 600 features for each joint entropy and mutual information. The ITE toolbox [13] was used for estimation of marginal and joint entropies which subsequently were used in estimating mutual information.

A. Feature Selection

A total of 600 features might have demanded the higher classifier complexity, and consequently, a high variance problem might exist – the curse of dimensionality. To overcome this, we selected the best features from the training data set in three steps: (1) calculated Pearson correlation coefficient r between features and removed highly redundant features if $|r| > 0.9$, (2) ranked the remaining features by Fisher scores and sorted in descending order (a large Fisher score corresponds to a large mean difference and a small overlap, and consequently, good separation between two classes), and (3) 3-fold cross-validated the top ranked-feature using LDA classifier and saved the average value of $\mathrm{AUC_{ROC}}$. The next feature was combined with the feature selected in the last iteration and was selected if the combination improved the $\mathrm{AUC_{ROC}}$, otherwise, it was discarded. The process was repeated and at every iteration, based on $\mathrm{AUC_{ROC}}$, a feature was either added or discarded. This ensured that features were only selected for which the combination improves the classification performance and resulted in 62 and 146 features of joint entropy and mutual information respectively.

A. Classification

A single LDA classifier was separately fed with joint entropy and mutual information features to predict microsleep states. Standard classifier performs better when the prevalence of different classes is approximately equal by minimizing the error [14, 15]. However, a simple algorithm-based approach referred to as cost-sensitive learning [16] was used by adjusting the decision threshold to account for average class imbalance ratio (microsleep: responsive) of about 1:137. LOOM was performed on the data to train the classifier and to test on independent subjects. Prediction of microsleeps was 0.25 s ahead of the gold standard as shown in Fig. 2.

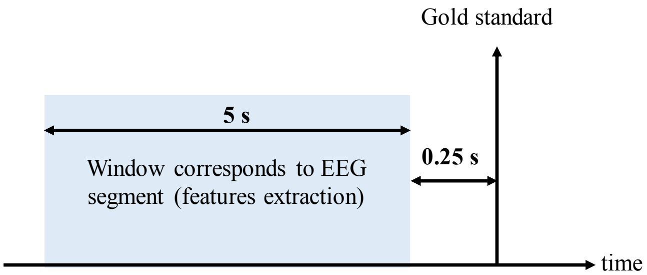

Fig. 2. Prediction of gold standard from corresponding epoch.

III. Results

Our analysis was limited to the 8 subjects who having had at least one definite microsleep over the two sessions. The data of one subject (both sessions) was held for testing and evaluation of prediction performance. The feature selection and training was done on the concatenated data of the other seven subjects. The cross-validation performance during the process of feature selection of both joint entropy and mutual information is shown Fig. 3. The classification process was repeated 8 times and different measures of performance were averaged for both joint entropy and mutual information. Prediction of the microsleep states on independent test data is presented in the TABLE I. Joint entropy on average gave substantially better $\mathrm{AUC_{ROC}}$, sensitivity, specificity, GM, precision, and ϕ than mutual information. Furthermore, on average, mutual information needed 146 features compared 62 features for joint entropy. Also, mutual information depends on both marginal entropies and joint entropy and, therefore, requires more computation than joint entropy.

TABLE I. Microsleep Prediction Performance

	Joint Entropy	Mutual Information
$\mathrm{AUC_{ROC}}$	0.93	0.81
Sensitivity	0.68	0.59
Specificity	0.90	0.83
Precision	0.33	0.22
GM	0.75	0.70
Phi (ϕ)	0.38	0.23

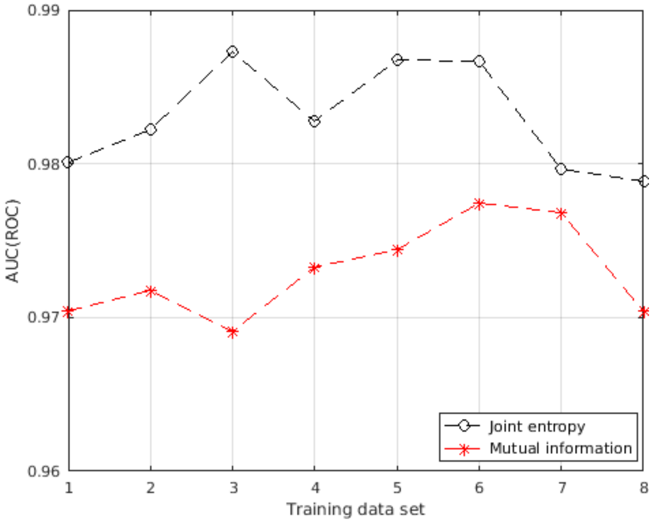

Fig. 3. Comparison of $\mathrm{AUC_{ROC}}$ of joint entropy and mutual information features at cross validation stage during feature selection.

Except for $\mathrm{AUC_{ROC}}$, all of the above mentioned performance measures are based on a threshold which in turn depends on class sizes. Therefore, average of such performance measures for imbalance data may be misleading as a classifier sets threshold on training data that have substantially different class sizes than test data.

We, therefore, present test performance of independent subjects with their class imbalance ratio for joint entropy and mutual information in Fig. 4 and Fig. 5 respectively.

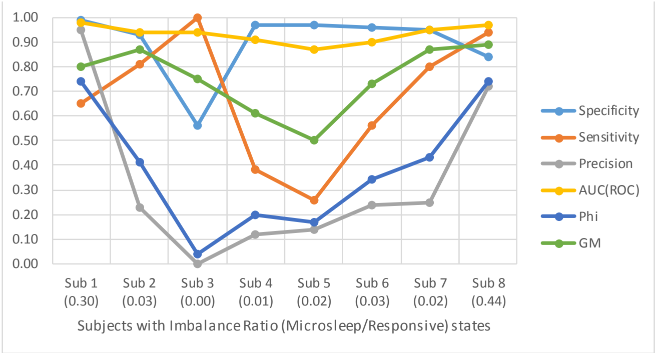

Fig. 4. Performance of joint entropy features against imbalance ratio.

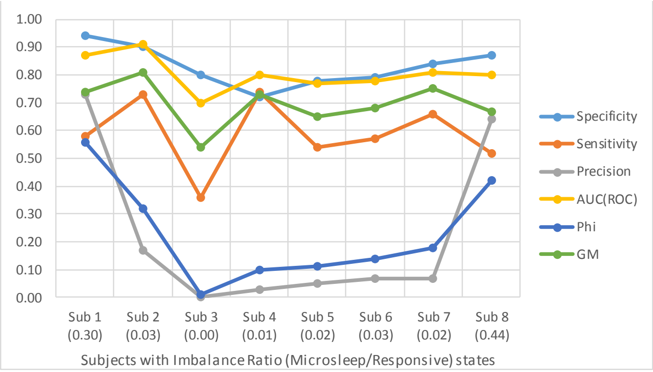

Fig. 5. Performance of mutual information features against imbalance ratio.

IV. Discussion

The cross-validation AUC_{ROC} during the process of feature selection and test AUC_{ROC} of mutual information was poorer than joint entropy. While on average, joint entropy performed substantially better than mutual information in all measures of performance and with small number of features. However, mutual information was slightly more consistent in specificity, and GM than joint entropy across all imbalance ratios.

Pairwise joint entropy of T6-O2 in theta frequency band showed the highest single-feature discrimination power in terms of Fisher score and AUC_{ROC} with average of 23.8 nats for responsive and 22.9 nats for microsleep states for all 8 iterations of training data. O1-O2 was dominant in all training iterations in which the training data contained only a small number of microsleeps with average of 23.6 and 22.7 nats for responsive and microsleep states respectively in theta frequency band. This indicates that temporal and occipital regions of the brain become more active during the responsive state as more neuronal communication (information) in delta frequency band occurs when a person is responsive and less communication during the microsleep states. However, there was no consistent electrode pair and frequency band for mutual information.

This indicates that mutual information on average contains less information than joint entropy presumably due to subtraction of joint entropy from sum of marginal entropies. Shared information (e.g., joint entropy, cross spectral powers) irrespective of relationship, account for the contribution of individual channels/sources and therefore represent the brain states more adequately.

Use of joint entropy and mutual information is, therefore, application dependent. For statistical group analysis to differentiate two groups in a large population, mutual information may be a preferable choice. However, joint entropy appears preferable over mutual information for continuously extracting features at epoch level from small number of data points.

V. Conclusion

Microsleep states were predicted one-step-ahead, i.e., 0.25 s prior to the gold standard. We tested both joint entropy and mutual information independently using a three-step feature selection method on a single LDA classifier. We could achieve an AUC_{ROC} of 0.93 that is the best reported performance for EEG-based microsleep prediction. Joint entropy compared to [6] gave better sensitivity, precision, and ϕ but substantial improvements are still needed for real-time real world implementation.

Future work will be focused more on feature extraction techniques that require less processing time and contain better information on brain states relating to microsleeps, e.g., asymmetric features. Fusion of orthogonal features containing non-redundant information.

References

[1] R. D. Jones, G. R. Poudel, C. R. H. Innes, P. R. Davidson, M. T. R. Peiris, A. M. Malla, *et al.*, "Lapses of responsiveness: characteristics, detection, and underlying mechanisms," presented at the 32nd Annual International Conference of the IEEE EMBS, Buenos Aires, Argentina, 2010.

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

[3] P. R. Davidson, R. D. Jones, and M. T. R. Peiris, "EEG-based lapse detection with high temporal resolution," *IEEE Trans. Biomed. Eng.,* vol. 54, pp. 832-839, May 2007.

[4] M. T. R. Peiris, P. R. Davidson, P. J. Bones, and R. D. Jones, "Detection of lapses in responsiveness from the EEG," *J. Neural Eng.,* vol. 8, Feb 2011.

[5] S. S. D. P. Ayyagari, R. D. Jones, and S. J. Weddell, "Optimized echo state networks with leaky integrator neurons for EEG-based microsleep detection," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, Aug. 2015, pp. 3775-3778.

[6] R. Shoorangiz, S. J. Weddell, and R. D. Jones, "Prediction of microsleeps from EEG: Preliminary results," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, Aug. 2016, pp. 4650-4653.

[7] E. Pereda, R. Q. Quiroga, and J. Bhattacharya, "Nonlinear multivariate analysis of neurophysiological signals," *Prog. Neurobiol.,* vol. 77, pp. 1-37, Sep-Oct 2005.

[8] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson, V. Protopopescu*, et al.*, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Phys. Rev. E,* vol. 76, Aug 2007.

[9] J. D. Bonita, L. C. C. Ambolode, B. M. Rosenberg, C. J. Cellucci, T. A. A. Watanabe, P. E. Rapp*, et al.*, "Time domain measures of inter-channel EEG correlations: a comparison of linear, nonparametric and nonlinear measures," *Cogn. Neurodyn.,* vol. 8, pp. 1-15, Feb 2014.

[10] R. Q. Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger, "Performance of different synchronization measures in real data: A case study on electroencephalographic signals," *Phys. Rev. E,* vol. 65, Apr 2002.

[11] K. J. Blinowska, "Review of the methods of determination of directed connectivity from multichannel data," *Med. Biol. Eng. Computi.,* vol. 49, pp. 521-529, May 2011.

[12] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest Neighbor Estimates of Entropy," *Amer. J. Math. Management Sci.,* vol. 23, pp. 301-321, Feb 2003.

[13] Z. Szabo, "Information Theoretical Estimators Toolbox," *J. Mach. Learn. Rese.,* vol. 15, pp. 283-287, Jan 2014.

[14] C. Elkan, "The foundations of cost-sensitive learning," in *Conf. Proc. IJCIA*, Seattle, WA, USA, Aug. 2001, pp. 973-978.

[15] J. J. Chen, C. A. Tsai, H. Moon, H. Ahn, J. J. Young, and C. H. Chen, "Decision threshold adjustment in class prediction," *SAR QSAR Environ. Res.,* vol. 17, pp. 337-352, Jun 2006.

[16] W. J. Lin and J. J. Chen, "Class-imbalanced classifiers for high-dimensional data," *Brief. Bioinform.,* vol. 14, pp. 13-26, Jan 2013.