Prediction of microsleeps using EEG inter-channel relationships

Abdul Baseer Buriro

A thesis presented for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering
at the
University of Canterbury,
Christchurch, New Zealand.

April 2019

ABSTRACT

A microsleep is a brief and involuntary sleep-related loss of consciousness of up to 15 s during an active and attention-demanding task. Such episodes of unresponsiveness are of particularly high importance in people who perform high-risk and monotonous activities requiring extended-attention and unimpaired visuomotor performance, such as car and truck drivers, train drivers, pilots, and air-traffic controllers, where microsleeps can, and do, result in catastrophic accidents and fatalities. Microsleep-related accidents can potentially be avoided and thereby lives saved, if microsleeps are noninvasively and accurately predicted.

The aim of this study was to explore various inter-channel relationships in the electroencephalogram (EEG) for detection/prediction of microsleeps. In addition to feature-level and decision-level data fusion techniques, ensemble classification techniques were investigated to improve microsleep detection/prediction accuracies.

The data used in this research were from a previous study in which 15 healthy non-sleep-deprived participants performed two 1-h sessions of 1-D continuous tracking task. Eight subjects were included for the analyses, all of whom experienced at least one microsleep during the two sessions of the task. A gold-standard was formed by integrating two independent measures of the video-rating from a human expert and tracking performance. An average duration of microsleep was 2.16 min/h.

Three univariate feature sets, namely variance, wavelet spectral power, and entropy features were extracted from individual channels of EEG. Seven bivariate and symmetric feature sets were directly extracted from pairs of EEG channels. Three of the 7 feature sets were non-normalized (i.e., covariance, wavelet cross-spectral power, joint entropy) and 4 were normalized (i.e., Pearson's correlation coefficient, wavelet coherence, mutual information, phase synchronization index (PSI)). Eleven feature sets were extracted from coefficients of a multivariate autoregressive (MVAR) model, fitted to the EEG time series. Five of the 11 feature sets were non-causal (i.e., cross-spectral power, coherence, imaginary part of the coherency (iCOH), partial coherence (pCOH), PSI), whereas 6 were causal (i.e., partial directed coherence (PDC), generalized PDC (GPDC), directed transfer function (DTF), normalized directed transfer function (nDTF), direct DTF (dDTF), full frequency DTF (ffDTF)). All features were extracted from a window length of 5 s of EEG . Feature sets extracted from EEG time series of each session were demeaned with respect to mean of their first 2 min features.

Two supervised feature-selection techniques were used to select discriminatory and relevant

features from each feature set. The first technique was classifier-dependent feature-selection, which comprised Pearson's correlation coefficient-based filter followed by Fisher's score-based ranking and a linear discriminant analysis (LDA)-based wrapper. The second technique was classifier-independent, which was a combination of relevance-based ranking (i.e., mutual information between a feature and class labels) and relevance-based sequential forward selection (SFS) method.

Two linear classifiers, namely LDA and linear support vector machine (LSVM), were used to validate the efficacy of a feature set at detecting/predicting microsleeps.

To improve detection/prediction accuracies of microsleeps, all 21 feature sets were fused at feature-level and decision-level. Two-stage feature-selection was performed in feature-level fusion and the homogeneous (same type) classifiers were used in decision-level fusion.

The data of each subject was assumed to be an individual cluster. Three ensemble classification techniques were proposed and, in addition to soft majority voting, compared with single classifier. In one ensemble classification technique, individual clusters, using divergence, were combined to form overlapping clusters.

Non-normalized features performed better than normalized and causal features. Baseline correction (demeaning) of non-normalized features substantially improved the performance metrics on microsleep states and onsets. For most of the feature types, performances of classifier-dependent feature-selection was better than classifier-independent. With classifier-dependent feature-selection, LDA classifier, compared to LSVM, gave slightly higher performance.

Overall, joint entropy was the best single feature set. Data fusion was no better than best performing single feature set. All of the ensemble techniques resulted in overall improved performances over single classifier.

The overall highest detection performance metrics (phi = 0.48, area under the curve of precision recall (AUC_{PR}) = 0.51, area under the curve of receiver operating characteristic (AUC_{ROC}) = 0.95) and (phi = 0.11, AUC_{PR} =0.09, AUC_{ROC} = 0.91) on microsleep states and onsets, respectively using single feature sets were achieved with joint entropy features.

Compared to the best performing single feature set, feature-level fusion resulted in similar detection performance metrics for microsleep states of (phi = 0.49, AUC_{PR}R = 0.49, AUC_{ROC} = 0.96) and for microsleep onsets of (phi = 0.10, AUC_{PR} = 0.09, AUC_{ROC} = 0.92).

Overlapping clusters resulted in the overall highest detection performance metrics of (phi = 0.50, AUC_{PR} = 0.52, AUC_{ROC} = 0.96) and (phi = 0.12, AUC_{PR} =0.10, AUC_{ROC} = 0.93) on microsleep states and onsets, respectively.

We report the overall highest microsleep (state and onset) detection and prediction performance metrics achieved on this data set. Our results are a way forward in using EEG inter-channel relationships as features at detecting/predicting microsleeps. Contrary to the literature on other classification tasks, our results favour non-normalized inter-channel features at classifying microsleeps from responsiveness. Our results do not support data fusion of single modality (e.g.,

EEG) data in classification of microsleeps. All of the proposed ensemble techniques improved detection/prediction accuracies of microsleeps, using EEG, over a single classifier.

Notwithstanding our achievements, the performances remain too low in general for real-life applications.

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from which the extract comes:
Section 6.2.1–Section 6.2.2, Section 6.3, and Section 6.4.1 are adapted from an accepted paper: Buriro, A.B., Shoorangiz R., Weddell, S.J., and Jones, R.D. (2018), 'Predicting microsleep states using inter-channel EEG relationships', In IEEE Transaction on Neural Systems and Rehabilitation Engineering (TNSRE), in Press.

Please detail the nature and extent (%) of contribution by the candidate:
The PhD candidate did the research and wrote all sections of the published manuscript. The co-authoring supervisory team suggested research directions and offered criticisms which the candidate worked into the manuscript.

Certification by co-authors

If there is more than one co-author then a single co-author can sign on behalf of all. The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work, he or she wrote the text.

Name: *Professor Richard D. Jones* Signature: Date: *04 November 2018*

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from which the extract comes:
Section 6.7.4 is adapted from a published conference paper: Buriro, A.B., Shoorangiz R., Weddell, S.J., and Jones, R.D. (2018), 'Ensemble learning based on overlapping clusters of subjects to predict microsleep states from EEG', In Proceedings of the 40th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3036–3039.

Please detail the nature and extent (%) of contribution by the candidate:
The PhD candidate did the research and wrote all sections of the published manuscript. The co-authoring supervisory team suggested research directions and offered criticisms which the candidate worked into the manuscript.

Certification by co-authors

If there is more than one co-author then a single co-author can sign on behalf of all. The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work, he or she wrote the text.

Name: *Professor Richard D. Jones* Signature: Date: *04 November 2018*

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from which the extract comes:
Section 6.2.3 is adapted from a published conference paper: Baseer, A., Weddell, S.J., and Jones, R.D. (2017), 'Prediction of microsleeps using pairwise joint entropy and mutual information between EEG channels', In Proceedings of the 39th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4495–4498.

Please detail the nature and extent (%) of contribution by the candidate:
The PhD candidate did the research and wrote all sections of the published manuscript. The co-authoring supervisory team suggested research directions and offered criticisms which the candidate worked into the manuscript.

Certification by co-authors

If there is more than one co-author then a single co-author can sign on behalf of all. The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work, he or she wrote the text.

Name: *Professor Richard D. Jones* Signature: Date: *04 November 2018*

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my supervisors, Prof. Richard D. Jones and Dr. Stephen J. Weddell, for their continuous support of my PhD. I am really impressed with mentoring of Richard, whose patience, motivation, enthusiasm, attention to details, and guidance helped me throughout research and writing of this thesis. I would also like to thank Steve for sharing his insightful comments, technical expertise, and encouragement.

I would like to acknowledge doctoral scholarship, received from the Sukkur IBA University, Pakistan. I also acknowledge travel grants received from the University of Canterbury and New Zealand Brain Research Institute to attend the IEEE Engineering in Medicine and Biology conferences in Jeju, South Korea, and Honolulu, USA, respectively.

My sincere thanks goes to my colleagues and friends at New Zealand Brain Research Institute for their support. In particular, I would like to thank Dr. Reza Shoorangiz for volunteering his IT expertise and sparing his time for technical discussions.

I am eternally indebted to my parents for supporting me spiritually during the writing of this thesis and throughout my life in general. I wish to thank by brothers and sisters for their support and encouragement.

Last not the least, I thank my wife and daughters for their patience specially during my PhD.

CONTENTS

PREFACE

This thesis is submitted for the degree of Doctor of Philosophy in Electrical and Computer Engineering at the University of Canterbury. The research for this thesis was completed between October 2015 and October 2018 while I was enrolled in the Department of Electrical and Computer Engineering at the University of Canterbury. The work was carried out as part of Christchurch Neurotechnology Research Programme at the New Zealand Brain Research Institute and was supervised by Professor Richard D. Jones and Dr. Stephen J. Weddell. The University of Canterbury and the New Zealand Brain Research Institute, each kindly granted a travel grant for conference attendance.

PUBLICATIONS

Papers

•**Buriro, A.B.,** Shoorangiz R., Weddell, S.J., Jones, R.D. (2018), 'Predicting microsleep states using inter-channel EEG relationships, In *IEEE Transaction on Neural Systems and Rehabilitation Engineering (TNSRE)*, in Press.

Conference papers

•**Buriro, A.B.,** Shoorangiz R., Weddell, S.J., Jones, R.D. (2018), 'Ensemble learning based on overlapping clusters of subjects to predict microsleep states from EEG', In *Proceedings of the 40th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3036–3039.

• **Baseer, A.,** Weddell, S.J., Jones, R.D. (2017), 'Prediction of microsleeps using pairwise joint entropy and mutual information between EEG channels', In *Proceedings of the 39th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4495–4498.

PRESENTATIONS

• October 2018, Department of Electrical and Computer Engineering, University of Canterbury, New Zealand. Oral Presentation.

• July 2018, IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, Hawaii, USA. Oral Presentation.
• July 2017, IEEE Engineering in Medicine and Biology Society (EMBC), Jeju Island, South Korea. Oral Presentation

ABBREVIATIONS

ADASYN	Adaptive synthetic sampling
ANN	Artificial neural network
aRT	Alert response time
ASR	Artefact subspace reconstruction
AUC_{PR}	Area under the curve of precision recall
AUC_{ROC}	Area under the curve of receiver operating characteristic
bagging	Bootstrap aggregating
BIC	Bayesian information criterion
BM	Definite behavioural microsleep
CAR	Common average reference
CTT	Compensatory tracking task
dDTF	Direct DTF
DFT	Discrete Fourier transform
DMV	Diversity-incorporated majority voting
DTF	Directed transfer function
EEG	Electroencephalogram
EOG	Electrooculogram
ERS	Event-related synchronization
ESN	Echo state neural network
FCBF	Fast-correlation-based filter
ffDTF	Full frequency DTF
FFT	Fast Fourier transform
FIR	Finite impulse response
Fm	F-measure
GM	Geometric mean
GPDC	Generalized PDC

ICA	Independent component analysis
iCOH	Imaginary part of the coherency
IDFT	Inverse discrete Fourier transform
IIR	Infinite impulse response
IP	Instantaneous phase
kNN	K-nearest neighbour
LDA	Linear discriminant analysis
LF	Left frontal
LO	Left occipital
LOSO-CV	Leave-one-subject-out cross-validation
LP	Left parietal
LSTM	Long short-term memory
LSVM	Linear support vector machine
LT	Left temporal
MIT-BIH	Massachusetts Institute of Technology-Beth Israel Hospital
MPC	Mean phase coherence
mRMR	Minimal-redundancy-maximal-relevance
MVAR	Multivariate autoregressive
nDTF	Normalized directed transfer function
OLVQ	Optimal learning vector quantization
PCA	Principal component analysis
pCOH	Partial coherence
PCs	Principal components
PDC	Partial directed coherence
pdf	Probability density function
PR	Precision-recall
Pr	Precision
PS	Phase synchronization
PSI	Phase synchronization index
PVT	Psychomotor vigilance task
QDA	Quadratic discriminant analysis
RBF	Radial basis function

REST	Reference electrode standardization technique
RF	Right frontal
RMV	Rank-based weighted majority voting
RNN	Recurrent neural network
RO	Right occipital
ROC	Receiver operating characteristic
RP	Right parietal
RT	Right temporal
SFS	Sequential forward selection
SMOTE	Synthetic minority oversampling technique
Sn	Sensitivity
SNR	Signal-to-noise ratio
SONFIN	Self-organizing neural fuzzy inference network
Sp	Specificity
SVM	Support vector machine
TAN	Tree augmented naïve Bayes
VBLR	Variational Bayesian logistic regressionl
WLMSF	Wavelet log mean squared features

Chapter 1

INTRODUCTION

1.1 OVERVIEW

Microsleeps are complete and unintentional, sleep-related losses of consciousness of up to 15 s. They are accompanied by behavioural signs of eye closure, droopy eyes, head nodding, and total loss of visuomotor responsiveness (Davidson et al. 2007, Jones et al. 2010, Peiris et al. 2006), which are quite distinctive from the more tonic states of drowsiness (tendency to fall asleep) and fatigue (disinclination to responsiveness) (Poudel et al. 2014, Vanlaar et al. 2008). Tonic activity refers to a cognitive state of a person on a longer time scale, ranging from several seconds to minutes (Huang et al. 2008).

Microsleeps, when a person momentarily falls asleep, are part of everyday life. They are generally harmless and many individuals are unable to recognize that they have microsleeps. Performing active and attention-demanding tasks, like driving, on other hand are everyday norms and generally done easily. Loss of sleep-related consciousness during an extended-attention monotonous job, however, can lead to an erroneous (impaired), delayed (increased reaction time), or a completely failed (absent) response, the consequences of which are often catastrophic.

Despite distinct behavioural signs, outcomes of drowsiness, fatigue, and microsleep, especially in transport sector, are usually the same (Vanlaar et al. 2008). Sleep restriction increases the propensity of falling asleep and probability of accidents and duration of microsleeps are highly correlated (Sirois et al. 2010). Studies, however, have shown that non-sleep-deprived and healthy people can also have frequent microsleeps (Peiris et al. 2006, Poudel et al. 2014). In addition, no correlation was found between the number of microsleeps in normally-rested and sleep-deprived people (Innes et al. 2013). Propensity to microsleep, interestingly, increases with decrease in complexity of the task (Buckley et al. 2016), which suggests that microsleeps are more likely to occur in routinely performed active tasks like driving, trucking, aviation, navigation, maritime, and process control (Borghini et al. 2014, Léger et al. 2014, Tefft 2012). These studies suggest that microsleeps can intermittently occur at any time without any warning and cannot be induced. Whereas, sleep-restriction (Peiris et al. 2006) and prolonged engagement on a cognitive task (Dimitrakopoulos et al. 2018, Pageaux et al. 2015) can respectively introduce drowsiness and fatigue. Despite drowsy and fatigued having different meanings, they have

similar symptoms and are used interchangeably in the literature (Correa et al. 2014, Lin et al. 2013, Liu et al. 2016, Vanlaar et al. 2008).

In short, both drowsiness and fatigue can increase the propensity of microsleeps and can be relieved by sleep and rest respectively (Borghini et al. 2014). In contrast, individuals having sufficient rest and sleep can have microsleeps.

1.2 MOTIVATION

As per a nationally conducted survey, 41% of United States drivers admitted to have fallen asleep while driving at least once in their lifetime, 11% in the past year, and 3.9% in the past month (Tefft 2010). Drowsy driving has been estimated the main cause of 13.1% non-fatal and 16.5% fatal road accidents on the processed data of more than 97000 individuals (Tefft 2012). A study investigating the main causes of accidents based on samples gathered from six European countries showed that 12% of accidents were due to a temporary personal factor of fatigue (Thomas et al. 2013). 58.6% of 750 Ontario drivers admitted that they occasionally driven while fatigued or drowsy. 14.5% of respondents admitted that they had fallen asleep behind the wheel and nearly 2% were involved in a fatigue or drowsy driving-related crash in the past year (Vanlaar et al. 2008).

In 2015, the World Health Organization estimated that road accidents cause more than 1.2 million deaths, world-wide, each year. Globally, road accidents are the ninth leading cause of death across all age groups. Furthermore, it estimated that road traffic accidents will become the seventh leading cause of death by 2030 (World Health Organization 2015). Drowsy driving claims thousand of lives each year. In the United States, the annual cost, including societal harm and hospitalization costs, of drowsy driving has been estimated at $109 billions (Higgins et al. 2017).

As mentioned earlier, in addition to sleep-deprived individuals, non-sleep-deprived are also vulnerable to having microsleeps (Peiris et al. 2006). This raises major safety concerns, especially for those in high-risk occupations that require extended unimpaired visuomotor performance such as driving, aviation, navigation, maritime, and process control. Imminent microsleeps predicted accurately therefore have the potential to save lives and prevent catastrophic accidents.

Detection of microsleeps has been the focus of several studies (Ayyagari et al. 2015, Davidson et al. 2007, Golz et al. 2007, 2016, Malla et al. 2010, Peiris et al. 2011). Similarly, prediction of microsleeps has been the subject of more recent studies (Shoorangiz 2018, Shoorangiz et al. 2016). These studies were a way forward in detection and prediction of microsleeps. However, the reported accuracies were too low to implement and use microsleep detection/prediction systems in real-life.

Most of the electroencephalogram (EEG)-based studies aimed at detecting fatigue, drowsiness, or microsleeps have done a lot from two aspects of the overall system: investigation of different feature reduction/selection and different classification algorithms (refer Figure 4.6). In

addition to generally used spectral features, various temporal and information theoretic features of the EEG have also widely been used. However, these features have generally been univariate, i.e., extracted from individual channels of EEG while ignoring the shared information. In contrast, EEG time series are generally multivariate (i.e., multichannel) and synchronously recorded. Inter-channel relationship of EEG may therefore have additional microsleep-related information, which can be useful in improving detection and prediction accuracies of microsleeps. Furthermore, information contained in the inter-channel relationships may be complementary to that contained in univariate features and their fusion at feature or classifier-level could improve detection/prediction accuracies of microsleeps.

1.3 OBJECTIVES

The overall objective of this thesis was to explore various bivariate and multivariate features of EEG in the detection/prediction of microsleeps. The EEG is considered to contain information related to sleep-related unresponsiveness – fatigue, drowsiness, and imminent microsleeps – and therefore can be used to predict and ultimately prevent microsleeps. More specifically, the objectives were as follows:

- Conduct a review of the literature and gain the current knowledge of EEG-based approaches used to detect and predict microsleeps and lapses.
- Understand the theory of different inter-channel relationships of EEG and their practical utilizations and gains.
- Implement various bivariate and multivariate inter-channel relationships to detect/predict microsleep states and onsets and to compare their classification accuracies with some of the univariate features.
- Investigate different techniques to fuse features to improve accuracies at detecting/predicting microsleeps.
- Understand the theory of ensemble classification/learning.
- Evaluate the performance of microsleep prediction using behavioural and EEG data of an earlier study.

1.4 THESIS ORGANIZATION

This thesis is organized into nine chapters and three appendices. The current chapter provides an overview of microsleeps, related problems, and the motivations behind their detection/prediction. Chapter 2 provides a literature review of EEG, microsleeps, and the approaches used to detect/predict microsleeps. Chapter 3 provides a theoretical overview of EEG inter-channel relationships. Chapter 4 describes the data and terminologies, and illustrates an overview of

microsleep detection/prediction system. Chapter 5 presents aims and hypotheses for conducting this research. Chapter 6 describes the methods used, which comprise four parts. The first part describes the overall microsleep detection/prediction system including feature extraction, feature selection, and classification of microsleeps. The second part describes data fusion at feature-level and decision-level. The third part describes ensemble classification/learning. The last part discusses performance evaluation. Chapters 7 presents results and can be divided into four parts. The first part presents multiple within-group and inter-group comparisons of individual features at predicting microsleep states, followed by performance comparison between feature-selection techniques, and single classifiers. Based on the best performing single feature type, feature-selection technique, and classifier, the second part presents detection/prediction performance of microsleep states and onsets. Third and the last part present detection and prediction performance metrics of microsleep state and onset using data fusion and ensemble classification techniques, respectively. Chapter 8 comments on microsleep state prediction performance achieved with 21 types of individual features and comparisons among different groups of features, between feature-selection techniques and classifiers. The chapter then comments on performance achieved with data fusion followed by ensemble classification techniques. Chapter 9 concludes and presents key findings of this research, a critique of the current study, and the future work.

The thesis contains three appendices. Appendix A presents performances of non-normalized and non-causal features with two feature-selection techniques and classifiers at predicting microsleep states and onsets. Appendix B presents prediction performances of feature-level and decision-level fusion techniques on microsleep states and onsets. Appendix C presents prediction performances of different ensemble techniques with joint entropy features.

Chapter 2

OVERVIEW OF EEG-BASED MICROSLEEP DETECTION AND PREDICTION

2.1 THE ELECTROENCEPHALOGRAM

The EEG is a measure of the electrical activity of the brain and recorded from different spatial locations on the scalp. EEG recordings from the scalp are noninvasive and involve multiple electrodes. To cover the whole head, different EEG montages (combinations of electrodes) of m electrodes are used. An EEG with a higher number of electrodes provides higher spatial resolution but at a cost of higher computation and subject comfort (Correa et al. 2014). To record brain activity, electrodes are placed on the scalp according to a specific spatial distribution and generally in accordance with commonly used and internationally recognized 10–20 (as shown in Figure 2.1) or 10–10 systems (Oostenveld and Praamstra 2001). Then either mono (referential)- or bipolar (differential) derivations are obtained. For the international 10-20 system, clinicians measure the distance from the nasion to the inion and the head circumference. Then they mark precisely based on 10%–20% intervals of those distances for electrode locations excluding ear electrodes. For reliable recordings, an electrode impedance is generally recommended to be $\leq$ 5 kΩ, with gel used to achieve high conductance (Duffy et al. 1989, Sanei and Chambers 2007, Tong and Thakor 2009).

In addition to high temporal and adequate spatial resolutions, EEG recordings are noninvasive, relatively inexpensive, and convenient in real-time (Kayser and Tenke 2015), and therefore have been widely used in diagnostics, clinical, and sleep-related research settings (Correa et al. 2014, Liu et al. 2016, Tong and Thakor 2009, Torres et al. 2014). In addition, EEG recordings provide a measures of changes in brain activity associated with various states of arousal and have been used in classification of cognitive states of vigilance, alertness, fatigue, drowsy , and automatic sleep-staging (Bajaj and Pachori 2013, Şen et al. 2014, Melia et al. 2015). However, due to a very small amplitude (order of tens of µV), EEG is easily contaminated with intrinsic (e.g., eye-moment and muscle activity) and extrinsic artefacts (e.g., electrical line noise at 50 Hz) (Chen et al. 2015).

EEG is commonly divided into five frequency bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz). These frequency bands show different

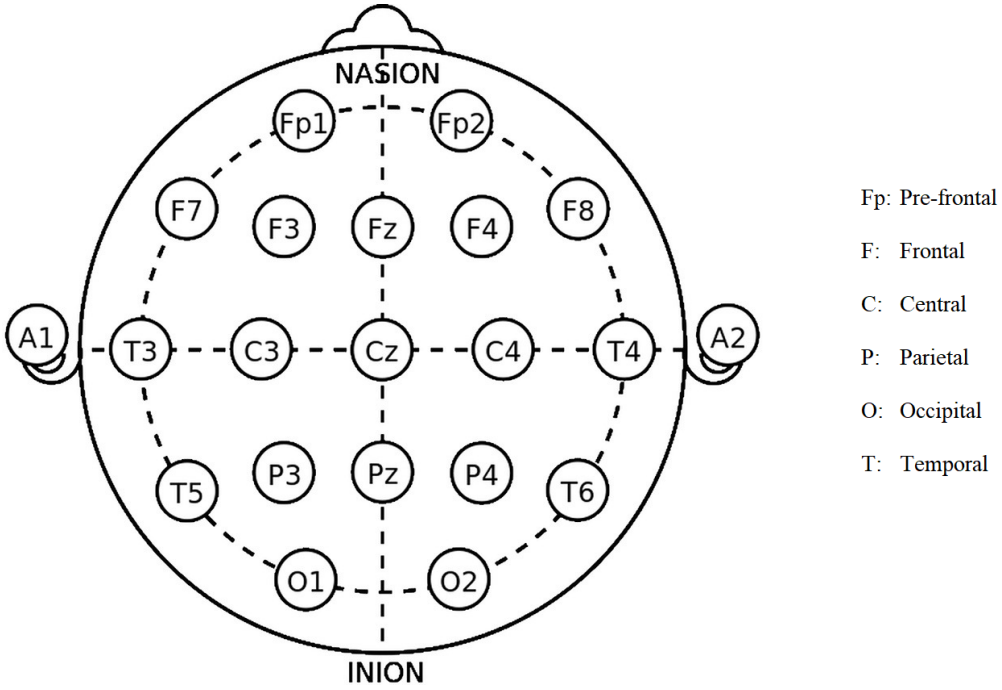

Figure 2.1 Spatial distribution of 10–20 international system of EEG electrode placement. Odd numbers refer to the left side of the scalp, even numbers to the right side of the scalp, and Z to mid-line placements. A signifies ear channels. Reprinted from (Rojas et al. 2018).

characteristics of cognition and mental state. Delta oscillations are primarily associated with deep sleep but can also be present in the waking state. Theta rhythms appear as consciousness slips towards drowsiness. Alpha oscillations are sinusoidal-shaped signals and commonly appear in the posterior scalp over the occipital region of the brain in awake individuals, and are attenuated by opening the eyes. The alpha band is the most prominent rhythm in the whole realm of brain activity. The beta band is associated with active thinking, alertness, and excitement. Gamma oscillations (sometimes called fast beta) though rarely occur but has been proved to be a good indication of event-related synchronization (ERS) (Burgess 2012, Sanei and Chambers 2007).

2.2 EEG-BASED CLASSIFICATION

2.2.1 Classification of sleep stages

Identification of sleep stages is useful in diagnosis of sleep-related disorders. Polysomnographic signals are generally used by experts to classify these stages. Such ratings are time consuming and often vary among the experts (Bajaj and Pachori 2013, Chapotot and Becq 2010, Memar and Faradji 2018).

To automatically classify sleep stages, Bajaj and Pachori (2013) used different features based on histograms of time-frequency images of the EEG time series. Using a multiclass least square support vector machine (SVM) classifier with Morlet kernel, they achieved a mean

10-fold cross-validation accuracy of 92.9% . Cross-validation is a statistical technique used to evaluate machine learning models on limited data samples.

Şen et al. (2014) compared various features, feature-selection algorithms, and classification algorithms at classifying sleep stages. Using features selected through a fast-correlation-based filter (FCBF), they achieved an accuracy of 97.3% with a random forest classifier. Various features in time, frequency, wavelet, and entropy domain were extracted from 30-s epochs of EEG, and SVM, artificial neural network (ANN), C4.5 decision tree, and random forest classifiers were investigated. Decision tree is a multistage classification approach, where each stage contains test condition(s) to separate data that have different characteristics. C4.5 is entropy-based and widely used algorithm to build a decision tree. Random forest is an ensemble of decision trees. For comprehensive coverage of different classification models, one can refer to Hastie et al. (2008).

Memar and Faradji (2018) extracted various features from EEG and achieved 5-fold cross-validation and leave-one-subject-out cross-validation (LOSO-CV) accuracies of 95.3% and 86.6% respectively. Minimal-redundancy-maximal-relevance (mRMR) and a random forest classifier were respectively used to select features and classify the sleep stages.

2.2.2 Detection of drowsiness and fatigue

Lin et al. (2012) proposed a generalized drowsiness detection system. Participants performed a driving simulation for one hour. They were asked to keep in the lane, while lane-departure events were randomly induced. EEG data of 1 s prior to the deviation onset was used as the baseline and the response time, i.e., time difference between response onset and deviation onset, as arousal state of the driver. After the response offset, the lane-departure events were repeated every 5–10 s. A short response time was considered to indicate alert and a long response time to be the drowsy state. EEG epochs of 1 s were decomposed by independent component analysis (ICA) and components thereof from the occipital region were Hamming windowed with 50% overlap and subsequently power spectral features, using fast Fourier transform (FFT), for theta and alpha bands were extracted. Their best subject-specific mean accuracy with a self-organizing neural fuzzy inference network (SONFIN), based on 10-fold cross-validation, was 97.2±1.6%, whereas best generalized detection accuracy based on LOSO-CV was 78.3±5.7%. Accuracy was the Pearson's correlation coefficient between the estimated and recorded response time.

Chen et al. (2015) wavelet-decomposed 8-s epochs of EEG and subsequently used the components thereof to extract different nonlinear features. Fusion of nonlinear features resulted in an accuracy of 95.6% in classifying alert and drowsy states, using EEG segments, defined by an expert. EEG data were recorded while participants were performing mental calculations in a cubicle. Independence of gold-standard from the behavioural cues and dependence on the EEG let the behavioural lapses go undetected and the detection is actually discrimination between the two groups of EEG epochs.

Liu et al. (2016) performed fatigue prediction during a 90-min driving simulation session using a similar paradigm to that of Lin et al. (2012). Unresponsiveness, as a consequence of falling asleep, of up to 2.5 s to randomly-induced lane deviations was considered as fatigue. 5-s EEG data prior to the deviation onset and during the response time were used to define the physiological and arousal states respectively. Power spectral features with a recurrent self-evolving fuzzy neural network resulted in an average LOSO-CV-based generalized accuracy of 90%.

Albalawi and Li (2018) developed a real-time drowsiness detection system using single-channel EEG. Relative powers (i.e., ratio of the power spectral density within the frequency band over the total power) of multiple 1-s epochs from 8 different frequency bands were nonlinearly mapped into 8 features. The nonlinear mapping was a sum of logical values obtained by comparing the relative powers against two thresholds, over 30 1-s epochs. These nonlinear features with a linear support vector machine (LSVM) resulted in a mean accuracy of 83.4% on EEG data of 16 subjects from the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) polysomnographic database. 30-s EEG epochs marked awake and stage-1 sleep (S1) were considered as alert and drowsy states respectively. The achieved accuracy is debatable as it is subject-specific and the number of cross-validations and amount of data used for training are not mentioned. System generalization is further limited due to inconsistent channel locations across the subjects. Earlier, Correa et al. (2014) using the same data, extracted different time, wavelet, and frequency-domain features from 5-s epochs and with an ANN achieved a specificity and sensitivity of 87.4% and 83.6% respectively. Specificity (true negative rate) refers to the ability of a classifier to correctly identify negative class instances, i.e., the ratio between correctly identified and actual negative class instances. Sensitivity (true positive rate also known as recall) refers to the ability of a classifier to correctly identify positive class instances, i.e., the ratio between correctly identified and actual positive class instances (refer Section 6.8).

2.3 LAPSES OF RESPONSIVENESS

Lapses of responsiveness ('lapses') are temporary episodes of complete failure to respond during active tasks. They differ in terms of their underlying cognitive mechanism and can appear as (1) failure to respond in time, i.e., prolonged reaction times, (2) failure to respond correctly, i.e., response error, and (3) complete phasic disruption of sensory-motor and cognitive performance (Buckley et al. 2016, Finkbeiner et al. 2015, Harrison and Horne 1996, Jones et al. 2010, Peiris et al. 2006).

Lapses can be categorized as microsleeps, attention lapses, and sleep events. A microsleep is an unintentional temporary ($\lesssim$ 15 s) loss of consciousness in which the person momentarily falls asleep. They are associated with lower arousal levels and can usually be identified by behavioural cues including slow eye-closure, loss of facial-tone, and head nodding (Davidson et al. 2007, Jones et al. 2010, Peiris et al. 2006, Poudel 2010).

An attention lapse is a brief disruption of goal-directed behaviour, without loss of consciousness, that generally results in slowed or absent response. Hence, it may be possible to continue on secondary tasks subconsciously, such as walking and driving, which don't require full-time conscious awareness/attention. Attention lapses can be identified by termination of performance on tasks (e.g., flat-spots during tracking) or substantially delayed responses on tasks (e.g., psychomotor vigilance task (PVT)).

Unresponsiveness of > 15 s is categorized as sleep. Sleep is physiologically distinct from microsleeps due to its duration (notwithstanding the arbitrariness of the 15-s demarcation) and the recovery of responsiveness after a brief period in microsleeps.

Microsleeps are negatively related to the complexity of a task, whereas attention lapses are positively related (Buckley et al. 2016).

2.4 DETECTION AND PREDICTION OF MICROSLEEPS/LAPSES

Golz et al. (2007) developed a microsleep detection system using 7 EEG, 2 electrooculogram (EOG), and 3 eye-tracking signals per eye (pupil size, x and y gaze coordinates). Participants were asked to drive in driving simulation. A monotonous task was intentionally selected to increase the likelihood of microsleeps. Microsleeps were identified from driving performance and behavioural clues of prolonged eye-closure and nodding-off. Power spectral features were calculated with a windowed periodogram of EEG segments and averaged over power in the standard EEG frequency bands. Power spectral and nonlinear features of delay vector variances were extracted from a 3-s window. The classification accuracy of power spectral features from EEG with optimal learning vector quantization (OLVQ) was similar to that of the fusion of both feature sets. Fusion of feature sets from all signals using an radial basis function (RBF) SVM classifier resulted in the best accuracy of 91% in classifying microsleep and non-microsleep (alert) events. Accuracy was defined as the total number of correct classifications relative to the total number of instances. Similarly, using RBF SVM classifier with power spectral features, they off-line predicted immediately occurring microsleep events with an accuracy of 88% (Golz et al. 2016). These promising but erroneous accuracies were achieved by balancing the test data and performing cross-validations on concatenated data from all the subjects. In doing so, independence of the test and training data, and hence generalization accuracy of the system, were lost. Their reported performances therefore do not reflect the true performances. What they called the detection/prediction was in fact the discrimination between manually selected epochs of alert and microsleep.

Krajewski et al. (2008) developed a subject-specific microsleep detection system using speech processing. Participants were asked to perform simulation driving, in a similar paradigm to Golz et al. (2007), and were instructed to engage in a verbal task similar to the navigation between a pilot and air-traffic controller. Microsleep events were detected using spectral features of the speech. They achieved an accuracy of 86.1%. Using the same speech data in a following study, the effect of various feature reduction methods was analyzed (Krajewski

et al. 2009). Neither supervised nor unsupervised feature selection techniques improved the performance. However, the practicability of speech-based microsleep detection in pilots and air-traffic controllers is debateable, and their approach is not suitable for applications in which subjects usually do not talk, such as drivers, and process controllers. Furthermore, microsleeps are less likely to occur during the conversation, instead attention lapses are more likely to occur. In fact, a silent period could be due to an episode of microsleep, which would be missed with this system.

Lin et al. (2013) used a sustained-attention driving task to detect 'behavioural lapses' and determine the effectiveness of providing feedback during these lapses. Participants were asked to maintain their lane position while events of lane perturbation were randomly induced. At the start of each experiment participants were fully alert and the first 5-min data were therefore considered as baseline to quantify the average alert response time (aRT). An event was called a lapse when the response time of subject was 3-times greater than the aRT. Time difference between deviation and the response onset was called the response time. Power spectral features were extracted from the ICA-decomposed EEG data. This study was further extended to an on-line lapse detection system with real-time feedback (Wang et al. 2014b). Although their results indicated that providing feedback can reduce reaction times, detection accuracies were not included. In addition to discrete task, their lapse identification depends on lane deviation and, consequently the lapse onset, for 8–10 s, remains unclear indicating limited practicability of the approach.

Compared to other ANNs, a long short-term memory (LSTM) recurrent neural network (RNN) with log-power spectral features resulted in a phi performance (defined in Equation (6.27)) of 0.38 (Davidson et al. 2007) at detecting lapses. Peiris et al. (2011) used different nonlinear, spectral power, and power spectral ratio features and achieved a slightly higher phi of 0.39 with a stacked generalization of 7 linear discriminant analysis (LDA) classifiers on pruned EEG data. Spectral power features, compared to nonlinear, normalized spectral power, and power spectral ratio features gave the highest performances. Ayyagari et al. (2015), using a stacking of 7 leaky echo state neural network (ESN) with power spectral features of the same EEG data achieved a reasonably higher phi of 0.51. However, using the same features and classifiers on unpruned EEG data, Ayyagari (2017) achieved a substantially lower phi of 0.44. The pruning involved exclusion of all 2-s epochs of EEG with $|z| > 3.0$, where z-score of each epoch was relative to mean and standard deviation of the first 2-min of baseline EEG (Peiris et al. 2011). In all of these studies, a 2-s window was used to extract features from EEG, principal component analysis (PCA) to reduce feature space, and LOSO-CV to evaluate the performance of their respective approaches. The resolution of detecting microsleep sates in these studies was 1.0 s. Microsleeps were identified using a 1-D continuous tracking task (Peiris et al. 2006). A continuous visuomotor task has a higher temporal resolution, which resulted in accurate identification of microsleeps.

Shoorangiz et al. (2016, 2017) predicted microsleep states 0.25 s ahead with a temporal resolution of 0.25 s. They used 5-s window to extract spectral features from the EEG. With a

single LDA classifier, a generalized mean phi of 0.33 was achieved with mutual information based greedy-forward-feature-selection algorithm (Shoorangiz et al. 2016), and 0.34 with Bayesian multi-subject factor analysis for feature reduction (Shoorangiz et al. 2017). They used both physiological and behavioural data of the previous study (Peiris et al. 2006) and rigorously revised the identification of microsleep (gold standard), described in Section 4.3 and shown in Figure 4.3.

Results of the aforementioned studies are a way forward in EEG-based, real-time, and continuous detection/prediction of microsleep states. However, the performances are still too low for real-life applications.

Chapter 3

EEG INTER-CHANNEL RELATIONSHIPS

3.1 INTRODUCTION

The human brain is functionally segregated at local levels while specialized functional brain areas/units are globally integrated and communicate within the network to perform cognitive or perceptual tasks (Bastos and Schoffelen 2016, Friston 2011). Such communication can be explored by using the inter-channel relationships of the EEG due to its multivariate property and simultaneously recorded signals. These inter-channel relationships are commonly known as functional and effective connectivity (Friston 2011) between EEG electrodes or brain anatomical regions and may or may not be normalized (Friston et al. 2014). Both connectivities can be model-based and model-free (data-driven) (Bastos and Schoffelen 2016). Functional connectivity is an observable phenomenon and is quantifiable in terms of synchrony/statistical dependencies between the brain units, and can be measured in time, frequency, or information theoretic domains. In contrast, effective connectivity corresponds to causal (directed) influence of one brain unit on another (as shown in Figure 3.1), is generally model-based, multivariate (Friston 2011, Wang et al. 2014a), and in frequency domain. Causally-related units are always correlated but the converse is not the true. Functional connectivity analyses make no or minimal assumptions about the underlying mechanisms, while effective connectivity analyses aim to find the simplest possible circuit diagram explaining observed responses (Friston et al. 2013).

For many experimental data (e.g., EEG), the underlying dynamics are generally unknown (Kreuz et al. 2007). In addition, different measures of connectivity are based on different underlaying mathematical assumptions and can therefore give different results (Bastos and Schoffelen 2016, Wang et al. 2014a). Choice for a particular inter-channel relationship to use at detecting/predicting microsleeps from EEG, eventually, becomes difficult to decide in advance. This thesis therefore focuses on extraction of various connectivity features and their performances at detecting/predicting microsleeps.

Throughout this thesis, EEG time series are of equal length (window size) while computing inter-channel relationships. Despite difference in terminology, both time series and signal represent information and, in the context of EEG, for the sake of clarity may occasionally be used interchangeably.

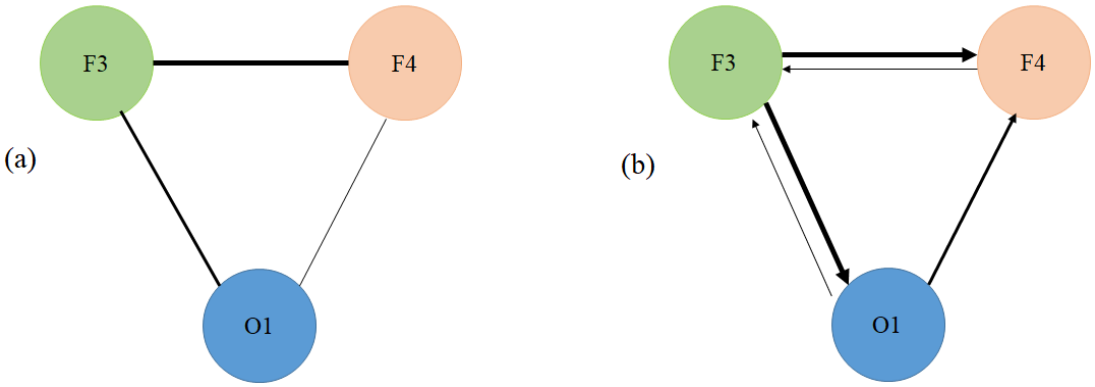

Figure 3.1 Illustration of connectivity among three channels of EEG. (a) Functional connectivity (e.g., coherence) whose the strength is represented by width of a line. (b) Effective connectivity (e.g., PDC). Arrows represent flow of information (source to sink) whose the width represents connectivity strength.

3.2 MODEL-FREE INTER-CHANNEL RELATIONSHIPS

Model-free inter-channel relationships are bivariate, symmetric, and extracted directly from each pair of the segmented EEG time series.

3.2.1 Covariance and Correlation

Covariance C describes the linear association (interdependence) between two EEG time series X and Y whereas correlation (Pearson's correlation coefficient) r describes the degree of such association between two EEG time series X and Y with N samples (window length). Thus are defined as

$$C_{XY} = \frac{1}{N-1} \sum_{i=1}^{N} (X(i) - \mu_X)(Y(i) - \mu_Y), \tag{3.1}$$

$$r_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y}, \tag{3.2}$$

where μ_X and μ_Y are the means and σ_X and σ_Y are the standard deviations of time series X and Y, respectively.

It should be noted that for two zero mean time series, covariance is equivalent to the cross-covariance function at zero lag. Similarly, for two zero mean and unit variance time series, correlation is equivalent to mean of the cross-correlation function at zero lag.

3.2.2 Cross-Spectral Power and Coherence

Cross-spectral power is a measure of interdependence based on total shared energy between two time series at a particular frequency. Coherence is the normalized cross-spectral power

with respect to individual auto-spectra. Both cross-spectrum and coherence describe linear interdependence between two time series and are sensitive to change in power and phase relationships. These measures are very effective when coupling is limited to a particular frequency band (Quiroga et al. 2002). The cross-spectral power S and coherence Co between two EEG time series X and Y, in a particular frequency band, are defined as

$$S_{XY} = |\langle X(\omega)Y^*(\omega)\rangle|, \tag{3.3}$$

$$Co_{XY} = \frac{S_{XY}}{\sqrt{\langle |S_{XX}(\omega)|^2\rangle \langle |S_{YY}(\omega)|^2\rangle}}, \tag{3.4}$$

where $*$, $\langle . \rangle$, and ω show the complex conjugate, expectation value over a specific frequency band, and an individual frequency respectively. $S_{XX}(\omega) = X(\omega)X^*(\omega)$ and $S_{YY}(\omega) = Y(\omega)Y^*(\omega)$ are the auto-spectra of X and Y respectively and $S_{XY}(\omega)$ is the cross-spectra between them at frequency ω. It is to be noted that auto-spectra and cross-spectral power, and coherence are respective frequency-domain equivalent of auto-covariance, cross-covariance, and cross-correlation functions. Frequency-domain linear interdependences can therefore indirectly be calculated by taking the Fourier transform of the corresponding time-domain interdependence followed by the expectation. Due to expectations over a particular frequency band, EEG time series need not to be band-pass filtered to calculate band-specific cross-spectral power and coherence between them.

3.2.3 Joint Entropy and Mutual Information

Unlike aforementioned linear measures of interdependencies, joint entropy and mutual information are also sensitive to and describe nonlinear interdependencies, which do not manifest themselves in the covariance. Joint entropy H and mutual information I between two EEG time series X and Y with N samples (window length) are defined as

$$H(X;Y) = -\sum_{i=1}^{N}\sum_{j=1}^{N} P_{XY}(i,j)\log P_{XY}(i,j), \tag{3.5}$$

$$I(X;Y) = \sum_{i=1}^{N}\sum_{j=1}^{N} P_{XY}(i,j)\log \frac{P_{XY}(i,j)}{P_X(i)P_Y(j)}, \tag{3.6}$$

where P_{XY} is the joint probability density function (pdf) between X and Y, P_X and P_Y are the marginal pdfs of X and Y, respectively. For two independent time series, their joint pdf becomes the product of their marginal pdfs.

It is to be noted that, due to logarithm, both joint entropy and mutual information, for perfectly interdependent time series, can approach infinity. Reliable estimates of pdfs require a large number of samples and measures of interdependence between two small length time series can therefore be inconsistent. For two time series being normal, the mutual information and Pearson's correlation coefficient are related as $I(X;Y) = -0.5\log(1 - r_{XY}^2)$. The units of both

joint entropy and mutual information are the same and these are defined corresponding to the base of logarithm, i.e., nats for ln, bits for $\log_2$, and Hartelys for $\log_{10}$.

3.2.4 Phase Synchronization Index

Phase synchronization (PS) quantifies rhythmic relationships between the two signals and ignores the effect of amplitude and therefore is less susceptible to artefacts and inter-session/inter-subject amplitude variability (Kabbara et al. 2016, Kong et al. 2017). It measures nonlinear interdependences in the frequency domain and is suitable for analyzing weak interactions between the two brain units that other measures hardly detect (Quiroga et al. 2002, Sun et al. 2012). The instantaneous phase (IP) of two time series involved in PS can be realized using an analytic time series Z, which for a univariate EEG time series X is defined as

$$Z = X + jH[X] = Ae^{j\theta}, \tag{3.7}$$

where A is the the instantaneous amplitude, $\theta = \tan^{-1}(\frac{H[X]}{X})$ is the instantaneous phase, and $H[X]$ is the Hilbert transform of X.

The analytic time series is calculated in three steps (Marple 1999, Sun et al. 2012) by: (1) taking the discrete Fourier transform (DFT) Y of the time series X, (2) doubling the positive frequency bins and zeroing the negative frequency bins, and (3) taking the inverse discrete Fourier transform (IDFT) Z of Y. This results in a phase shift (delay) of $\frac{\pi}{2}$ but does not alter the spectral distributions of the signal. The Hilbert transform can therefore be considered a unity-gain all-pass filter (Quyen et al. 2001).

The IP difference (relative phase) between the two EEG time series X and Y is defined as $\Delta\theta = \theta_X - \theta_Y$, which is then used to estimate phase synchronization index (PSI) either via entropy or mean phase coherence (MPC). For meaningful IP, the time series require filtering over a frequency band of interest . It is to be noted that right-hand-side of Equation (3.3) without $|.|$ can also be used to estimate $\Delta\theta$ and subsequently PSI.

3.3 MODEL-BASED INTER-CHANNEL RELATIONSHIPS

These inter-channel relationships are generally multivariate but may be symmetric (non-causal) or asymmetric (causal). A time series Y is said to be causal if its past and present terms (information) can be used to predict another time series X (caused).

Model-based (also called parametric) inter-channel relationships generally involve a multivariate autoregressive (MVAR) model of order p, described as

$$\mathbf{x}_t = \mathbf{v} + \sum_{k=1}^{p} \mathbf{A}_k \mathbf{x}_{t-k} + \mathbf{u}_t, \tag{3.8}$$

where $\mathbf{x}_t$ is the observed M-dimensional EEG data at time t, $\mathbf{v}$ is the $M \times 1$ vector of intercept terms, $\mathbf{A}_k$ is the $M \times M$ matrix of k^{th} autoregressive parameters (coefficients), and $\mathbf{u}_t$ is the innovation process, which is a multivariate uncorrelated white noise process with mean μ and nonsingular covariance matrix Σ. The elements $a_{i,j}(k)$ of $\mathbf{A}_k$ describe the contribution of channel j with lag k to channel i and are asymmetric (show direction of the connection), i.e., $a_{i,j} \neq a_{j,i}$. Modeling of time-delayed (i.e., $k > 0$) contributions represent causal relationships and MVAR is capable of describing such relationships.

Assuming a zero-mean process (i.e., $\mathbf{v} = 0$), Equation (3.8) can be written as

$$\mathbf{u}_t = \sum_{k=0}^{p} \hat{\mathbf{A}}_k \mathbf{x}_{t-k}, \tag{3.9}$$

where $\hat{\mathbf{A}}_k = -\mathbf{A}_k$ and $\hat{\mathbf{A}}_0 = -\mathbf{I}$. Performing Fourier transform on Equation (3.9) results in

$$\mathbf{U}(f) = \mathbf{A}(f)\mathbf{X}(f) \implies \mathbf{X}(f) = \mathbf{A}(f)^{-1}\mathbf{U}(f) = \mathbf{H}(f)\mathbf{U}(f). \tag{3.10}$$

Both $\mathbf{A}(f)$ and $\mathbf{H}(f)$ are asymmetric, show causal dependencies, and can, with Σ, be used to derive a battery of different inter-channel relationships. $\mathbf{H}(f)$ is generally called the transfer function of the system.

3.3.1 Multivariate Spectral Density and Coherence

The multivariate spectral density matrix is defined as

$$\mathbf{S}(f) = \mathbf{X}(f)\mathbf{X}(f)^* = \mathbf{H}(f)\mathbf{\Sigma}\mathbf{H}(f)^*, \tag{3.11}$$

where $S_{ii}(f)$ is the auto-spectra of channel i and $S_{ij}(f)$ is the cross-spectrum between channel i and j. The coherency $C_{ij}(f)$ is the normalized multivariate spectral density matrix with respect to auto-spectrum $S_{ii}(f)$, i.e.,

$$Co_{ij}(f) = \frac{S_{ij}(f)}{\sqrt{S_{ii}(f).S_{jj}(f)}}. \tag{3.12}$$

3.3.2 Imaginary part of Coherence

It is commonly acceptable to assume that EEG is quasi-static and volume conduction is instantaneous. Quasi-static refers to no (zero) time-lag between the scalp potential and underlying source activity whereas, volume conduction refers to a single brain activity observed in many channels. The imaginary part of the coherency (iCOH) depends on time-lag and is, therefore, insensitive to both volume conduction and non-interacting sources. Consequently, it accurately captures the true synchronization between two processes, time-lagged to each other

(Nolte et al. 2004). It is defined as

$$iCOH_{ij}(f) = im(C_{ij}(f)). \tag{3.13}$$

3.3.3 Partial Coherence

Partial coherence (pCOH) measures the intrinsic direct relationship (synchronization) between the channels and removes/discounts the effect of all other channels/common components (e.g., volume conduction) (Korzeniewska et al. 2003). Thus, pCOH can be regarded as the conditional coherence between i and j with respect to all other measured variables and is defined as

$$P_{ij}(f) = \frac{\hat{S}_{ij}(f)}{\sqrt{\hat{S}_{ii}(f).\hat{S}_{jj}(f)}}, \tag{3.14}$$

where $\hat{\mathbf{S}} = \mathbf{A}(f)\Sigma^{-1}\mathbf{A}(f) = \mathbf{S}(f)^{-1}$. pCOH is also defined in terms of minor $M_{ij}(f)$ produced by removing i-th row and j-th column from the spectral density matrix $\mathbf{S}$ as

$$P_{ij}(f) = \frac{M_{ij}(f)}{\sqrt{M_{ii}(f)M_{jj}(f)}}. \tag{3.15}$$

3.3.4 Partial Directed Coherence

Partial directed coherence (PDC) (Baccalá and Sameshima 2001) extends the concept for pCOH by measuring the directional influences and is defined as

$$\pi_{ij}(f) = \frac{A_{ij}(f)}{\sqrt{\sum_{i=1}^{M} |A_{ij}(f)|^2}} = \frac{A_{ij}(f)}{\sqrt{A_{:j}^H(f)A_{:j}(f)}}, \tag{3.16}$$

where superscript H indicates Hermitian operator (transposed complex conjugate) and subscript colon (:) indicates all row of matrix $\mathbf{A}$.

PDC is the direct and normalized outflow of information from channel j to i with respect to all the outflows from the source channel j and, subsequently, emphasizes the sinks. It holds the normalization properties, i.e., $0 \leq |\pi_{ij}(f)|^2 \leq 1$ and $\sum_{i=1}^{M} |\pi_{ij}(f)|^2 = 1, \forall\ 1 \leq j \leq M$.

3.3.5 Generalized Partial Directed Coherence

Generalized PDC (GPDC) is a modified PDC and is robust against finite time series samples (Baccalá et al. 2007). It is defined as

$$\pi_{ij}(f) = \frac{\frac{1}{\sigma_i}A_{ij}(f)}{\sqrt{\sum_{i=1}^{M} |\frac{1}{\sigma_i}A_{ij}(f)|^2}} = \frac{\frac{1}{\sigma_i}A_{ij}(f)}{\sqrt{\frac{1}{\sigma_i^2}A_{:j}^H(f)A_{:j}(f)}}, \tag{3.17}$$

where σ_i^2 refers to the variances of the innovations processes.

3.3.6 Normalized Directed Transfer Function

Normalized directed transfer function (nDTF) (Kamiński and Blinowska 1991) has the form

$$\gamma_{ij}(f) = \frac{|H_{ij}(f)|}{\sqrt{\sum_{k=1}^{M} |H_{ij}^2(f)|}} = \frac{|H_{ij}(f)|}{\sqrt{H_{i:}^H(f) H_{i:}(f)}}, \tag{3.18}$$

where $\mathbf{H}(f)$ is the directed transfer function (DTF) and subscript colon (:) indicates all columns of matrix $\mathbf{H}$.

nDTF, contrary to PDC, indiscriminately encompasses both direct and cascade (indirect) inflow from channel j to i and is normalized with respect to all inflows to sink channel i. It emphasizes the sources and has been used in localization of epileptic/seizure source (Kamiński and Blinowska 1991).

3.3.7 Full Frequency Directed Transfer Function

Full frequency DTF (ffDTF) is another way of normalizing DTF, i.e., with respect to the whole frequency band (Korzeniewska et al. 2003). It is given by

$$\eta_{ij}(f) = \frac{|H_{ij}(f)|}{\sqrt{\sum_f \sum_{k=1}^{M} |H_{ij}^2(f)|}} = \frac{|H_{ij}(f)|}{\sqrt{\sum_f H_{i:}^H(f) H_{i:}(f)}}. \tag{3.19}$$

The summation over the whole frequency band guarantees that the denominator remains insensitive to the frequency and, consequently, its spectrum depends only on the outflow from the channel j. In general, ffDTF shows similar peaks for frequencies represented in coherence and therefore measures direct causal connections (Korzeniewska et al. 2003).

3.3.8 Direct Directed Transfer Function

Direct DTF (dDTF) combines the direct information from pCOH with information on direction of influence from ffDTF. It is an extension of DTF that discriminates between direct and indirect causal relationships (Korzeniewska et al. 2003) and is given by

$$\delta_{ij}(f) = \eta_{ij}(f) P_{ij}(f). \tag{3.20}$$

3.4 CLASSIFICATION BASED ON EEG INTER-CHANNEL RELATIONSHIPS

Lee and Hsieh (2014) compared three model-free and symmetric inter-channel relationships namely correlation, coherence, and PSI in classifying different emotional states. Using quadratic discriminant analysis (QDA) classifier, they achieved the best accuracy of 82% with PSI.

Melia et al. (2015) using mutual information between O1 and its delayed version with an LDA classifier achieved a sensitivity of 65% and specificity of 75% in differentiating alert and sleepy groups. Similar values were also achieved with mutual information between O2 and its delayed version. Mutual information between O1–O2, however, without reducing the sensitivity, resulted in the highest specificity of 95%.

Wang et al. (2016) used the outflow by aggregating the PDC-based information flow. The outflow information was regarded as the input vector of RBF-SVM to discriminate ictal and interictal periods of the EEG. They achieved mean sensitivity of 91.44%, specificity of 99.34%, and precision of 67.88% on 5-fold cross-validation. The suitability of 2-s epoch, which resulted in the highest performances, however is debatable, as number of data points required to fit an MVAR model are insufficient. In addition, the model order used is unclear.

Dimitrakopoulos et al. (2018) extracted GPDC features from 5-min epochs of 62-channel EEG and with RBF-SVM achieved mean LOSO-CV accuracies of 0.92 and 0.97, respectively, on the classification of mental fatigue and alert states on a 1-h simulated driving task and on a PVT. The first and last 5 min of the respective task were considered as state of alert and fatigue respectively. Both proportion of data used in feature selection and model order are unclear.

These accuracies are impressive but are limited to class-balanced data, with much of the majority class data not processed and/or subject-specific. Binary class-balanced data refer to the similar representation (prior probabilities) of both classes with respect to each other. Majority class is an over-represented (higher prior probability) class relative to another class, called minority class. The ratio between minority class to majority class is referred to as imbalance ratio (refer Table 4.1). Subject-specific data refer to the data from a subject, one portion of which is used to train a classification model and other to test it.

3.5 SUMMARY

This chapter provided an introduction to EEG inter-channel relationships, which in neuroscience are known as functional and effective connectivity. A brief description of different model-free and model-based inter-channel relationships was presented. Finally, some literature related to EEG inter-channel relationship-based classification was reviewed.

Chapter 4

DATA

4.1 DATA

The original data were from an earlier study (Peiris et al. 2006) that comprised behavioural and EEG recordings from 15 non-sleep-deprived healthy male subjects, aged 18–36 (mean 26.5) years, and with combined vision of both eyes of 6/9 or better. Gender and age restrictions were to minimize their potential influences and, consequently, variations in the data. All subjects slept normally and the average sleep in the night prior to the test was 7.8 ± 1.2 h. Data in the current study were from a subset of 8 subjects who had at least one definite microsleep over the two sessions.

Participants continuously tracked a 1-D pseudo-random target, using a steering wheel of 395 mm in diameter, as accurately as possible for two 1-h sessions, one week apart. The pseudo-random target was generated by summation of 21 sinusoids with random phases but evenly spaced frequencies at 0.00781 Hz intervals. The pseudo-random target (bandwidth 0.164 Hz, period 128 s) scrolled down the screen at a rate of 21.8 mm/s and its 8-s preview was provided to the participants. During the task, EEG at 256 Hz, facial video at 25 fps, and tracking error at 64 Hz were recorded. Sixteen EEG electrodes, namely Fp1, Fp2, F3, F4, F7, F8, C3, C4, O1, O2, P3, P4, T3, T4, T5, and T6 were placed per the international 10–20 system (refer Figure 2.1). Horizontal and vertical EOG signals were also recorded to facilitate removal of eye-artefacts. The reference electrode was linked ears and ground electrode was on the forehead.

4.2 EEG PREPROCESSING

The EEG signals collected from the scalp of each individual were re-referenced to a common average reference (CAR) to improve the signal-to-noise ratio (SNR) (McFarland et al. 1997). CAR has proven to be an optimal high-pass spatial filter in which the average activity of the entire electrode montage is subtracted from individual electrodes (McFarland et al. 1997, Yu et al. 2014). The reference electrode standardization technique (REST) has been reported to be the best EEG reference in terms of recordings and connectivity patterns, but requires prior and accurate knowledge of the head model (Chella et al. 2016). CAR, in contrast, does not require

head model information, is insensitive to noise and number of EEG channels. Compared to the central (e.g., Cz) and other cephalic references (e.g., forehead reference), CAR resulted in the best performances (i.e., lowest errors between standard EEG potentials and transformed ones). Compared to REST, it has similar connectivity patterns and, in the case of high instrument/sensor noise, has similar error (Chella et al. 2016, Hu et al. 2018). On two datasets of P300 speller, compared to 11 different reference techniques including cephalic, bipolar, and surface Laplacian, CAR resulted in the best classification accuracies (Alhaddad 2012).

EEG signals were then band-pass filtered from 0.5 to 45 Hz using a zero-phase finite impulse response (FIR) filter. As mentioned in Chapter 3, this thesis was to investigate the performances of different inter-channel features at predicting microsleeps. A zero-phase filter preserves the shape (i.e., does not distort the phase) of the time series and therefore was used for unbiased performance comparisons between different feature sets at predicting microsleeps. It is to be noted that zero-phase filter being non-causal can't be used for real-time applications but linear-phase or minimum-phase filter being causal can be used. A filter is said to be causal if its output depends on past or present inputs. Minimum-phase refers to a causal and stable filter.

Artefacts were removed using artefact subspace reconstruction (ASR) (Mullen et al. 2015) followed by a canonical correlation analysis blind source separation (Clercq et al. 2006). ASR requires clean data to be used as calibration/base data to remove artefacts from the rest of the data. The clean data were selected based on z-score ≤ 5 of EEG. The noisy segments of the data were decomposed into principal components (PCs) which were then projected into the calibration data's space by using its covariance matrix. PCs which represented high-amplitude artefacts were removed based on a threshold derived from the calibration data. The remaining PCs were then back-projected into EEG channel space as shown in Figure 4.1. Considering non-stationarity associated with EEG, the EEG was segmented into 2-min epochs with 50% overlap and ASR was applied to each epoch independently. Calibration data of each epoch were found and used to clean the same epoch. The epochs were then concatenated to achieve a cleaned set of original EEG data. The overlapping parts of consecutive epochs were averaged to avoid discontinuity (Shoorangiz et al. 2016).

It should be noted that 2-s epochs containing artefacts were excluded/rejected from the original preprocessed EEG (Peiris et al. 2006, 2011). Whereas the EEG, used in this thesis, was preprocessed, without rejecting any epoch, to remove artefacts (Shoorangiz 2018).

The preprocessed EEG was decomposed into 5 sub-bands (i.e., delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13-30 Hz), and gamma (30–45 Hz)), then decimated to 128 Hz to reduce the processing time. EEG, however, was not decomposed into sub-bands for spectral features.

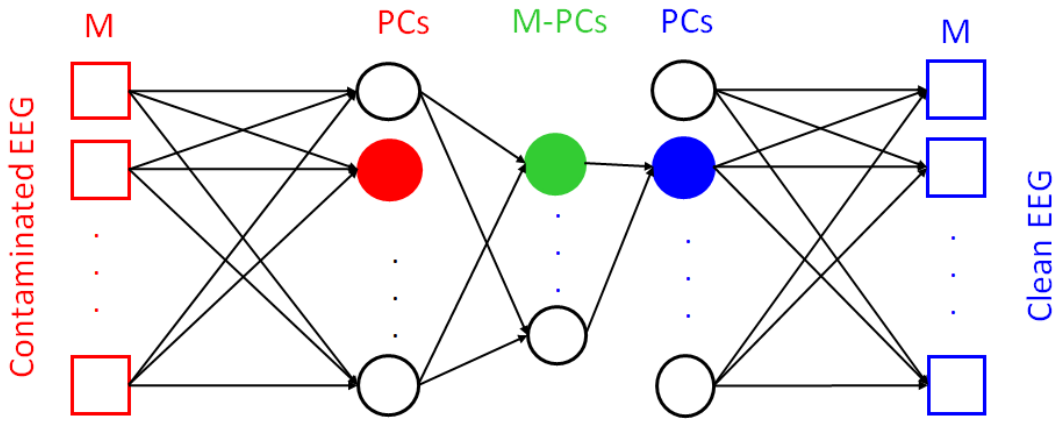

Figure 4.1 Illustration of ASR method. Red circle indicates artefact components, wich are discarded. Blue circle (imputation of each of the removed component) is a linear combination of activity of the remaining nonartefact components. M-PCs indicate calibration data space. M is the number of EEG channels. Adapted from (Mullen et al. 2015).

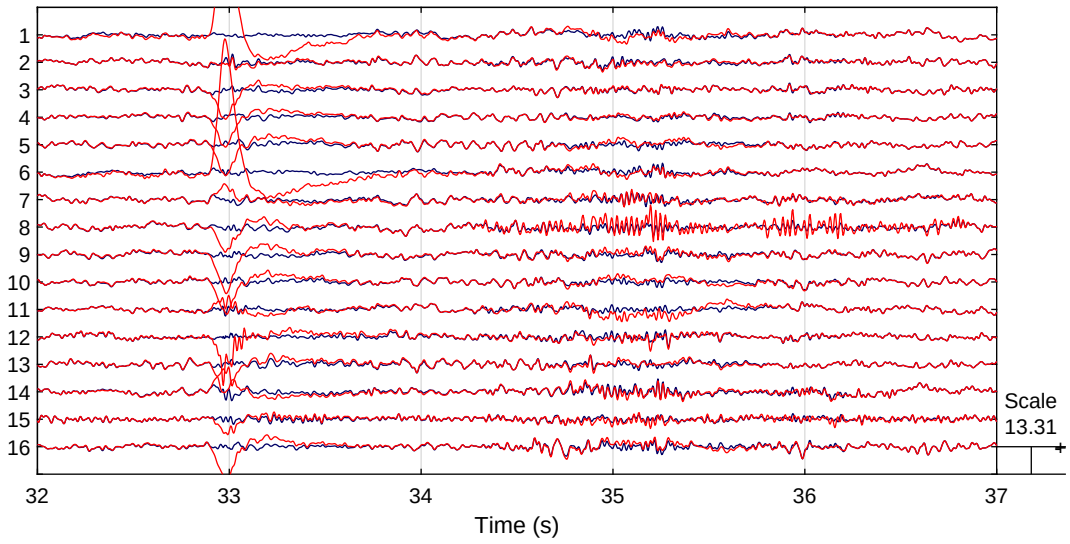

Figure 4.2 5 s of 16 channel EEG time series following ASR method, red trace (original data) is superimposed on blue trace (cleaned data). Scale (due to high amplitude artefacts) is set to mV.

4.3 GOLD STANDARD

The original gold standard was a combination of two independent measures, i.e., video ratings and flat-spots. The video recordings were conservatively rated by Peiris et al. (2006), without knowledge of the corresponding tracking performances. He rated the video on a 6-level scale using criteria like those of Wierwille and Ellsworth (1994). Levels 1–6 were marked alert, distracted, forced eye closure while alert, drowsy, deep drowsy, and sleep respectively. Flat-spots were found by a threshold-dependent automatic algorithm. Based on logical operators, video ratings and flat-spots were combined to generate two gold standards, i.e., a lapse index and a definite behavioural microsleep (BM). The lapse index was a logical *OR*, whereas BM was a

logical *AND* of the two independent measures. Therefore, a lapse was either a video-lapse and/or a tracking flat-spot and BM was the simultaneous occurrence of both flat-spot and video-lapse.

Shoorangiz (2018) pointed out shortcomings in the original gold standard and revised it. The tracking performances were regressively reanalyzed, which together with the original video ratings were used to revise the gold standard, comprising 3 classes – responsive, microsleep, and uncertain, as shown in Figure 4.3. The responsive label was defined as the coherent tracking performance irrespective of the video ratings. The microsleep label was the union of erroneous tracking and unresponsiveness along with video rating of deep drowsy or lapse. A mean absolute error of more than 30 mm on the screen that lasted for at least 1 s was marked as erroneous (incoherent) tracking. A drop in tracking velocity of more than 10% of the mean velocity of the corresponding target (1.2 mm/s) was marked as unresponsive. Epochs that did not fall contextually into either of these classes were labeled as Uncertain.

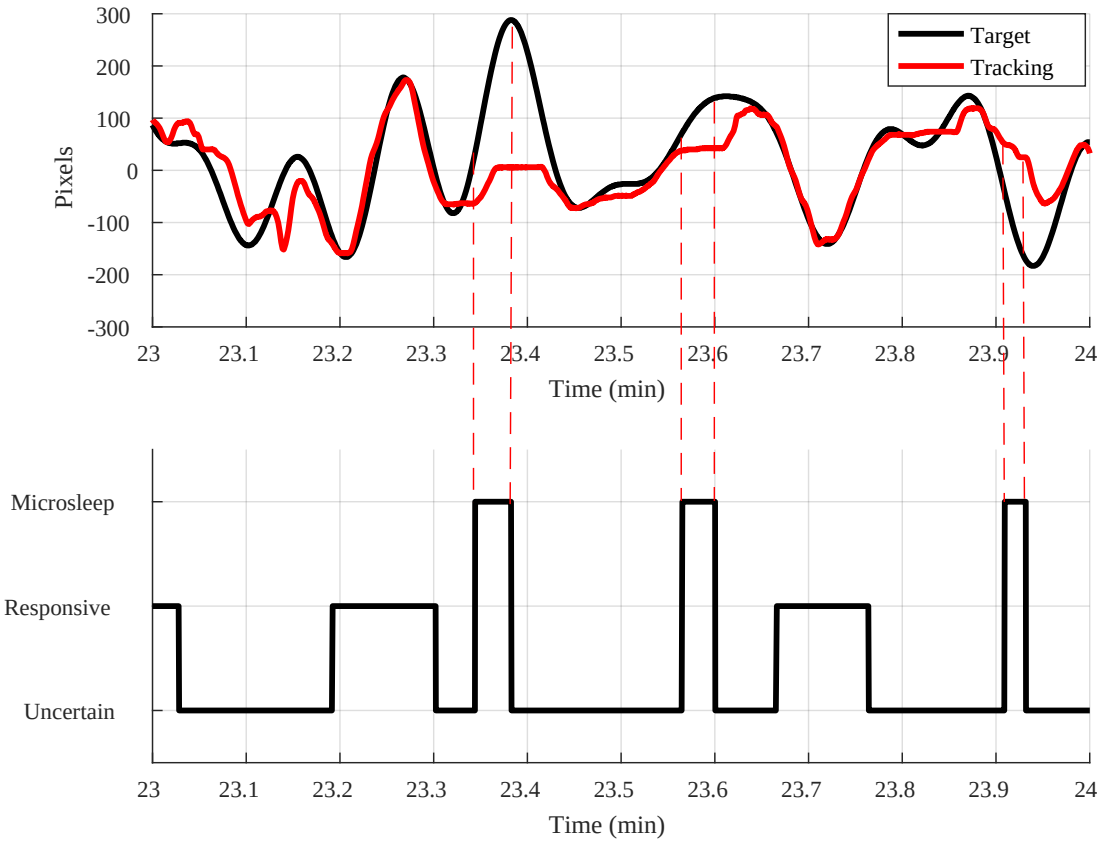

Figure 4.3 Illustration of tracking performance and corresponding continuous gold standard, where each label is an event. Microsleeps are instances of unresponsiveness as indicated by essentially flat tracking or incoherent tracking.

The gold standard is the key component of the overall microsleep prediction system and, as shown in Figure 4.6, was used in feature selection, classification, and performance evaluation steps.

4.4 TERMINOLOGY

Events, as shown in Figure 4.3, are continuous, whose the number and corresponding length vary across the sessions and subjects. Labels used in classification generally need to be discrete. In this thesis, two terminologies of state and onset were used. In earlier studies (Ayyagari 2017, Davidson et al. 2007, Peiris et al. 2006, 2011), states were events discretized/sampled at 1 Hz. However, due to uncertain labels, shown in Figure 4.3, it was difficult to identify all responsiveness and microsleep events. Like the recent work of Shoorangiz (2018), states in this thesis are referred to the gold standard sampled/discretized at 4 Hz, as shown in Figure 4.4, following which the uncertain labels were discarded. Therefore, the gold standard for states comprised all of the responsive and microsleep states.

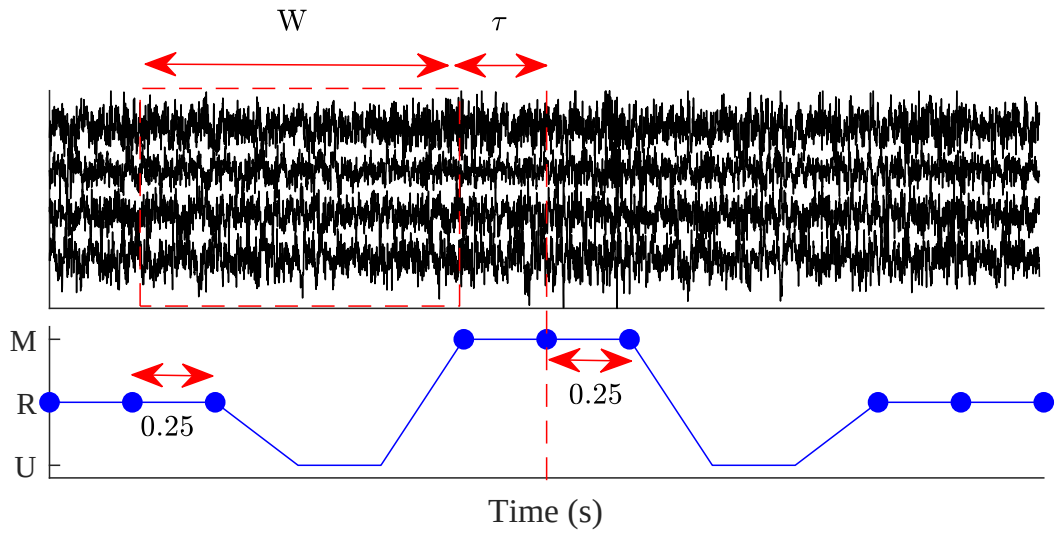

Figure 4.4 Illustration of microsleep state prediction. M, R, and U correspond to microsleep, responsive, and uncertain events of the gold standard. • indicates both responsive and microsleep states, and ↔ on the blue trace indicates the temporal resolution of states. Features extracted from the EEG of window length (W) s were used to predict the states at prediction time of (τ) s ahead of the gold standard.

Microsleep onset was referred to as the first state of a microsleep event, as shown in Figure 4.5. The gold standard for onset prediction unequivocal comprised all of the responsive states and the microsleep onsets.

Each subject had a different number and length of microsleep events, which results in quite different imbalance ratios. Subject-wise incidence and duration of microsleeps, together with imbalance ratios of states and onsets, are shown in Table 4.1.

The classification of both microsleep states and onsets from the corresponding EEG epoch (W) at $\tau = 0$ was referred to as detection, and for all $\tau > 0$ as prediction. Unless explicitly mentioned, in this thesis onward, microsleeps refer to both states and onsets. Similarly, the word 'prediction' is collectively used for both detection (prediction at $\tau = 0$) and prediction ($\tau > 0$). Classification was done with a temporal resolution of 0.25 s.

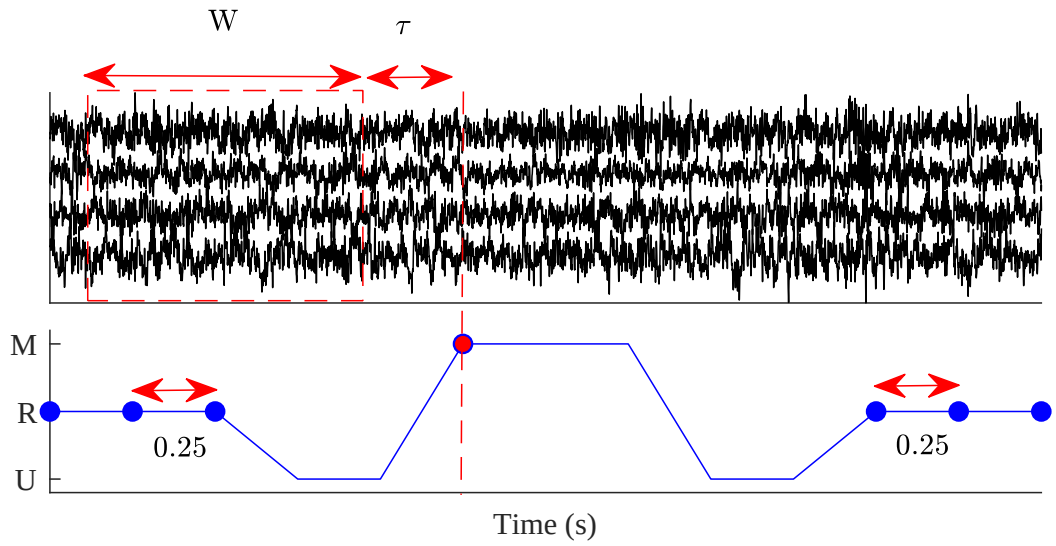

Figure 4.5 Illustration of microsleep onset prediction. M, R, and U correspond to microsleep, responsive, and uncertain events of the gold standard. • indicates responsive states, • indicates microsleep onset, and ↔ on the blue trace indicates the temporal resolution of responsive states. Features extracted from the EEG of window length (W) s were used to predict responsive states and microsleep onsets at prediction time of (τ) s ahead of the gold standard.

4.5 OVERVIEW OF MICROSLEEP DETECTION/PREDICTION SYSTEM

An overview of the overall microsleep prediction system is shown in Figure 4.6. The first stage is to collect EEG from the scalp of an individual, which, as mentioned in Section 4.2, is then preprocessed to remove various artefacts. Features (attributes), like those discussed in Section 6.2, of the preprocessed EEG are then extracted. A feature reduction or selection method is generally used to reduce the dimensionality of features. The selected (relevant) features together with gold standard (labels) from the training data are used to train a classifier/model. The selected features of test data are then fed to a trained classifier to classify (identify) the labels, e.g., imminent microsleeps. Various performance measures are finally used to evaluate a particular classifier by comparing its detected/predicted labels with the true labels as mentioned in Section 6.8.

Beside single feature set and classifier, data fusion involving multiple feature sets (known as feature-level fusion) and multiple classifiers (known as decision-level fusion), discussed in Section 6.6, are also widely used to cope with wide variability and scalability in the data and ultimately to improve overall system reliability.

Similarly, large and high-dimensional data are partitioned into multiple non-overlapping or overlapping groups (known as clusters) and the data of each cluster is used to train different classifiers of same (homogeneous) or different (heterogeneous) types. This approach is know as ensemble classification, discussed in Section 6.7 and, like data fusion, is aimed at handling data variability and improving overall system accuracy.

Table 4.1 Total microsleeps (both sessions) for each subject and corresponding imbalance ratio, i.e., microsleep:responsive. States correspond to a discrete gold standard, where the number of states = 4*Duration of events (temporal resolution of 0.25 s corresponds to a prediction frequency of 4 Hz). Onsets correspond to a gold standard, where the numbers of microsleep onsets and events are equal.

Subject	Microsleeps		Responsive	Microsleep:Responsive	
	Events	Durations (s)	durations (min)	States	Onsets
1	60	1011.3	56.6	4045:13570	60:13570
2	29	121.6	74.2	486:17796	29:17796
3	5	6.3	84.7	25:20335	5:20335
4	18	57.3	95.5	229:22905	18:22905
5	36	106.6	91.7	426:22004	36:22004
6	25	103.6	68.8	414:16510	25:16510
7	41	92.3	79.3	369:19043	41:19043
8	33	573.3	21.7	2293:5184	33:5184

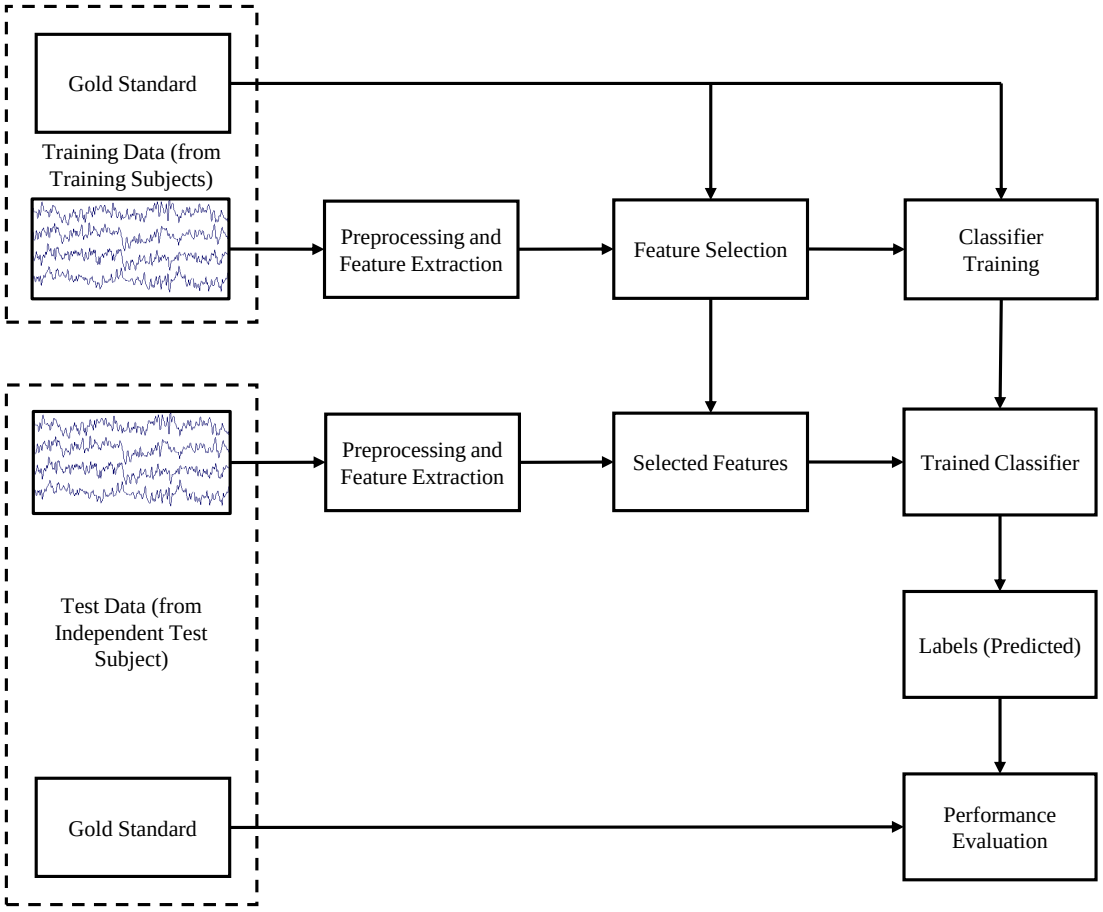

Figure 4.6 Block diagram of overall microsleep detection/prediction system.

4.6 SUMMARY

This chapter discussed the data used in this thesis. Different techniques used in EEG preprocessing and refinements of gold standard were subsequently discussed. Two terminologies of state and onset were illustrated and their corresponding imbalance ratios were provided. Finally, the overall microsleep prediction system was discussed.

Chapter 5

AIMS AND HYPOTHESES

5.1 AIMS

The aim of this research was to detect/predict behavioural microsleep states and onsets from inter-channel relationships of EEG time series, acquired from the scalp.

5.2 HYPOTHESES

This research posed two questions and hypotheses.

5.2.1 Inter-channel Relationships

•**Question:** Spectral features extracted from mono (referential) or bipolar EEG configuration have commonly been used in EEG-based analysis and classification of microsleeps. Neural communication, however, involves different brain units/regions, which raises the question of whether using inter-channel relationships features can lead to improved detection/prediction accuracies of microsleeps?

•**Hypothesis:** Inter-channel relationships can improve the detection/prediction accuracy of microsleep states and onsets over spectral features extracted from independent channels.

• **Rationale:** EEG signals are generally multivariate (i.e., multichannel) and recorded synchronously. The brain is a large-scale complex network, locally segregated and globally integrated, often referred to as the "connectome". Such local and global networks are based on assemblies/connections of neurons. Cognitive functions and information processing are mainly based on the interactions between specialized distant brain units and at multiple time scales (Bastos and Schoffelen 2016, Fell and Nikolai 2011, Friston 2011, Kabbara et al. 2016). Such communication/information exchange can better represent the brain state and be picked up by inter-channel relationships of the EEG than commonly used spectral features extracted independently from individual channels. Inter-channel features being higher in number than

individual channel features, from a machine learning perspective, may be advantageous in reducing the bias and, subsequently, improve the detection/prediction accuracies of microsleeps.

Different inter-channel relationships have been used in characterization of drowsiness (Awais et al. 2017, Dissanayaka et al. 2015), fatigue (Dimitrakopoulos et al. 2018, Kong et al. 2017), and microsleeps (Toppi et al. 2016), where they have been reported to be significantly different from alert. Information theoretic inter-channel relationships have also been used in characterization of daytime sleepiness (Melia et al. 2015). Although these studies are limited to characterization or classification of discrete events, which may not be used in continuous prediction of microsleeps, their results favour inter-channel features over conventional single channel spectral features. Characterization refers to the statistical comparison between two discrete events, e.g., Awais et al. (2017) performed t-test between coherences computed from 5-min EEG data during drowsy and consecutive alert events. Similarly, Toppi et al. (2016) performed t-test between connectivity patterns (i.e., PDC) obtained during microsleep events and baseline. Dimitrakopoulos et al. (2018) classified maximum alterness and maximum fatigue using connectivity features (i.e., GPDC) calculated from first and last 5-min window of the experiment.

5.2.2 Feature Fusion

• **Question:** Feature fusion, irrespective of level, has generally been shown to improve accuracy of EEG-based classification. However, Peiris et al. (2011) reported that fusion of spectral features with nonlinear features did not improve the detection accuracy of lapses. Similarly, fusion of spectral features with delay variance vector has been reported to marginally improve accuracy of classifying microsleep and alert events (Golz et al. 2007). Fusion of four wavelet-based nonlinear feature sets from the EEG resulted in 1.3% higher mean accuracy than the single best performing feature set (Chen et al. 2015) in classifying alert and drowsy states. These studies pose a question that instead of features extracted from individual channels, could fusion of connectivity patterns with either spectral power features or features extracted from individual channel improve microsleep detection/prediction accuracies.

• **Hypothesis:** Fusion of multiple feature sets can improve detection/prediction accuracy of microsleep states and onsets over single feature set.

• **Rationale:** Different feature extraction approaches have different mathematical formulations and therefore can have partially or fully orthogonal (complementary) information. In addition, EEG amplitudes fluctuate with respect to circadian rhythms (Rocca et al. 2014) and are directly correlated with the behavioural performances (Melia et al. 2015). Features extracted from individual channels in time or frequency domain may, therefore, vary within sessions and across subjects. In contrast, connectivity patterns are insensitive to such fluctuations; e.g., two signals may have different amplitudes and/or phases but high coherence is achieved when the

phase difference is relatively constant. It is, therefore likely that both features contain different aspects of the information. Fusion of such features, having orthogonal microsleep-related information, can improve the detection/prediction accuracies of microsleeps. Harvy et al. (2018) used 2880 connectivity (PDC) features and 55 power spectral features to discriminate driving fatigue from alert. Despite mean accuracy of connectivity features over 30 subjects being lower than power spectral features, their fusion at both feature-level and decision-level significantly improved the accuracy. Similarly, Rocca et al. (2014) used 56 power spectral features and 1540 connectivity (coherence) features in automatic EEG-based biometric identification. Power spectral resulted in higher accuracies than connectivity features. However, feature-level fusion resulted in perfect recognition rate (i.e., 100%) in both eye open and eye close conditions.

Chapter 6

METHODS

In addition to three univariate features, this chapter describes the methods/techniques used to extract various inter-channel relationships (refer Chapter 3) from EEG time series and to fuse them in Section 6.2 and Section 6.6, respectively. The chapter proposes three ensemble classification techniques in Section 6.7 and then describes metrics to evaluate the performance of a microsleep prediction system in Section 6.8.

6.1 SEGMENTATION

Window size (epoch length) is an important parameter in EEG inter-channel relationships (Wang et al. 2014a). It needs to be of adequate length to capture transient and tonic changes (Huang et al. 2008) to predict microsleeps, fit the model (Schlögl 2006) (if required), and computationally be handled easily (e.g., extraction of distance-based features). In contrast to earlier work (Ayyagari 2017, Davidson et al. 2007, Peiris et al. 2011), epochs of 5 s were used to ensure smooth and reliable connectivity estimates and to compare microsleep state and onset detection/prediction performances with spectral features (Shoorangiz 2018, Shoorangiz et al. 2016, 2017). Furthermore, the best classification accuracy in detecting ongoing microsleep events has been reported with an epoch of 5-s EEG (Golz et al. 2007). The preprocessed EEG was therefore segmented into 5-s epochs to extract channel-wise and inter-channel features. The step size was set to 0.25 s to match the temporal resolution of the gold standard which defines the lower limit of microsleep events (i.e., microsleeps $\geq$ 250 ms can be picked up) was used in this study. To identify microsleeps as early as possible, a high temporal resolution is required.

6.2 FEATURE EXTRACTION

Three univariate features (variance, spectral power, entropy), 7 bivariate features (covariance, cross-spectral power, joint entropy, correlation, coherence, mutual information, PSI), and 11 multivariate features (cross-spectral power, coherence, iCOH, pCOH, PSI, DTF, nDTF, ffDTF, dDTF, PDC, GPDC), shown in Figure 6.1, were extracted from individual epochs of EEG.

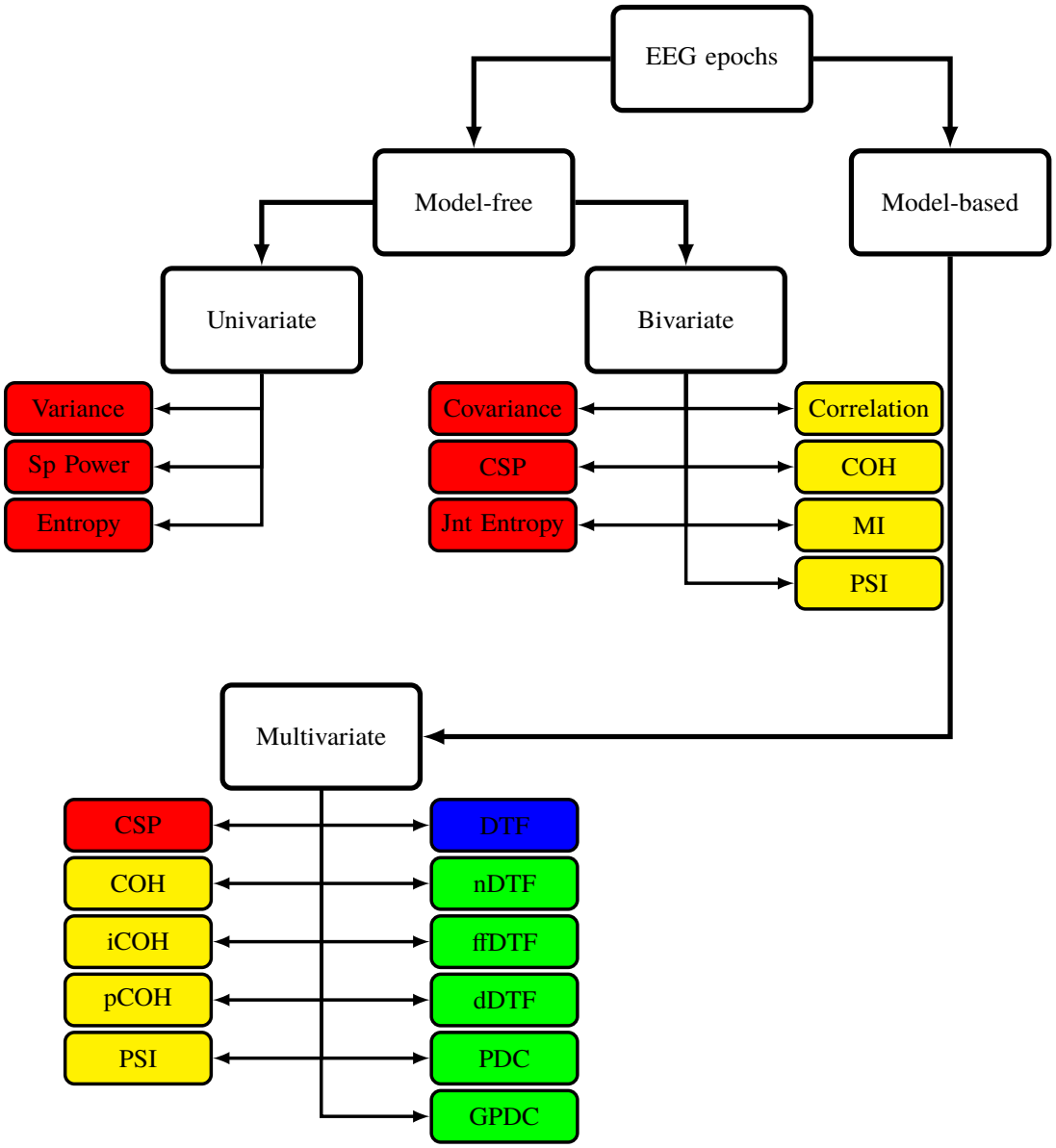

Figure 6.1 Taxonomy of feature extraction techniques used in this study. Non-normalized and non-causal, Normalized and non-causal, Non-normalized and causal, Normalized and causal.

All univariate, and bivariate features were model-free (or data-driven) and were directly extracted from the EEG epochs. However, coefficients of a fitted MVAR model to EEG epochs were used to extract multivariate (model-based) features. Model-free features can further be categorized according to the domain of computation as temporal, spectral, or information-theoretic. However, multivariate features are generally computed in the frequency domain.

6.2.1 Temporal Features

Channel-wise temporal features were the variances extracted from individual channels of EEG. The variance of a EEG time series X is defined as

$$V_X = \frac{1}{N-1}\sum_{i=1}^{N}(X(i)-\mu_X)^2, \tag{6.1}$$

where μ_X is the mean of X and N is the number of samples (window length). Non-normalized and normalized temporal features were extracted according to Equation (3.1) and Equation (3.2), respectively.

6.2.2 Spectral Features

In practice, the Fourier, Hilbert, and wavelet transforms are mathematically equivalent and yield similar results for neuronal signals (Bruns 2004, Quyen et al. 2001). However, the wavelet transform, due to its variable window size and better immunity to noise over the Fourier transform (Brittain et al. 2007, Klein et al. 2006), was used to extract channel-wise and inter-channel bivariate frequency-domain features. The wavelet spectral power feature S_X of an EEG time series X was defined as

$$S_X = \langle\langle|W^X(n,f)|^2\rangle_f\rangle_N, \tag{6.2}$$

where W^X is the wavelet transform of X, n is the number of time points, and f is the number of frequency points. $\langle.\rangle_f$ is the expectation, used as a smoothing function, over frequency (EEG sub-bands). $\langle.\rangle_N$ is expectation over time samples (window length) used to obtain spectral power feature for the corresponding window. Wavelet cross-spectral power S_{XY} and wavelet coherence Co_{XY} features between two EEG time series X and Y were defined as

$$S_{XY} = \langle|\langle W^X(n,f)W^{Y*}(n,f)\rangle_f|\rangle_N, \tag{6.3}$$

$$Co_{XY} = \left\langle \frac{|\langle W^X(n,f)W^{Y*}(n,f)\rangle_f|}{\sqrt{\langle|W^X(n,f)|^2\rangle_f\langle|W^Y(n,f)|^2\rangle_f}} \right\rangle_N. \tag{6.4}$$

The numerator and denominator of Equation (6.4) are required to be smoothed separately, otherwise the quantity will always be unity. Smoothing can be carried out over time, scale, or both and can be a simple averaging (Brittain et al. 2007, Grinsted et al. 2004, Plett 2007, Torrence and Webster 1999). Smoothing in scale/frequency direction has been empirically found to be more effective than smoothing in time direction (Maraun and Kurths 2004). For continuously long-time series (e.g., EEG), performing wavelet transform prior to the segmentation avoids edge effects and is suitable for on-line implementation. For consistency, both wavelet spectral and cross-spectral powers were smoothed over the frequency bands.

The Morlet wavelet (with $\omega_0 = 6$), due to its good frequency-time resolution, is a good

choice for feature extraction (Grinsted et al. 2004) and was therefore used as the mother wavelet

$$\psi_0(\eta) = \pi^{0.25} e^{-j\omega_0\eta} e^{-0.5\eta^2}, \tag{6.5}$$

where ω_0 and η are dimensionless frequency and dimensionless time respectively. Frequency (f)-to-scale (s) conversion was performed according to Torrence and Compo (1998) by

$$s = \frac{\sqrt{\omega_0^2 + 2}}{4\pi f}. \tag{6.6}$$

The Hilbert transform, however, was used to extract phase synchronization index features, as mentioned in Section 3.2.4. MPC is the most commonly used and suitable PSI for analyzing EEG signals at low sampling rates (Sun et al. 2012) and is defined as

$$\lambda = |\langle e^{j\Delta\theta} \rangle_N|, \tag{6.7}$$

where λ is the PSI feature, $\Delta\theta$ is the IP difference between two EEG time series X and Y, and $\langle . \rangle_N$ is the expectation over samples (window size).

6.2.3 Information-theoretic Features

Information-theoretic features generally involve the estimation of pdfs either by a non-parametric or parametric approach. The former is data-dependent and model-free, while the latter assumes the data distribution and is independent of data size (Walters-Williams and Li 2009). The k-nearest neighbour (kNN), which is a commonly used non-parametric estimator, was used to extract information-based channel-wise and inter-channel bivariate information-theoretic features. It has been reported to outperform for a small number of data points typically of the order of 100–1000 across noise levels (Khan et al. 2007). The entropy H feature of an EEG time series X via kNN is estimated as (Singh et al. 2003)

$$H(X) = \frac{1}{N} \sum_{i=1}^{N} \log(\varepsilon_i) - \psi(k) + \log(\nu) + \log(N), \tag{6.8}$$

where N is the number of sample points in X. $\varepsilon_i = ||x_i - kx_i||$ is the distance between a point and its k-th neighbourhood, and ψ is digamma function. The volume of the d-dimensional unit ball (ν) is defined as

$$\nu = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)}, \tag{6.9}$$

where Γ is the gamma function. Kraskov et al. (2004), using kNN, also ended up with a similar estimate of entropy, except $\log(N)$ was replaced by $\psi(N)$ and $\varepsilon_i = 2||x_i - kx_i||$ in Equation (6.8).

Similarly, the joint entropy feature between two univariate EEG time series X and Y is defined as

$$H(X,Y) = \frac{2}{N}\sum_{i=1}^{N}\log(\varepsilon_i) - \psi(k) + \log(\nu_X\nu_Y) + \log(N). \tag{6.10}$$

In practice, continuous data results in small values of distances (ε_i) which, due to the log factor in Equation (6.8), can result in highly fluctuating (variance) estimates of entropy. Such fluctuations can be avoided by using higher values of k but at the cost of increased bias. To account for bias-variance trade-off, a value of $k = 3$ from optimal values of 2–4 (Khan et al. 2007, Kraskov et al. 2004, Singh et al. 2003) was used. Mutual information between EEG time series X and Y was estimated in terms of their marginal and joint entropies as

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{6.11}$$

6.2.4 Multivariate Spectral Features

Model-based multivariate features are based on the estimation of MVAR coefficients, which in turn requires a number of delays (model order) to use. An MVAR of overly low order cannot capture the whole signal (under-fit) while an overly high order captures too much noise (over-fit) and results in inconsistent estimates. Fitting a p-order MVAR model to M-channel EEG requires an estimation of M^2p coefficients (free parameters), which in practice requires at least $10M^2p$ data points (Broersen 2009, Korzeniewska et al. 2008, Schlögl and Supp 2006). Better estimates of MVAR coefficients can therefore be obtained by using a longer time window but at a cost of low time resolution and stationarity. For reliable estimates and without increasing the window size, EEG signals were selected and averaged according to the brain regions, defined in Table 6.1. Neural activity in these brain regions is correlated with microsleeps (Poudel et al. 2010, 2018), central left (C3) and right (C4) channels were therefore not selected to reduce the number of MVAR coefficients.

Table 6.1 Brain regions and corresponding averaged EEG signals.

Region	Channels
Left frontal (LF)	Fp1, F3, F7
Right frontal (RF)	Fp2, F4, F8
Left temporal (LT)	T3, T5
Right temporal (RT)	T4, T6
Left occipital (LO)	O1
Right occipital (RO)	O2
Left parietal (LP)	P3
Right parietal (RP)	P4

The MVAR model order for each epoch (assuming local stationarity) was selected from 4 to12

using Bayesian information criterion (BIC). It is a consistent estimator, i.e., $\lim_{N\to\infty} Pr(P_{sel} = p_{true}) = 1$, and is defined as

$$BIC(p) = \log|\tilde{\Sigma}(p)| + \frac{\log(N)M^2p}{N}, \tag{6.12}$$

where $p, M, N, \tilde{\Sigma}$ are the model order, data dimension, window size, and a non-singular noise covariance matrix, respectively (Lütkepohl 2007). The model order p that minimizes the BIC (i.e., $\arg\min BIC(p)$) was selected.

Besides model order, coefficients of the MVAR model also depend on the estimator. A superior estimator provides a small prediction error and describes the data properties more accurately. All MVAR estimators are based on minimizing the residual errors, but for a small number of observations (data points), the Nuttall-Strand (multivariate Burg) (Strand 1977) method was found optimal because of stationarity of the estimated model and favourable numerical properties (Schlögl 2006). Prior to the frequency transformation, the stability and stationarity of the MVAR model were confirmed from its coefficients. Stability guarantees that the model will not diverge to infinity while stationarity confirms that its first and second moments are time-invariant. Stability implies stationarity, but the converse is not true. A model is stable if all eigenvalues λ of its coefficient matrix have a modulus less than one, alternatively $(\log|\lambda|) < 0$ (Lütkepohl 2007). The coefficients of MVAR, which were unstable for any epoch, were zeroed.

Multivariate connectivity measures, mentioned in Section 3.3, were averaged over the frequency bands.

6.3 FEATURE PREPROCESSING

Power-like features either in the frequency or time domain are generally skewed and the log transform is used to get their normal-like distributions. Therefore, variance, spectral power, and bivariate and multivariate cross-spectral power, and DTF features were log-transformed. Some of the covariance features were positively, while other were negatively, skewed. Therefore, to get normal-like distributions and maintaining/preserving the sign, covariance features were transformed as $sign(x).\log(|x|)$.

There can be a considerable difference in EEG time series and reaction times over different subjects and sessions. These result in varying and session and subject-specific distributions and subsequently, poor classification performances. A short segment of data from an individual subject is usually used to adapt a subject-specific classifier. Data collection and classifier retention before every session, however, become impractical. Subjects were generally alert and responsive at the start of each session. For the same task, homogeneity in EEG signals and reaction times can therefore be assumed. The first 2 min were thus arbitrarily treated as baselines and averages thereof were subtracted from the respective feature set and session. This was aimed at matching the data means without affecting their distributions. In contrast, the

standard normalization of features changes data distributions and can, therefore, remove some of the information. Thus, features were not divided by the standard deviation of the first 2 min.

6.4 FEATURE SELECTION

M channels of EEG decomposed into B (5) sub-bands gives $B \times M$ (i.e., 80) channel-wise, and $\frac{B \times M \times (M-1)}{2}$ (i.e., 600) bivariate inter-channel features per epoch. Similarly, EEG from R (8) brain regions decomposed into B sub-bands results in $B \times R^2$ (i.e., 320) multivariate features per epoch. A large number of features can be computationally expensive and introduce high variance in test results (over-fitting). Redundant and irrelevant features tend to substantially degrade classification performances. In contrast to dimensionality reduction techniques (e.g., PCA), feature selection techniques do not alter the original variable representation and maintain their semantics (Saeys et al. 2007). Two three-stage supervised classifier-dependent and classifier-independent approaches were therefore used to select the features from each training data set (e.g., concatenated data from 7 subjects for a single classifier approach). The use of classifier-independent feature selection was to avoid any classifier-related bias.

6.4.1 Classifier-dependent

This approach was a serial combination of a correlation coefficient, a Fisher score-based filter, and an LDA-based wrapper. The features were selected by using the sequential forward selection (SFS) (May et al. 2011) method. Initially, the training data were pruned by discarding the redundant (linearly-correlated) features ($|r| > 0.9$). A feature ranking (filtering) is a simple, computationally efficient, and scalable preprocessing step that shows the intrinsic relationships between individual features and the class (general goodness of features). A Fisher score (Duda et al. 2000) based filter was therefore used to rank the features per their individual discriminatory powers and the best feature was selected. A high Fisher score indicates a high discrimination (i.e., large difference of means relative to the sum of variances) between the classes. The Fisher score F of a feature set $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}] \in \mathbb{R}^{d \times n}$ was computed as

$$F(\mathbf{x}^j) = \frac{\sum_{k=1}^{c} n_k \times (\mu_k^j - \mu^j)^2}{v^j}, \tag{6.13}$$

where $\mathbf{x}^j$ denotes the jth row of $\mathbf{X}$. μ^j and v^j are the mean and variance of the feature j, μ_k^j is the mean of feature j corresponding to the class k and n_k is the number of sample points (observations) belonging to class k (Gu et al. 2011).

A filter, however, generally ignores feature dependencies and does not consider combined discriminatory power and, consequently, features selection by ranking can be suboptimal (Gu et al. 2011, Saeys et al. 2007). An LDA-based wrapper was therefore used to sequentially select the ranked features. The area under the curve of receiver operating characteristic (AUC_{ROC}) was used as the objective function (performance metric) due to its robustness under skewed-class

distribution (Fawcett 2006). The mean AUC_{ROC} of the 5-fold cross-validation with the top-ranked feature was calculated and saved. The successive feature was then combined with the top-ranked feature and selected if their combined mean AUC_{ROC} improved, otherwise it was discarded. The process was iterated until a stopping criterion was met. In this study, the stopping criterion was a logical *OR* of maximum number of features (70) and number of successive iterations (70) in which no performance improvement was observed. As at every iteration, a feature based on relative AUC_{ROC} was either selected or discarded. It was therefore possible for the feature selection process to complete before reaching the stopping criterion.

6.4.2 Classifier-independent

This approach was a serial combination of a relevance-based ranking (i.e., mutual information between a feature $\mathbf{x}$ and the class labels c) and relevance-based SFS method. The features were first ranked per their individual relevances. A high relevance indicates that a feature contains a high information about the class labels.

The relevance between a feature $\mathbf{x}$ and class labels was calculated as (Kwak and Choi 2002)

$$I(\mathbf{x}, c) = H(c) - H(c \mid \mathbf{x}) \tag{6.14}$$

and ranked. $H(c)$ is the entropy of class labels and is defined as

$$H(c) = -\sum p(c) \log p(c), \tag{6.15}$$

where $p(c)$ is the probability function of class labels. $H(c \mid \mathbf{x})$ is the conditional entropy and is defined as

$$H(c \mid \mathbf{x}) = -\sum p(c \mid \mathbf{x}) \log p(c \mid \mathbf{x}), \tag{6.16}$$

where the conditional probability $p(c \mid \mathbf{x})$, using the Bayesian rule, can be written as

$$p(c \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid c)p(c)}{p(\mathbf{x})}. \tag{6.17}$$

Relevant features, however, can either be similar and can degrade the classification performances or their combined relevance is poorer than their individual relevances. The relevance of the top-ranked feature was calculated and saved. The successive feature was then combined (union) with the top-ranked feature and selected if their combined relevance with the class labels improved , otherwise it was discarded. The process was iterated until a stopping criterion, i.e., a logical *OR* of the maximum number of features (70) and the number of successive iterations (70) in which no improvement in the relevance was observed, was met.

6.5 CLASSIFICATION

By incorporating a kernel, nonlinear classifiers generally map the data instance $\mathbf{x}$ to a higher dimensional vector $\Phi(\mathbf{x})$. In the worse case, the computational complexity of kernel classifiers is $O(N^3)$, where N is the number of training instances (Chapelle 2007, Lawrence and Schölkopf 2001). Data mapping to higher dimensions using basis transformation can lead to over-fitting (Hastie et al. 2008). The selection of optimal values for kernel and regularization parameters subsequently lead to a nested cross-validation which, for a highly demanding classifier with a large amount of data ($\sim$ 121,000) per training and feature set, becomes infeasible to use. Linear classifiers, in contrast, work directly in feature space and, consequently, are faster and often have similar performances to nonlinear classifiers for high dimensional and large data sets (Garrett et al. 2003, Yuan et al. 2012). Two linear classifiers, i.e., LDA and LSVM, were therefore used to explore and validate the efficacy of different channel-wise and inter-channel features of EEG in predicting microsleep states and to compare their prediction accuracies. For both classifiers, an algorithm-based approach referred to as cost-sensitive learning (Lin and Chen 2013) was employed by considering both classes as equally important (or equal misclassification cost) to account for class imbalance ratio in the training data sets. Equal prior probabilities (0.5) were therefore assigned to both classes in the decision.

6.5.1 Linear Discriminant Analysis

LDA has been widely used in brain-imaging to discriminate different brain states (Lemm et al. 2011, Lotte et al. 2007). LDA is based on the assumption that data of all the classes are normally distributed with a common covariance matrix Σ. However, due to stable estimates and simple boundaries, LDA generally works well even for data distributions deviated from normal. LDA discriminates two or more classes by a hyperplane (i.e., the decision boundary) given by $f(\mathbf{x}) = \mathbf{b} + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x} = 0$, which minimizes the inter-class variances and maximizes the distance between class means. For k classes with means μ_k and prior probabilities π_k, the LDA function is defined as (Hastie et al. 2008)

$$f_k(\mathbf{x}) = \mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\Sigma^{-1}\boldsymbol{\mu}_k + \log(\pi_k). \tag{6.18}$$

In practice, the parameters of Equation (6.18) are empirically estimated from the training data set. The test instance is classified according to $\hat{C} = \arg\max_k f_k(x)$. For unequal π_k, the error rate can be improved by moving the threshold towards minority class (Hastie et al. 2008).

6.5.2 Linear Support Vector Machine

For separable data, there can be multiple separating hyperplanes. The LSVM – a widely used classification algorithm – attempts to find an optimal hyperplane ($\mathbf{b} + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x} = 0$) that maximally separates the instances of two classes according to their distance (margin) as shown in Figure 6.2.

Maximization of margin on training data leads to better generalization capability, i.e., better classification performances on test data (Hastie et al. 2008, Kumar et al. 2014, Lotte et al. 2007, Quitadamo et al. 2017).

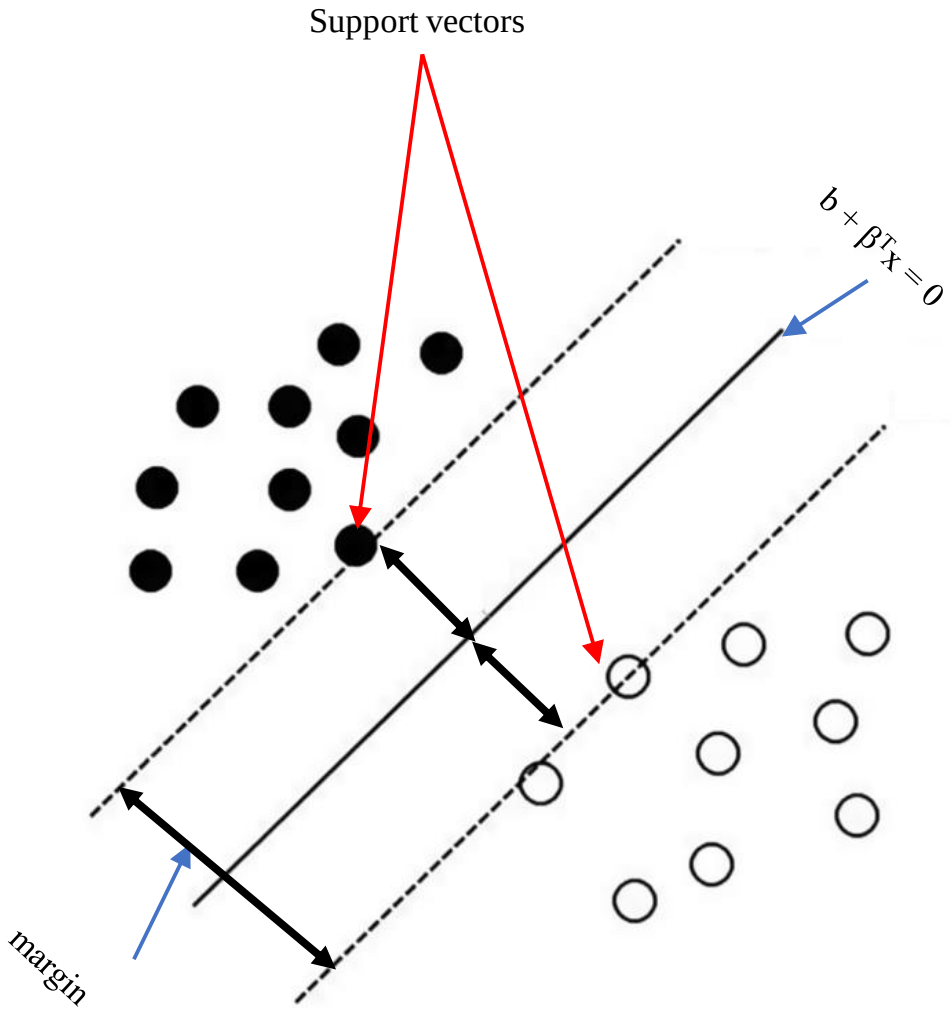

Figure 6.2 Illustration of margin and support vectors (i.e., instances on the boundary of slab). Filled and unfilled circles are the instances of two classes (labels).

6.6 DATA FUSION

Data fusion is a multi-level process that integrates data/information from several sources to achieve improved information (Castanedo 2013). Given a number of feature sets and/or classification models, data fusion in this research was aimed at achieving higher classification accuracy in predicting microsleeps. Data fusion can be performed at different levels/stages of the overall classification system. Three types of fusion strategies, namely, raw data (low-level), feature (intermediate-level), and decision (high-level) are common (Dasarathy 1997, Mangai et al. 2010). Low-level fusion involves multiple modalities/sensor sets with the goal to obtain a low detection error and high reliability and degree of freedom. The motivation behind

low-level fusion is that a single modality can hardly give a complete knowledge of a complex process/system. Its practicability, due to different incompatibilities among modalities, is limited (Lahat et al. 2015). Both feature-level (different feature sets) and classifier-level (multi-classifiers) fusion, as shown in Figure 6.3, involve a single modality/sensor set and are the widely used fusion strategies (Chen et al. 2018, Golz et al. 2007, Harvy et al. 2018, Planet and Iriondo 2012, Yang et al. 2003).

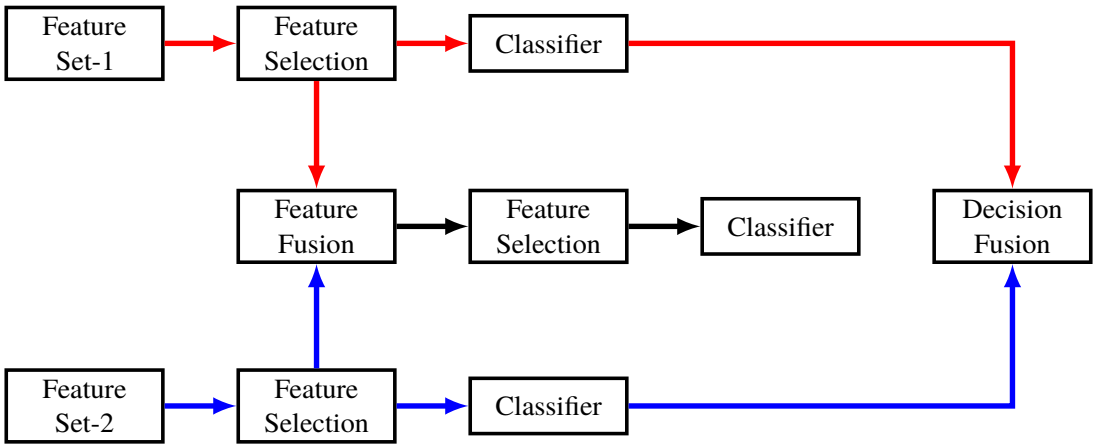

Figure 6.3 Schematic diagram of data fusion at feature-level and decision-level.

Feature-level fusion uses most discriminatory information from the original multiple feature sets while eliminating the redundant features (Mangai et al. 2010, Yang et al. 2003). This strategy, however, requires preprocessing (rescaling) of differently scaled-feature sets and generally a 2-stage feature selection for a very large and high dimensional data set. At the first stage, features are selected from individual sets followed by a next-stage feature selection to avoid any redundancy caused due to the fusion of independently selected features from different sets. A single classifier, used in feature-level fusion, may not handle a wide variability and scalability in different feature sets (Mangai et al. 2010). Whereas, decision-level fusion using a combination of multiple classifiers (of same or different types) is simple and stable and can, therefore, provide a better and unbiased performance. Each classifier of the same type, in this strategy, uses a different set of appropriate features, whereas the same feature sets can be used for different classifier types. This generally is to achieve diversity (uncorrelated errors or different decision boundaries) among the classifiers, which is the key issue in classifier combination. The final output (decision) is obtained by merging decisions of each classifier using a variety of methods, e.g., average, minimum, maximum, or median of the posterior probabilities or widely used majority voting of the output labels (Ali et al. 2015, Kuncheva and Whitaker 2003, Lahat et al. 2015, Rahman and Verma 2010).

6.6.1 Feature-level fusion

Fusion of various feature sets, explored in this thesis, will result in a very high dimensional (i.e., 80×3 univariate + 600×7 bivariate + 320×11 multivariate = 7,960) feature set, where

selection of the most discriminatory features could computationally be very expensive. Features were therefore selected in two stages. Features from various feature sets were separately selected and then fused, which could have resulted in redundant and non-complementary features and consequently may not have improve prediction accuracy. A second-stage feature selection was used to ensure that the fused features are orthogonal (complementary) to each other and to avoid redundancy. Features extracted using different techniques were numerically unbalanced. Min-max normalization, defined in Equation (6.19), was therefore used to address the numerical unbalances among the training features. The min-max values thereof were subsequently used to normalize the test features. Min-max normalization was preferred over the widely used feature standardization, i.e., $\frac{X-\mu}{\sigma}$, due to different feature sets exhibiting different distributions.

$$\hat{F} = \frac{F - F_{min}}{F_{max} - F_{min}}. \tag{6.19}$$

The final feature-selected features were used to train both the classifiers, which were then used to predict microsleeps from previously unseen test features.

6.6.2 Decision-level fusion

Feature types explored in this research, shown in Figure 6.1, were assumed to be uncorrelated, and subsequently the same type of (homogeneous) classifiers were used. Features selected from each of the 21 feature types were used to train the same number of classifiers. Decisions thereof at predicting microsleeps on unseen test features were obtained and subsequently fused as

$$P_{fuse} = \frac{1}{N} \sum_{i=1}^{N} P(C \mid X, F_i), \tag{6.20}$$

where N shows the number of classifiers assigned to feature types F and P shows the posterior probabilities of individual classifiers. C and X are respectively the test class and features of the respective set. The final decision on classifying a microsleep was made by comparing P_{fuse} against the threshold of 0.5.

6.7 ENSEMBLE CLASSIFICATION

An ensemble classification refers to a combination of decisions obtained from multiple classifiers that learn differently from the training dataset and have therefore distinct class boundaries. A combination of such multiple and accurately trained classifiers is generally believed to be more accurate than individual classifiers. Factually, each learning algorithm has some limitations and associated risks, where ensemble classification is aims to improve the accuracy and confidence, while minimizing the risks (Ali et al. 2015, Dietterich 2000, Valdovinos and Sanchez 2009, Zhou 2012). A classifier is said to be accurate if its error rate is better than random guessing on a new feature value (Moacir 2011, Zhou 2012). Distinct class boundaries are generally

obtained by either training heterogeneous classifiers on the same data or the homogeneous classifiers on different data partitions. The two important aspects in the construction of ensemble classification are the distinct decision boundaries of each participating classifier and either the combination or selection of their decisions. Ensemble classification can therefore be regarded as a decision-level fusion. Data partitions are generally achieved through clustering techniques and fusion though majority voting. Ensemble classification is categorized as supervised learning and in the literature has also been named as *multiple classifier system*, *mixture of experts*, and *committees of learners*. Bootstrap aggregating (bagging), boosting, and stacked generalization are widely used ensemble techniques in machine learning literature.

In some applications, like continuous classification of microsleeps or sleep stages, there are multiple training subjects and the data of each subject is large. In such applications, due to data variability, a single classifier may be suboptimal. Multiple classifiers are generally trained on different portions of the data and their decisions are fused to get improved accuracies. To get different subsets of data, clustering is performed on the concatenated data. In applications, where the data of each subject is large, both concatenation and clustering may become redundant and time-consuming processing steps. Furthermore, variability in the concatenated data generally exists. In this situation, the data of each subject or session intuitively assumed to be an individual cluster. This assumption often gives two benefits, i.e., inexpensive computation and diversity among classifiers.

In this thesis, beside the above mentioned non-overlapping clusters (i.e., data of each individual training subject), overlapping clusters (i.e., data of one or more subjects are present in multiple clusters) were proposed. Standard and two variants of majority (soft) voting were proposed to fuse the decisions of all classifiers. Both LDA and LSVM classifiers were used separately as weak/base classifiers. Features for each cluster were selected using both classifier-independent and classifier-dependent techniques.

6.7.1 Majority voting

Due to its simplicity and effectiveness, majority voting is the most commonly used method for combining decisions of individual classifiers. It can directly be performed on decisions (e.g., binary outputs) or on posterior probabilities of the classifiers, also known as *soft voting*. However, binary output-based majority voting requires the number of classifiers to be odd. In this thesis, soft voting, as it is generally used for homogeneous ensembles (Zhou 2012), was used and defined as

$$P_{maj} = \frac{1}{N} \sum_{i=1}^{N} P(C \mid X, k_i), \tag{6.21}$$

where N is the number of k clusters and P shows the posterior probabilities of individual base classifiers. C and X are respectively the test class and features. Each base classifier was trained on data of the corresponding cluster (training subject). To obtain the class labels, average

posterior probability P_{maj} was compared against the threshold of 0.5.

6.7.2 Diversity-incorporated majority voting

Unlike majority voting, base classifiers were trained on concatenated data of all training subjects. Features, however, were selected from data of each training cluster (subject). The same concatenated data were used to obtain the decisions of each trained classifier. Q statistic (Yule 1900) was used to measure the diversity between each pair of contributing classifiers D_i. Q statistic is symmetric and defined as

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}. \tag{6.22}$$

Statistically-independent classifiers result in Q of zero. A correct decision from both classifiers leads to a positive value of Q, while disagreement between them in committing errors leads to a negative value of Q. The value of Q varies between -1 and 1 (Kuncheva and Whitaker 2003, Zhou 2012), where fusion of classifiers showing higher absolute values of Q are unlikely to improve performances. All classifiers, however, were trained and tested on the same data and, consequently, their decisions were likely to be correlated (i.e., high positive Q). The mean Qs were obtained from matrix **Q** showing pairwise statistics among all of the base classifiers and the classifier with minimum mean Q showing disagreement on correct decisions was discarded. The remaining mean Qs were subsequently used as static weights W_i, assigned to the decisions of individual base classifiers as

$$P_{Dmaj} = \frac{1}{N}\sum_{i=1}^{N} P(C \mid X, k_i)W_i. \tag{6.23}$$

To obtain class labels, the final decisions i.e., P_{Dmaj}, were compared against the threshold of 0.5.

For a label data set $Z = [z_i, ..., z_N]$ and the output $y_i = [y_{1,i}, ..., y_{N,i}]$ of a classifier D_i, N^{ab} is the number of instances for which two classifiers make decisions, as shown in Table 6.2.

Table 6.2 A 2×2 table showing the relationship between a pair of classifiers.

	D_j correct (1)	D_j incorrect (0)
D_i correct (1)	N^{11}	N^{10}
D_i incorrect (0)	N^{01}	N^{00}
Total, $N = N^{11} + N^{00} + N^{01} + N^{10}$		

6.7.3 Rank-based weighted majority voting

Like Section 6.7.2, base classifiers were trained on concatenated data of all training subjects and the same classifiers were used to test and subsequently obtain the posterior probabilities.

These, together with the training class labels, were used to calculate the Fisher score (refer Equation (6.13)) between them. A high Fisher score shows the relevance of classifier decision with the class labels. Fisher scores were normalized with respect to their maximum value and used as static weights, assigned to the decisions of individual classifiers as

$$P_{Rmaj} = \frac{1}{N}\sum_{i=1}^{N} P(C \mid X, k_i)W_i, \tag{6.24}$$

where W_i are the weights. Class labels were obtained by comparing P_{Rmaj} against a threshold of 0.5.

6.7.4 Ensemble of overlapping clusters

The formation/creation of the previously mentioned ensembles were based on data of non-overlapping clusters. However, in general, there exist some degree of correlation between data of two non-overlapping clusters or an individual subject performing the same task. By exploiting such correlation, formation of overlapping clusters was proposed. In this approach, multiple classifiers trained on some part of similar data can classify a new instance with high accuracy. Similarly, diversity among the classifier was maintained by having data of different sizes together with separate features from each cluster.

Kullback–Leibler divergence is an asymmetric measure of dissimilarity, where 0 indicates complete similarity between the two pdfs. This was calculated between each pair of the training subjects (concatenated data from both sessions). Assuming multivariate normal distributions for responsive data of each subject, Kullback–Leibler divergence from probability distribution Q to P was defined as (Duchi 2007)

$$D_{KL}(P \mid\mid Q) = 0.5\left(tr({\Sigma_2}^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T {\Sigma_2}^{-1}(\mu_2 - \mu_1) + \log\left(\frac{\det\Sigma_2}{\det\Sigma_1}\right) - d\right), \tag{6.25}$$

where tr is the trace of covariance matrices Σ, μ is means, and d is dimension of the feature sets. Symmetric Kullback–Leibler divergence was defined as

$$D_{KL} = 0.5\Big(D_{KL}(P \mid\mid Q) + D_{KL}(Q \mid\mid P)\Big). \tag{6.26}$$

The symmetric divergence matrix was normalized with respect to the maximum value. Subjects of each row were clustered if their normalized divergences were below a threshold of 0.70. Divergences among the training subjects were estimated prior to feature selection. Therefore, due to redundancy among features, and consequently, small determinants of covariance matrices, values of divergences were likely to be high. Correspondingly, a high threshold was manually selected. Redundant clusters, i.e., clusters containing the same subjects, were discarded. Features were selected from data of each cluster, which then were used to train the respective base classifier. Majority voting was used to fuse the decisions of individual classifiers

and class labels were obtained by comparing the final decisions against a threshold of 0.5.

6.8 PERFORMANCE EVALUATION

Performance metrics were based on data from the 8 independent test subjects who had at least one microsleep in one of the two sessions. Performance evaluations, shown in Figure 4.6, were according to LOSO-CV as: (1) one subject as an independent test-subject was reserved, (2) the remaining 7 subjects were used to select features and to train a classifier/model, (3) data of the test-subject were fed to the classifier and performance metrics were obtained, (4) steps 1–3 were repeated until all subjects had been the test-subject, and (5) performance metrics from all 8 subjects were averaged to obtain the overall performance metrics. The use of LOSO-CV was to attain a generalized microsleep prediction system and, subsequently, an estimate of the true performance metrics for data from an unforeseen subject.

For a binary-class problem, various performance metrics can be formulated from elements of a confusion matrix, shown in Table 6.3. sensitivity (Sn) (recall) $\left(\frac{TP}{TP+FN}\right)$, specificity (Sp) $\left(\frac{TN}{TN+FP}\right)$, and Precision (Pr) $\left(\frac{TP}{TP+FP}\right)$ are three basic measure of performances.

Table 6.3 Confusion matrix for a binary-class problem

	Predicted class	
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

These metrics, however, by themselves are incomplete. Different combinations of these metrics are, in practice, used to summarize the entire performance information into a scalar metric. For imbalanced binary learning problems, F-measure (Fm) $\left(\frac{(1+\beta)^2 \times Sn \times Pr}{\beta^2 \times Sn + Pr}\right)$ and geometric mean (GM) $(\sqrt{Sn \times Sp})$ have been widely used (Lin and Chen 2013, López et al. 2013, Parker 2011, Powers 2011, Straube and Krell 2014). Fm shows the compromise between sensitivity and precision, while GM equally weighs the classification performances on both minority and majority classes. The relative importance between precision and sensitivity in Fm is indicated by β. GM, however, lacks consideration of errors (incorrect decisions) and in practice may be deceptively convenient to report high classification performances, specifically for highly imbalanced data. In contrast, Fm is biased towards the minority class and completely ignores the performance on majority class (Powers 2011).

Matthew's correlation coefficient (phi) simultaneously handles both classes (Straube and Krell 2014) by incorporating all elements of the confusion matrix and has been used for imbalanced biomedical data (Boughorbel et al. 2017, Parker 2011):

$$\text{phi} = \varphi = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FP) \times (TP + FN) \times (TP + FP) \times (TN + FN)}}. \tag{6.27}$$

Phi weighs equally the classification accuracies and errors on both classes, and consequently, can be relatively straightforward to interpret and is a preferable single measure of performance on imbalanced data. Shoorangiz (2018) reported F-measure with $\beta = 1$ to be highly correlated ($r = 0.98$) with phi.

In addition, two curve-based threshold-independent performance evaluation metrics of AUC_{ROC} and area under the curve of precision recall (AUC_{PR}) have also been used for imbalanced learning problems (Gong and Kim 2017, Saito and Rehmsmeier 2015). The former is insensitive to class imbalance (Fawcett 2006) while the later is sensitive to the class imbalance (Saito and Rehmsmeier 2015).

A receiver operating characteristic (ROC) is a two-dimensional graph between true postive rate (also called sensitivity) and false positive rate. An ROC graph shows relative trade-off between benefits (true positve rate) and cost (false positive rate). Point (0,0) in right panel of Figure 6.4 indicates no benefit-no cost, whereas, point (1,1) indicates all benefits at all the costs (all positive class). The point (1,0) shows perfect classification (i.e., sensitivity = 1, specificity = 1). The diagonal red line $y = x$ represent random guess. A classifier appearing in upper/lower triangle performs better/worse than random guessing.

To compare performances of multiple classifiers, ROC graphs need to be reduced to scalar values. A common method is to calculate AUC_{ROC} that is the portion of area of the unit square.

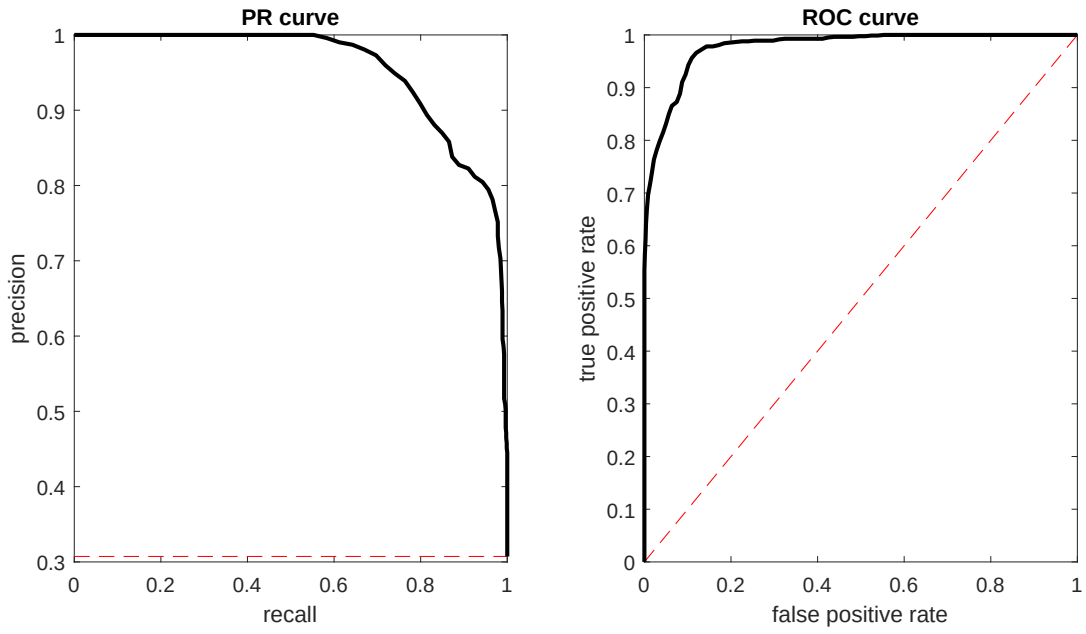

Figure 6.4 Precision-recall and receiver operating characteristic curves. False positive rate = 1- specificity.

A precision-recall (PR) graph is another way of visualizing and comparing performances of multiple classifiers on class-imbalanced data sets. A PR curve is a two-dimensional graph between precision and at the corresponding sensitivity (recall). In contrast to ROC graph, the baseline (i.e., red line in left panel of Figure 6.4) of PR graphs moves with the imbalance ratio

and the perfect classification is achieved at point (1,1).

To compare performances of multiple classifier, like ROC graphs, PR graphs also need to be reduced to scalar values called AUC_{PR}s, i.e., the portion of area of the unit square.

As shown in Table 4.1, the imbalance ratio between subjects is highly variable. Performance metrics like sensitivity, precision, F-measure, phi, and AUC_{PR}, however, are sensitive to the imbalance ratio of the data. Consequently, there might be performance measure inaccuracies when LOSO-CV is applied, and subject-independent performance metrics are computed. Therefore, in this project, the three metrics of phi, AUC_{ROC}, and AUC_{PR} were used in comparing performance among different feature sets.

The paired non-parametric Wilcoxon signed-rank test (Gibbons and Chakraborti 2003) was used to compare prediction performances between feature selection methods, of each feature type within the group (i.e., univariate, bivariate, and multivariate as shown in Figure 6.1. The comparisons were repeated for each feature selection technique and each classifier. For ease of comparison, bivariate group was divided into two sub-groups: normalized (correlation, coherence, mutual information, and PSI) and non-normalized (covariance, cross-spectral power, and joint entropy).

The revised gold standard was used in this study and, consequently, the results presented in this thesis (i.e., Chapter 7) are not directly comparable to earlier microsleep studies (Ayyagari 2017, Davidson et al. 2007, Peiris 2008). Results, however, are directly comparable to more recent work (Shoorangiz 2018, Shoorangiz et al. 2016, 2017).

6.9 SUMMARY

This chapter can be divided into four parts. The first part, provided an outline of the procedures used in microsleep prediction systems, including different feature extraction, and two feature-selection techniques and classification methods. Moreover, a brief description of different classifiers chosen for microsleep prediction was provided. Part two discussed about two fusion techniques. Ensemble classification techniques were discussed in part three. Finally, procedures to evaluate overall performance and the metrics were described.

Chapter 7

RESULTS

This chapter presents microsleep state prediction (τ = 0.25 s) performance of all 21 individual features (refer Figure 6.1) with both of the feature-selection techniques and both classifiers described in Section 6.4 and Section 6.5, respectively. This is followed by a comparison of performance metrics between feature-selection techniques and classifiers. Based on this, the best performing single feature set, feature-selection technique, and classifier are chosen and used to determine microsleep (state and onset) detection and prediction performance for a single feature set, multiple features (i.e., data fusion), and ensemble classifications, respectively.

All results, presented in this chapter, are the average values of 8 test subjects. In all tables and figures, Var, SP, Ent, Cov, CSP, JE, Cor, COH, and MI represent features extracted using variance, wavelet spectral power, entropy, covariance, wavelet cross-spectral power, joint entropy, correlation, wavelet coherence, and mutual information, respectively.

7.1 COMPARISONS

The mean performance metrics of all 21 types of features (refer Figure 6.1), with both feature-selection techniques and with single LDA and LSVM classifiers are shown in Table 7.1 and Table 7.2, respectively. The aim of this section is to compare within and inter-group (i.e., univariate, bivariate, multivariate as shown in Figure 6.1) performance, across both feature-selection techniques and classifiers. For ease of comparison, bivariate group has been divided into two subgroups: normalized (correlation, coherence, mutual information, PSI) and non-normalized (covariance, cross-spectral power, joint entropy). Similarly, multivariate group has been divided into non-causal (cross-spectral power, coherence, iCOH, pCOH, PSI) and causal (PDC, GPDC, DTF, nDTF, dDTF, ffDTF) subgroups. All results and the comparisons from Section 7.1.1 to Section 7.1.8 are based on prediction of microsleep states (refer Figure 4.4) at τ = 0.25 s.

All non-normalized and non-causal features, compared to baseline log-spectral power features with information-based greedy feature-selection method and an LDA classifier (Shoorangiz et al. 2016), gave higher mean prediction (τ = 0.25 s) performance on microsleep states. Similarly, mean prediction performance metrics achieved with non-normalized and non-causal

features and across both feature-selection techniques and classifiers, compared to baseline log-power spectral features with multi-subject factor analysis-based feature reduction method and an LDA classifier (Shoorangiz et al. 2017), were higher.

7.1.1 Univariate features

With the classifier-independent feature-selection, entropy features with an LDA, in terms of AUC_{ROC}, were marginally better than variance (i.e., 0.95 vs 0.92, $p = 0.063$) and spectral power (i.e., 0.95 vs 0.91, $p = 0.060$). Similarly, with an LSVM classifier, entropy features, in terms of AUC_{ROC}, were marginally better than variance (i.e., 0.95 vs 0.92, $p = 0.063$) and spectral power (i.e., 0.95 vs 0.92, $p = 0.063$). Apart from this, there was no within-group significant performance difference.

In the univariate group, the overall highest performance metrics (phi = 0.44, AUC_{PR} = 0.48, AUC_{ROC} = 0.95) were achieved with entropy features, classifier-independent feature-selection and an LDA classifier. The same highest phi of 0.44 was also achieved with variance features, classifier-dependent feature-selection technique, and an LDA classifier. The same highest mean AUC_{PR} and AUC_{ROC} of 0.48 and 0.95, respectively, were also achieved with entropy features, classifier-independent feature-selection, and an LSVM classifier.

7.1.2 Bivariate features

In terms of phi, AUC_{PR}, and AUC_{ROC}, non-normalized features (covariance, cross-spectral power, joint entropy), irrespective of feature-selection technique and classifier, were superior ($p \leq 0.047$) to their corresponding normalized counterparts (correlation, coherence, mutual information).

In the bivariate and non-normalized subgroup, with classifier-dependent feature-selection technique and an LDA classifier, joint entropy features compared to cross-spectral power were marginally better in phi (i.e., 0.47 vs 0.44, $p = 0.078$) and AUC_{ROC} (i.e., 0.50 vs 0.46, $p = 0.078$). Apart from this, joint entropy features with both feature-selection techniques and classifiers were superior ($p \leq 0.047$) to their counterparts (i.e., covariance and cross-spectral power) on all metrics of performances.

The highest performance metrics (phi = 0.47, AUC_{PR} = 0.50, AUC_{ROC} = 0.95), in the subgroup, were achieved with joint entropy features, classifier-dependent feature-selection technique, and an LDA classifier. The same mean AUC_{ROC} of 0.95 was also achieved with joint entropy features, the classifier-independent feature-selection technique, and an LDA classifier.

In the bivariate and normalized subgroup, correlation features with classifier-dependent feature-selection technique and an LDA classifier were superior ($p \leq 0.047$) to PSI, on all metrics of performances. Compared to mutual information, correlation features resulted in marginally better phi (i.e., 0.26 vs 0.18, $p = 0.093$) and compared to coherence, correlation

Table 7.1 State prediction (τ = 0.25 s) performances (mean ± SE) of feature sets with an LDA classifier. A bold value indicates the highest performance of each feature set. Wilcoxon signed-rank tests were performed to identify significant performance difference between classifier-dependent and classifier-independent feature-selection methods.

Feature set	Feature type	Classifier-dependent			Classifier-independent		
		Phi	AUC_{PR}	AUC_{ROC}	Phi	AUC_{PR}	AUC_{ROC}
Univariate	Var	0.44±0.09	0.45±0.12*	0.93±0.02	0.42±0.09	0.42±0.12	0.92±0.02
	SP	0.43±0.09*	0.44±0.12	0.93±0.02	0.39±0.09	0.41±0.11	0.91±0.02
	Ent	**0.44 ± 0.09**	**0.47 ± 0.12**	**0.94 ± 0.01**	**0.44 ± 0.09**	**0.48 ± 0.12**	**0.95 ± 0.01**
Bivariate	Cov	0.39±0.09*	0.37±0.12	0.91±0.02	0.35±0.09	0.37±0.11	0.89±0.02
	CSP	0.44±0.09~	0.46±0.11	0.94±0.01	0.40± 0.09	0.43±0.12	0.92±0.02
	JE	**0.47 ± 0.10**	**0.50 ± 0.12**	**0.95 ± 0.02**	**0.42 ± 0.09**	**0.49 ± 0.12**	**0.95 ± 0.01**
	Cor	0.26±0.08	0.29±0.11	0.83±0.04*	0.21±0.09	0.24± 0.12	0.73±0.06
	COH	0.18±0.05*	0.19±0.09*	0.79±0.03*	0.10±0.05	0.15±0.08	0.64±0.05
	MI	0.19±0.07*	0.24±0.10*	0.78±0.04*	0.14±0.06	0.19±0.08	0.70±0.05
	PSI	0.18±0.06	0.20±0.10	0.76±0.04	0.15±0.05	0.18±0.08	0.71 ±0.03
Multivariate	CSP	**0.44 ± 0.09**	**0.45 ± 0.11**	**0.95 ± 0.01**	**0.45 ± 0.09**	**0.49 ± 0.12**	**0.94 ± 0.02**
	COH	0.21±0.06	0.22±0.09	0.79±0.02~	0.14 ±0.04	0.17±0.08	0.73±0.02
	iCOH	0.10±0.04	0.15±0.06	0.67±0.04	0.10±0.03	0.14±0.06	0.66±0.04
	pCOH	0.13±0.06	0.17±0.09	0.67±0.03	0.04±0.03	0.10±0.05	0.59±0.05
	PSI	0.15±0.04	0.17±0.08	0.72±0.02*	0.12±0.04	0.15±0.07	0.69±0.05
	PDC	0.14±0.05*	0.18±0.08	0.68±0.04*	0.09±0.04	0.14±0.07	0.61±0.05
	GPDC	0.11±0.02~	0.14±0.07~	0.67±0.02*	0.05±0.03	0.11±0.06	0.55±0.03
	DTF	0.19±0.06	0.23±0.09	0.78±0.03~	0.16±0.06	0.18±0.08	0.72±0.05
	nDTF	0.11±0.05	0.15±0.08	0.65±0.04	0.08±0.05	0.14±0.08	0.60±0.04
	ffDTF	0.08±0.04*	0.13±0.06	0.60±0.04	0.05±0.04*	0.11±0.05	0.54±0.04
	dDTF	0.09±0.04*	0.13±0.06~	0.62±0.04*	0.04±0.04	0.10±0.05	0.54±0.04

Wilcoxon signed-rank test: ~$p < 0.1$,* $p < 0.05$

features gave marginally better AUC_{PR} (i.e., 0.29 vs 0.19, $p = 0.09$). However, with an LSVM classifier, there was no within this subgroup significant performance difference.

With the classifier-independent feature-selection technique and an LDA classifier, correlation features, in terms of phi and AUC_{PR}, were superior ($p \leq 0.047$) to coherence. However, with an LSVM classifier, correlation features compared to coherence were marginally better in phi (i.e., 0.21 vs 0.09, $p = 0.062$) and AUC_{PR} (i.e., 0.23 vs 0.15, $p = 0.062$).

The highest mean performance metrics (phi = 0.26, AUC_{PR} = 0.29, AUC_{ROC} = 0.83), in the subgroup, were achieved with correlation features, classifier-dependent feature-selection, and an LDA classifier.

7.1.3 Multivariate features

In terms of phi, AUC_{PR}, and AUC_{ROC}, cross-spectral power features across both feature-selection techniques and classifiers were superior ($p \leq 0.009$) to both multivariate non-causal and causal counterparts. The highest phi (0.45) and AUC_{PR} (0.49) were achieved with cross-spectral power features, classifier-independent feature-selection, and an LDA classifier. Whereas, the highest AUC_{ROC} of 0.95 was achieved with classifier-dependent feature-selection and with both

Table 7.2 State prediction (τ = 0.25 s) performances (mean ± SE) of feature sets with an LSVM classifier. A bold value indicates the highest performance of each feature set. Wilcoxon signed-rank tests were performed to identify significant performance difference between classifier-dependent and classifier-independent feature-selection methods.

Feature set	Feature type	Classifier-dependent			Classifier-independent		
		Phi	AUC_{PR}	AUC_{ROC}	Phi	AUC_{PR}	AUC_{ROC}
Univariate	Var	0.42±0.10	0.42±0.12	0.93±0.02	0.41±0.09	0.41±0.12	0.92±0.02
	SP	0.40±0.09	0.40±0.12	0.91±0.02	0.40±0.09	0.41±0.12	0.92±0.02
	Ent	**0.41 ± 0.09**	**0.44 ± 0.12**	**0.93 ± 0.02**	**0.43 ± 0.10**	**0.48 ± 0.12**	**0.95 ± 0.01**$^{\sim}$
Bivariate	Cov	0.38±0.10$^{\sim}$	0.35±0.12	0.90±0.02	0.34±0.10	0.34±0.12	0.88±0.03
	CSP	0.42±0.10*	0.44±0.12	0.93±0.02	0.39±0.10	0.42±0.12	0.91±0.02
	JE	**0.44 ± 0.10**	**0.47 ± 0.13**	**0.94 ± 0.02**	**0.42 ± 0.10**	**0.48 ± 0.13**	**0.94 ± 0.02**
	Cor	0.24±0.08	0.25±0.11	0.81±0.04	0.21±0.09	0.23±0.12	0.74±0.05
	COH	0.17±0.05*	0.19±0.09*	0.78±0.03*	0.09±0.04	0.15±0.08	0.63±0.05
	MI	0.21±0.07*	0.24±0.10*	0.80±0.03*	0.15±0.07	0.20±0.09	0.71±0.05
	PSI	0.19±0.05*	0.21±0.10*	0.80±0.03*	0.15±0.05	0.18±0.08	0.72±0.02
Multivariate	CSP	**0.44 ± 0.09**	**0.45 ± 0.11**	**0.95 ± 0.01**	**0.42 ± 0.10**	**0.47 ± 0.12**	**0.94 ± 0.02**
	COH	0.22±0.07	0.23±0.01	0.81±0.02*	0.15±0.04	0.17±0.07	0.74±0.03
	iCOH	0.11±0.04	0.15±0.07	0.69±0.04	0.10±0.03	0.14±0.06	0.67±0.05
	pCOH	0.12±0.05	0.16±0.08	0.69±0.02	0.03±0.03	0.10±0.05	0.60±0.06
	PSI	0.13±0.04	0.16±0.08	0.69±0.02	0.08±0.02	0.14±0.05	0.68±0.04
	PDC	0.14±0.06*	0.18±0.09*	0.70±0.04*	0.05±0.03	0.13±0.07	0.61±0.05
	GPDC	0.12±0.04$^{\sim}$	0.15±0.07 $^{\sim}$	0.69±0.02*	0.05±0.03	0.11±0.06	0.54±0.04
	DTF	0.21±0.06	0.24±0.09$^{\sim}$	0.81±0.02*	0.17±0.06	0.19±0.09	0.74±0.04
	nDTF	0.10±0.05	0.14±0.08	0.64±0.05	0.08±0.04	0.14±0.07	0.60±0.05
	ffDTF	0.07±0.04	0.12±0.06	0.58±0.04*	0.05±0.03	0.11±0.05	0.55±0.04
	dDTF	0.07±0.04	0.13±0.06	0.58±0.04*	0.05±0.03	0.10±0.05	0.55±0.04

Wilcoxon signed-rank test: $^{\sim}p < 0.1$,* $p < 0.05$

classifiers.

Coherence features with classifier-dependent feature-selection and both classifiers, in terms of all metrics of performances, were superior ($p \leq 0.047$) to their non-causal and normalized (iCOH, pCOH, PSI) counterparts. Similarly, with classifier-independent feature-selection and an LSVM classifier, phi and AUC_{PR} of coherence were better ($p \leq 0.031$) than both pCOH and PSI.

The highest mean performance metrics (phi = 0.22, AUC_{PR} = 0.23, AUC_{ROC} = 0.81), in the subgroup (normalized and non-causal), were achieved with coherence features, classifier-independent feature-selection, and an LSVM classifier.

In the causal (PDC, GPDC, DTF, nDTF, dDTF, ffDTF) subgroup, the highest mean performance metrics (phi = 0.21, AUC_{PR} = 0.24, AUC_{ROC} = 0.81) were achieved with DTF features, classifier-independent feature-selection, and an LSVM classifier. Except for PDC, performances of DTF features with both feature-selection and classifiers were better ($p \leq 0.05$) than their causal counterparts. Irrespective of the feature-selection technique and classifier, both ffDTF and dDTF gave the lowest performance metrics.

7.1.4 Comparison between univariate and non-normalized bivariate features

With the classifier-dependent feature-selection and both classifiers, variance features, in terms of phi and AUC_{PR}, were superior ($p \leq 0.047$) to covariance features. With the classifier-independent feature-selection and an LDA classifier, phi of variance features was marginally better (i.e., 0.42 vs 0.35, $p = 0.07$) than covariance features. With an LSVM classifier, variance features in terms of AUC_{PR} were marginally better (i.e., 0.41 vs 0.34, $p = 0.063$) than covariance features.

The performances of both spectral and cross-spectral power features, irrespective of feature-selection technique and classifier, were similar.

With the classifier-dependent feature-selection, compared to entropy, performances of joint entropy features with an LDA were higher, i.e., phi (0.47 vs 0.44), AUC_{PR} (0.50 vs 0.47), and AUC_{ROC} (0.95 vs 0.94). Similarly, with an LSVM classifier, performances of joint entropy were higher than entropy features, i.e., phi (0.44 vs 0.41), AUC_{PR} (0.47 vs 0.44), and AUC_{ROC} (0.94 vs 0.93). However, the difference in performance between entropy and joint entropy was not statistically significant. With the classifier-independent feature-selection and both classifiers, performances of entropy and joint entropy were similar.

7.1.5 Comparison between bivariate and multivariate features

AUC_{ROC} of multivariate cross-spectral power features, with the classifier-independent feature-selection and an LSVM classifier, was better ($p = 0.016$) than corresponding bivariate cross-spectral power. Apart from this and despite apparent differences in the mean performance metrics, both bivariate and multivariate cross-spectral power and coherence features were statistically no different. However, all performance metrics of bivariate PSI features with classifier-dependent feature selection and an LSVM classifier were superior ($p \leq 0.009$) to its multivariate counterpart. AUC_{ROC} of bivariate PSI with the classifier-dependent feature-selection and an LDA classifier was marginally better (i.e., 0.76 vs 0.72, $p = 0.078$) than its multivariate counterpart.

7.1.6 Comparison between classifiers

Performances of both classifiers, across 21 feature sets and both feature-selection techniques, are presented in Table 7.1, Table 7.2, and Figure 7.1. With the classifier-dependent feature-selection technique, the LDA classifier, in terms of AUC_{PR} was superior ($p = 0.013$) to the LSVM classifier. Whereas, phi and AUC_{ROC} of both classifiers were no different. Similarly, with classifier-independent feature-selection, the performances of both classifiers were similar.

7.1.7 Feature-selection techniques

The average number of features selected from 8 training sets (i.e., concatenated data from 8 sets of 7 training subjects) of each feature type by classifier-dependent and classifier-independent techniques, are shown in Table 7.3.

In terms of phi, AUC_{PR}, and AUC_{ROC}, classifier-dependent feature-selection, across both classifiers and most of the normalized feature sets, performed better than its counterpart, i.e., classifier-independent. However, on non-normalized and non-causal features, performances of both feature-selection techniques and across both classifiers were similar, as shown in Table 7.1, Table 7.2, and Figure 7.1.

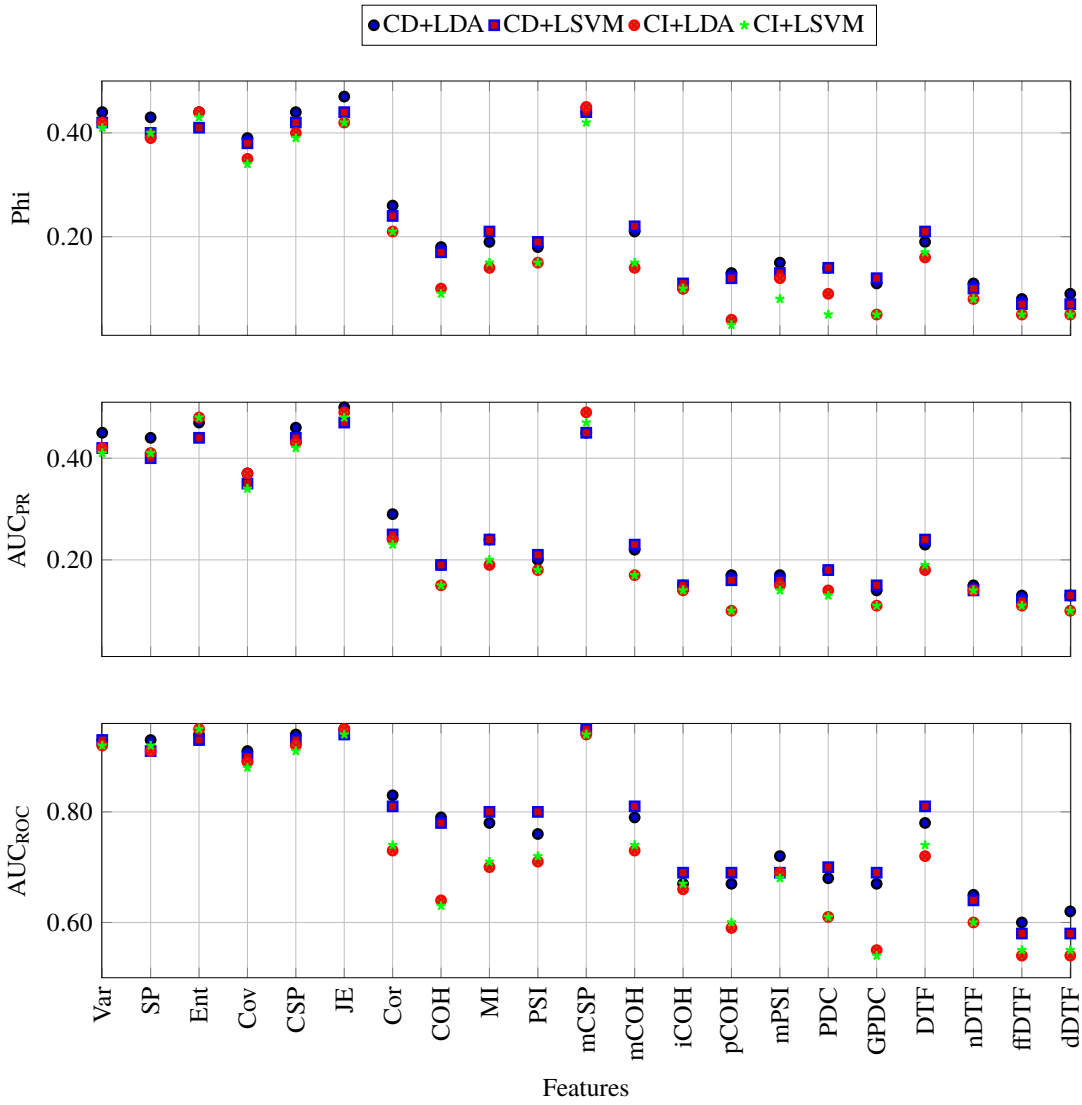

Figure 7.1 Comparison between feature-selection techniques and classifiers on various features at predicting (τ = 0.25 s) microsleep states. Classifier-dependent and classifier-independent were abbreviated as CD and CI, respectively.

7.1.8 Discriminatory features

Event duration and, subsequently, the number of states for both responsive and microsleeps, shown in Table 4.1 varied considerably between sessions and subjects. This leads to a quite imbalance ratios in the training data, which resulted in different features being selected in each

Table 7.3 Average number of features selected by each feature-selection technique for each feature set.

Feature set	Feature type	Classifier-dependent	Classifier-independent
Univariate	Var	33	47
	SP	36	47
	Ent	34	5
Bivariate	Cov	69	51
	CSP	64	48
	JE	47	20
	Cor	69	57
	COH	69	62
	MI	70	62
	PSI	70	69
Multivariate	CSP	40	47
	COH	69	62
	iCOH	70	70
	pCOH	64	19
	PSI	69	26
	PDC	70	62
	GPD	69	28
	DTF	30	12
	nDTF	69	46
	ffDTF	28	6
	dDTF	31	6

training iteration. The top 5 discriminatory features selected at least 6 times out of 8 training sets were used to represent the brain regions and the corresponding frequency band at classifying microsleep and responsive states. In addition, only top features from the feature types which resulted in their mean test $AUC_{ROC} > 0.75$ are presented in this section.

As mentioned in Section 6.4, classifier-dependent feature-selection involved Fisher-score based ranking and used AUC_{ROC} as the cost function. Whereas, classifier-independent feature-selection ranked features per their relevances with class labels and used mutual information as the cost function.

For univariate features (variances, spectral powers, and entropies) O1, O2, P3, and P4 in the theta band were always the top informative/discriminatory features. The top features selected by both classifier-independent and classifier-dependent feature-selection techniques were the same.

From the bivariate and non-normalized features, covariances between O1-O2, O1-P3, O1-T5, and O1-F4 from theta frequency band were the top discriminatory features across all training iterations. Except for the covariance between O1-F4, top feature selected by both feature-selection techniques were the same. Cross-spectral powers between O2-O1 and O2-P3 from theta frequency band were always the top discriminatory features, selected by both feature-selection

techniques. However, the top discriminatory cross-spectral powers between O2-P4, O2-C4, and P3-T6 from the theta band were only selected by classifier-dependent feature-selection. Joint entropies between T6-O2, T6-P4, P3-P4, and F7-O2 from theta frequency band, as per classifier-dependent feature-selection technique, were the top features. Whereas, as per classifier-independent feature-selection, joint entropies between O1-O2, O1-P4, O2-P3, O2-T6, and O2-P4 from the theta band were the top features.

Correlations between F7-F8 and T6-O2 from the delta and theta bands respectively were the top discriminatory features. Coherence between F7-F8 from the beta band and between F7-T4, and F8-T3 from the delta band were top features. Mutual information between F7-F8 from delta, T6-O2 from theta, and P3-P4 from the alpha bands were the most frequently selected top discriminatory features. PSIs between F7-F8 and F3-C4 from gamma and between T6-O2 from the theta band were the most frequently occurring informative features. All of these features were selected by classifier-dependent feature-selection, as it, for bivariate and normalized features, resulted in $AUC_{ROC} > 0.75$.

From multivariate group, spectral powers at left and right occipital, and at left parietal regions from the theta band and across all training folds were the most discriminatory features, selected by classifier-dependent feature-selection. However, as per classifier-independent feature-selection technique, cross-spectral powers between left occipital and left parietal region were the top features. Spectral powers (diagonal of $\mathbf{S}$, see Equation (3.11)) left and right parietal, and left and right occipital regions, from the theta band, were also the top features.

Multivariate coherence between the right frontal region and occipital region, from theta band, consistently across training folds, was the top discriminatory feature. Self-loop DTFs at right parietal, occipital, and temporal regions from the theta band, across the training folds, were the top discriminatory features. All of these features were selected by classifier-dependent feature-selection technique.

In summary, irrespective of feature-selection technique and feature-type, most discriminatory features for prediction of microsleep states were from the theta band and occipital, parietal, and temporal regions of the brain.

7.2 PERFORMANCE OF A SINGLE FEATURE SET

Irrespective of the feature-selection technique and classifier, prediction of microsleep states from bivariate normalized (correlation, coherence, mutual information, PSI), multivariate normalized (coherence, iCOH, pCOH, PSI), and causal (PDC, GPDC, DTF, nDTF, dDTF, ffDTF) features was relatively poor, as shown in Table 7.1 and Table 7.2. Joint entropy features resulted in the overall highest performance metrics (phi = 0.47, $AUC_{PR} = 0.50$, $AUC_{ROC} = 0.95$) at predicting microsleep states.

On normalized and non-causal features, performance metrics of both feature-selection techniques were comparable. However, the overall prediction performance of classifier-dependent

feature-selection, and across both classifiers as shown in Table 7.1 and Table 7.2 was better than the classifier-independent feature-selection. On the other hand, compared to the LSVM classifier, mean (over individual feature types) phi and AUC_{PR} of LDA were slightly higher, i.e., phi (0.25 vs 0.24) and AUC_{PR} (0.27 vs 0.26).

For clarity and ease of reading, detection and prediction performances of joint entropy features with classifier-dependent feature-selection, and an LDA classifier on microsleep states and onsets, are only reported. The interested reader, however, can find microsleep state and onset prediction performances of non-normalized and non-causal features, with both classifier-independent feature-selection and classifiers in Appendix A.

7.2.1 Detection and prediction of microsleep states

A decline in performance metrics (phi, AUC_{PR}, AUC_{ROC}) with respect to increasing prediction time τ on microsleep states, are shown in Figure 7.2. The mean detection and prediction (τ = 1 s) performance metrics were (phi = 0.48, AUC_{PR} = 0.51, AUC_{ROC} = 0.95) and (phi = 0.44, AUC_{PR} = 0.46, AUC_{ROC} = 0.94), respectively. With increasing prediction time τ, compared to AUC_{ROC}, drop in both phi and AUC_{PR} was faster.

7.2.2 Detection and prediction of microsleep onsets

A decline in performance metrics with respect to increasing prediction time τ on microsleep onset (refer Figure 4.5) are shown in Figure 7.3. The mean detection and prediction (τ = 5 s) performance metrics were (phi = 0.11, AUC_{PR} = 0.09, AUC_{ROC} = 0.91) and (phi = 0.05, AUC_{PR} = 0.01, AUC_{ROC} = 0.79), respectively. With increasing prediction time τ, both phi and AUC_{ROC} dropped linearly, while for AUC_{PR}, the drop was exponential.

Compared to log-power spectral features with an LSVM classifier (Shoorangiz 2018), microsleep onset detection phi achieved in this research was slightly higher (i.e., 0.11 vs 0.08) , while both detection AUC_{PR} and AUC_{ROC} were the same, i.e., 0.09 and 0.91, respectively.

7.3 PERFORMANCES OF MULTIPLE FEATURES

All of the 21 features (i.e., 3 univariate, 7 bivariate, and 11 multivariate) shown in Figure 6.1, Table 7.1, and Table 7.2 were fused at feature-level and decision-level, as shown in Figure 6.3.

The performances of multiple features with both feature-selection techniques and classifiers can be found in Appendix B.

7.3.1 Detection and prediction of microsleep states

The overall performance of the microsleep state prediction system, with single joint entropy features, feature-level fusion, and decision-level fusion, and with respect to increasing prediction times τ, are shown in Figure 7.4.

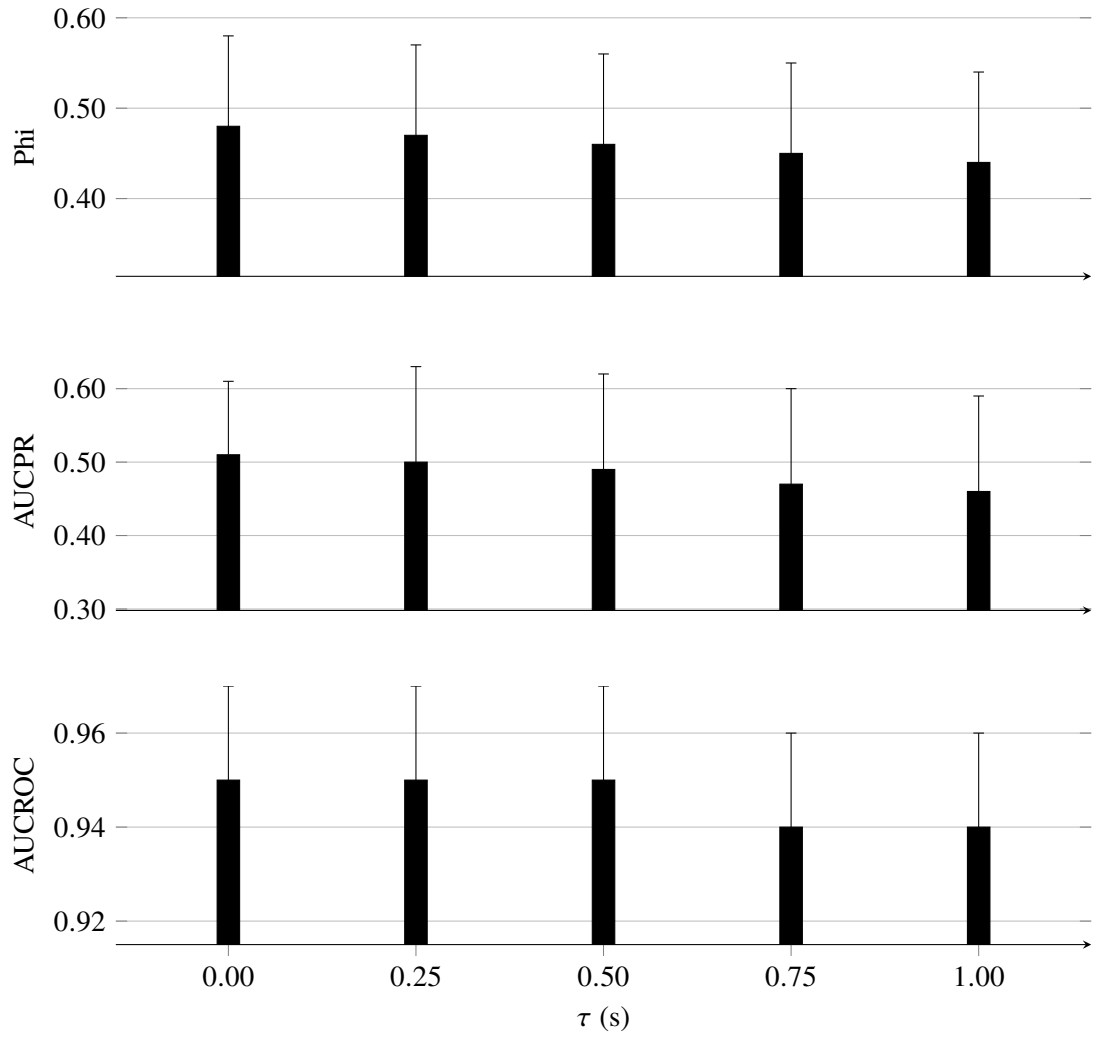

Figure 7.2 Performances (mean ± SE) of microsleep state prediction system for τ = 0.00–1.00 s.

In terms of phi, feature-level fusion, across the prediction time τ, resulted in the highest performance. Compared to a single feature set, feature-level fusion resulted in detection and prediction (τ = 1 s) phi of (0.49 vs 0.48) and (0.45 vs 0.44), respectively. In contrast, a single feature set resulted in the highest detection and prediction AUC_{PR} of 0.51 and 0.46, respectively. Both single feature set and feature-level fusion resulted in the same AUC_{ROC} and across the prediction time, but the highest detection AUC_{ROC} of 0.96 was achieved with feature-level fusion.

In summary, decision-level fusion across the prediction times τ, and compared to the single feature set resulted in the poorer performance metrics on microsleep states. A slightly higher phi achieved with feature-level fusion was at a comparative cost of a lower AUC_{PR}.

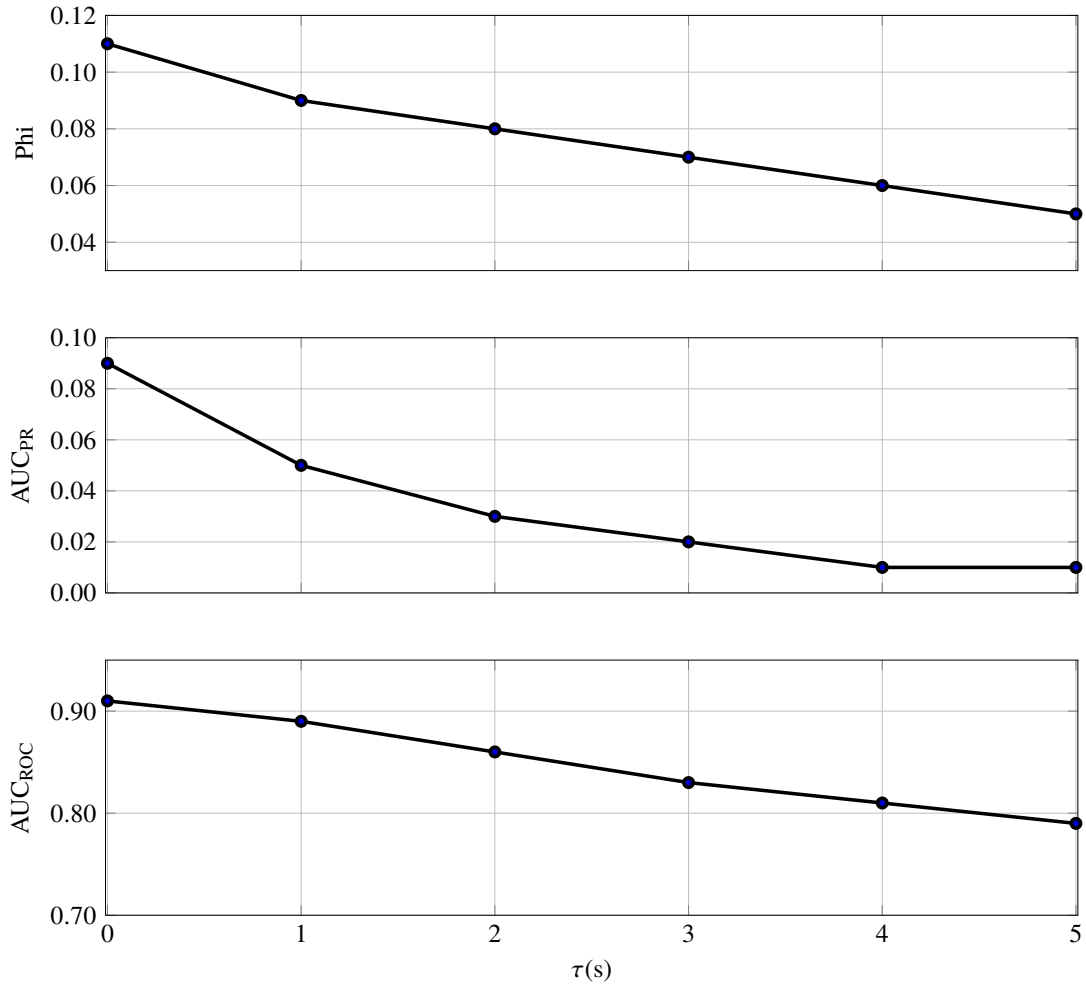

Figure 7.3 Performances of microsleep onset prediction system, for τ = 0–5 s.

7.3.2 Detection and prediction of microsleep onsets

Figure 7.5 shows the performances of the microsleep onset prediction system with single joint entropy features, feature-level fusion, and decision-level fusion, and with respect to increasing prediction times τ.

Contrary to state prediction and except for τ = 1–2 s, decision-level fusion, in terms of phi, gave the best performance. AUC_{PR} achieved with both single feature set and feature-level fusion was higher than the decision-level fusion for τ = 0–1 s but lower for τ = 4–5 s.

The best detection onset phi of 0.11 was achieved with both single feature set and decision-level fusion. Whereas all of these gave the same prediction (τ = 1 s) phi of 0.09. Both single feature set and feature-level fusion gave the highest detection and prediction AUC_{PR} of 0.09 and 0.05, respectively and detection and prediction AUC_{ROC} of 0.91 and 0.89, respectively.

As with microsleep states, the overall performance of both of the fusion techniques, on microsleep onsets, were not better than the performance achieved with a single feature set.

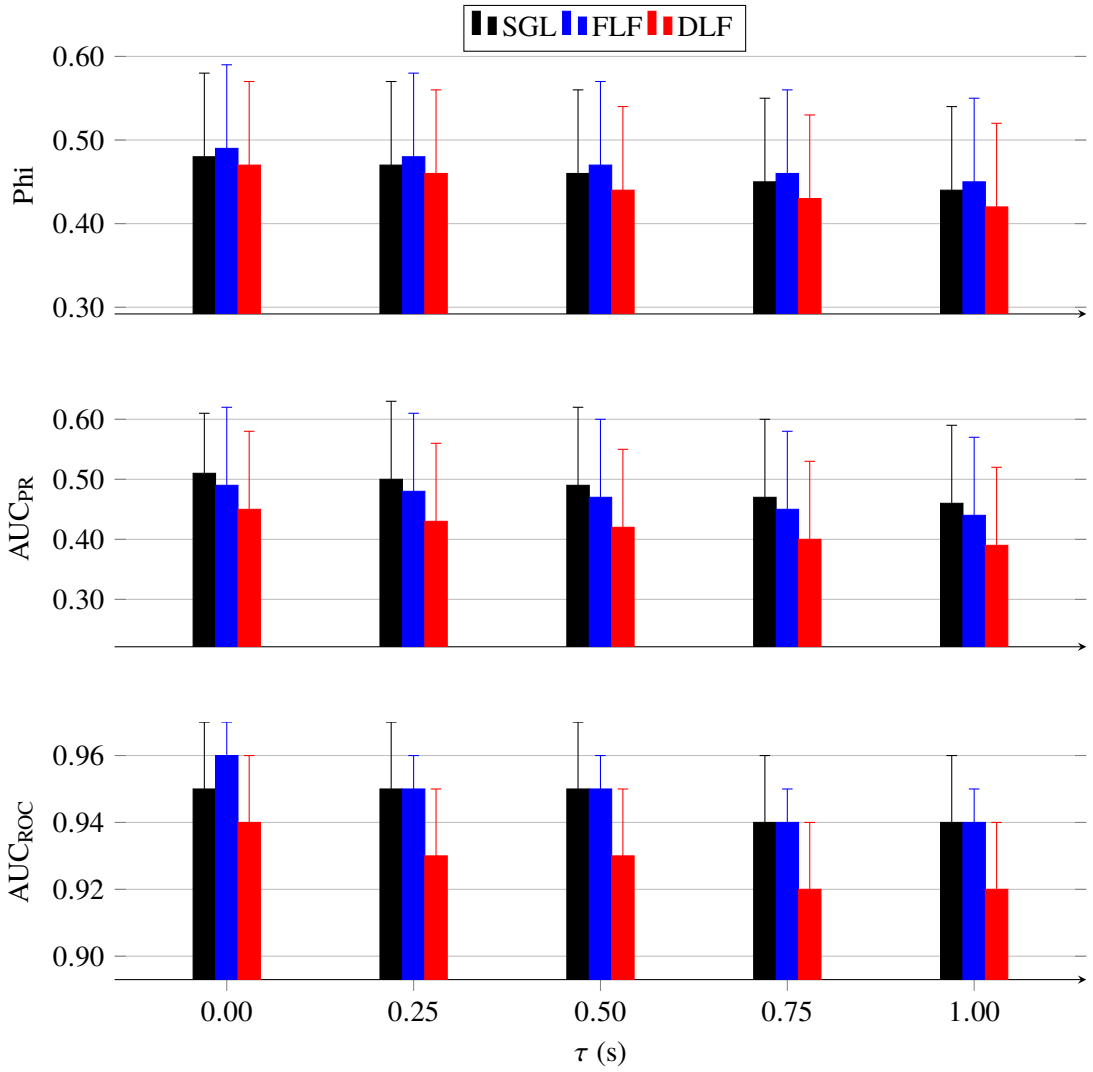

Figure 7.4 Performances (mean ± SE) of microsleep state prediction system with single feature set (SGL), feature-level fusion (FLF), and decision-level fusion (DLF) for τ = 0.00 –1.00 s.

7.4 PERFORMANCES OF ENSEMBLE CLASSIFICATIONS

Four ensembles (i.e., majority voting, diversity-incorporated majority voting (DMV), rank-based weighted majority voting (RMV), and overlapping clusters) as mentioned in Section 6.7 were formed, where the data of individual subjects were the clusters.

The performance of all of the ensemble classification techniques with joint entropy features, both feature-selection techniques, and classifiers can be found in Appendix C.

7.4.1 Detection and prediction of microsleep states

Mean performance metrics of a classifier together with four ensemble techniques (i.e., majority voting, DMV, RMV, and overlapping clusters) at predicting microsleep states for increasing prediction time τ = 0–1 s are shown in Figure 7.6.

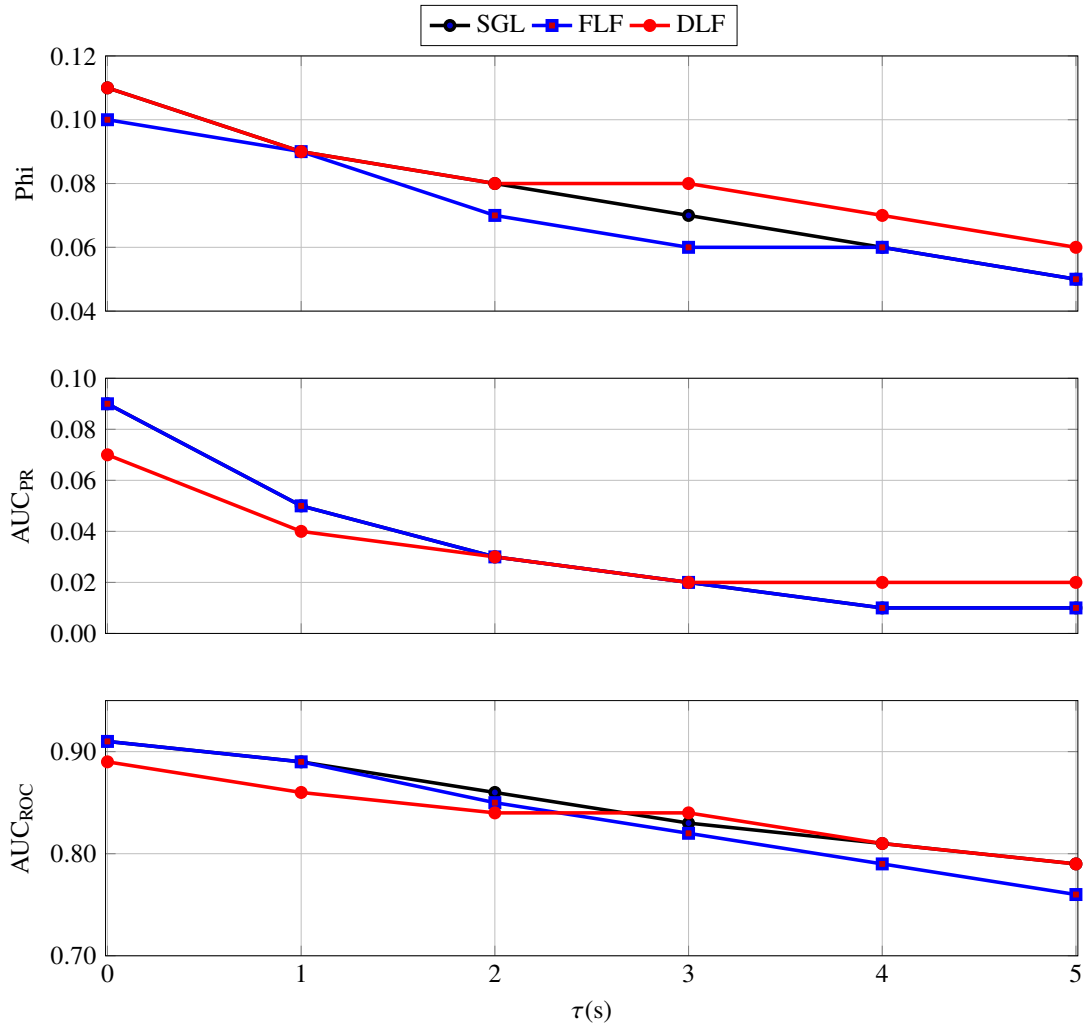

Figure 7.5 Performances of microsleep onset prediction system with single feature set (SGL), feature-level fusion (FLF), and decision-level fusion (DLF) for τ = 0–5 s.

Compared to a single classifier and except for the majority voting, mean detection phi of all of the proposed ensemble classification techniques were higher (i.e., 0.50 vs 0.48). Whereas, the highest mean prediction (τ = 1 s) phi of 0.47 was achieved with the overlapping clusters. The highest detection AUC_{PR} of 0.52 was achieved with both majority voting and overlapping clusters, whereas, the highest prediction AUC_{PR} of 0.47 was achieved with overlapping clusters. Compared to the single classifier, detection AUC_{ROC} of all of the ensemble classifications was the higher (i.e., 0.96 vs 0.95). However, all of the ensemble classifications and the single classifier resulted in the same AUC_{ROC} of 0.94.

In summary, the overall best detection and prediction performances on microsleep states were achieved with overlapping clusters.

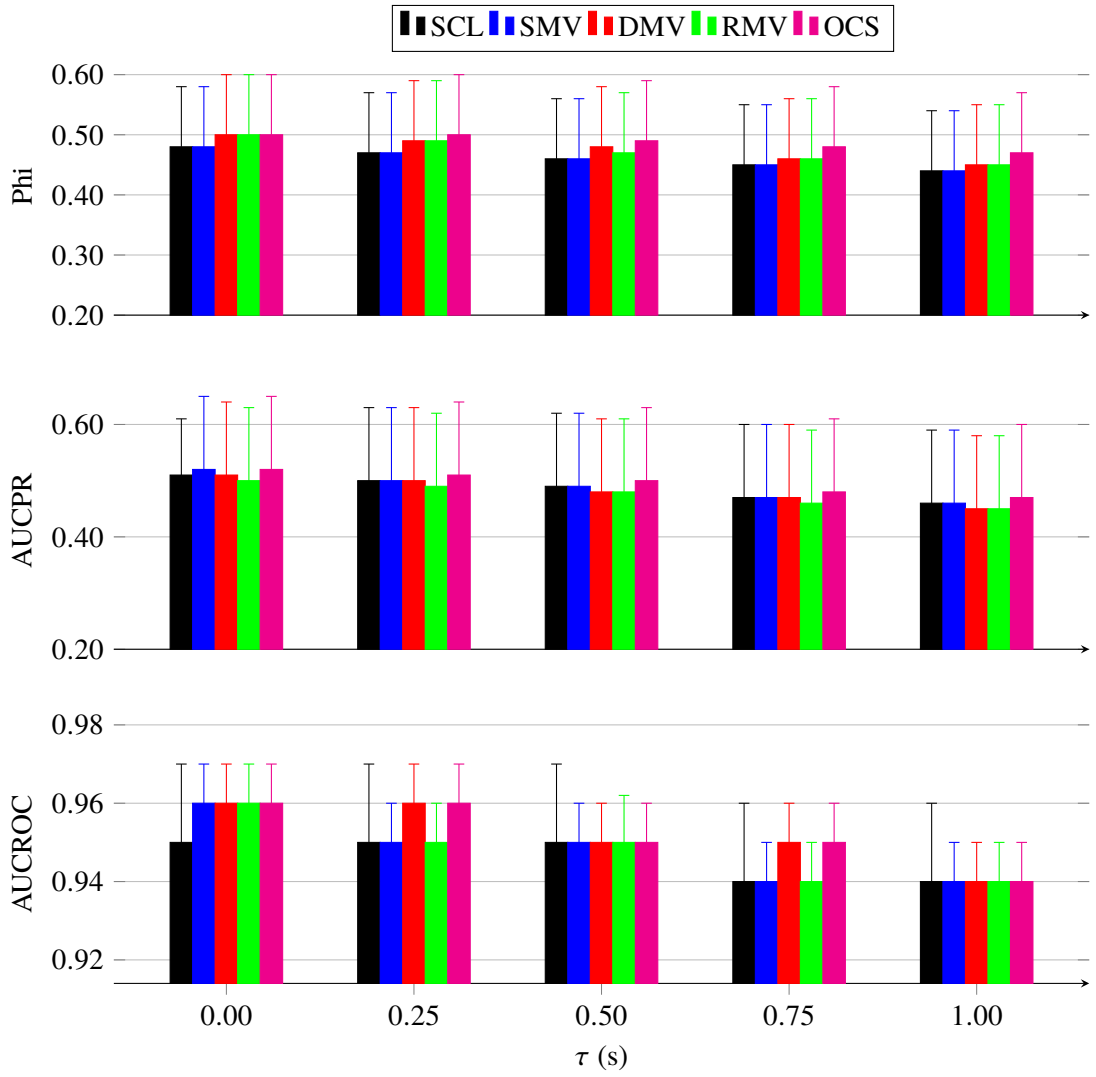

Figure 7.6 Performances (mean±SE) of microsleep state prediction system for τ = 0–1 s. Single classifier, standard majority voting, and overlapping clusters are abbreviated as SCL, SMV, and OCS, respectively.

7.4.2 Detection and prediction of microsleep onsets

Figure 7.7 shows the performance of a microsleep onset prediction system with the single classifier and all of the ensemble techniques with respect to increasing prediction times τ.

Unlike on microsleep states, majority voting gave the highest detection phi of 0.13, whereas, prediction (τ = 1 s) phi of 0.10 was achieved with all of the ensemble techniques. An overall drop in phi was linearly related to the prediction time. Overlapping clusters resulted in the highest detection and prediction AUC_{PR} of 0.10 and 0.07, respectively. However, at τ = 5 s, the single classifier and all of the ensemble classifiers gave the same AUC_{PR} of 0.01. The highest detection and prediction AUC_{ROC}s of 0.93 and 0.91, respectively, were achieved with overlapping clusters. At τ = 5 s, all of the proposed techniques gave the same AUC_{ROC} of 0.80.

In summary, the overall best detection and prediction performance metrics on microsleep

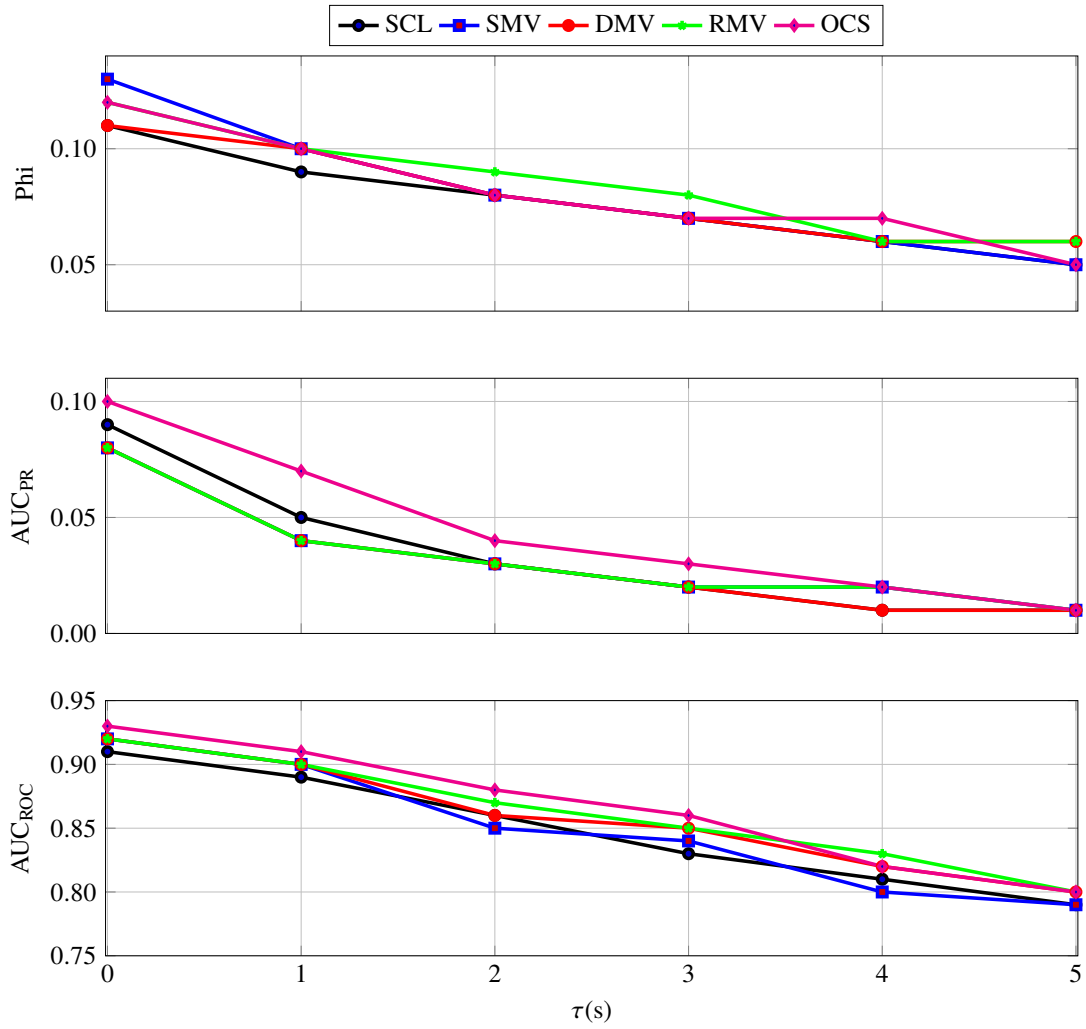

Figure 7.7 Performances of microsleep onset prediction system for τ = 0–5 s. Single classifier, standard majority voting, and overlapping clusters are abbreviated as SCL, SMV, and OCS, respectively.

onsets were achieved with overlapping clusters.

7.5 EFFECT OF FEATURE PREPROCESSING

Feature preprocessing (baseline correction mentioned in Section 6.3) resulted in improved prediction (τ = 0.25 s) performance on microsleep states. Compared to log-power spectral features with mutual information-based greedy feature-selection method and an LDA classifier (Shoorangiz et al. 2016), wavelet spectral power features with classifier-independent feature-selection and an LDA classifier resulted in 18%, 8%, and 1% higher mean phi, AUC_{PR}, and AUC_{ROC}, respectively. Similarly, compared to baseline log-power spectral features with multi-subject factor analysis-based feature reduction method and an LDA classifier (Shoorangiz et al. 2017), performances of wavelet spectral power features with classifier-independent feature-selection and an LDA classifier were higher, i.e., phi (0.39 vs 0.34), AUC_{PR} (0.41 vs 0.38), and

AUC_{ROC} (0.91 vs 0.90).

7.6 SUMMARY

This chapter presented the overall performance metrics of all 21 types of individual features with both feature-selection techniques and both single classifiers at predicting (τ = 0.25 s) microsleep states. The chapter then presented within-group (i.e., univariate, bivariate, and multivariate) performance comparisons of individual feature sets, followed by inter-group comparisons. In addition, comparisons between the feature-selection techniques and classifiers, followed by a discussion on discriminatory features were also presented. Based on these comparisons, only single joint entropy features with classifier-dependent feature-selection and an LDA classifier were selected and then presented thereof detection and prediction performances on microsleep states and onsets. This was followed by detection and prediction performances of two data fusion techniques together with the single feature set on both microsleep states and onsets. Finally, detection and prediction performances of four ensemble techniques together with the single classifier on microsleep states and onsets were presented.

Demeaning of non-normalized features substantially improved the overall prediction performance on microsleep states. Compared to normalized features, all non-normalized and non-causal features were better at predicting microsleep states. Overall, joint entropy was the single best feature set (detection phi = 0.48).

On normalized features, classifier-dependent was better than classifier-independent feature-selection technique, whereas, on non-normalized features, the performance metrics of both feature-selection techniques were similar.

With classifier-independent feature-selection, the performance metrics of both classifiers were similar. With classifier-dependent feature-selection and across all types of features, phi and AUC_{PR} of LDA classifier were higher (0.25 vs 0.24) and (i.e., 0.27 vs 0.26, p = 0.013) than its LSVM counterpart.

Feature-level fusion, compared to single joint entropy feature set, gave higher phi but lower AUC_{PR} on microsleep states. Overall, both fusion techniques were rather poorer than single feature set at predicting microsleep states and onsets.

The overall best microsleep states and onsets performance metrics were achieved with the overlapping clusters (detection phi = 0.50).

Overall mean detection performance metrics of the best performing single feature set (joint entropy) with a single classifier, fusion technique (feature-level), and the ensemble classifications (overlapping clusters) on detection and prediction (τ = 1 s) of microsleeps (state and onset) are respectively presented in Table 7.4 and Table 7.5.

For the completeness, in addition to phi, AUC_{PR}, and AUC_{ROC}, 5 metrics (i.e., specificity (sp), sensitivity (sn), precision (pr), F-measure (Fm), geometric mean (GM)) based on confusion matrix (refer Table 6.3) were presented.

Table 7.4 Summary of the detection performance. All performance metrics are the mean values of 8 test subjects.

	Sn	Sp	Pr	Fm	GM	Phi	AUC_{PR}	AUC_{ROC}
Microsleep states								
Single Feature	0.72	0.96	0.43	0.46	0.82	0.48	0.51	0.95
Fusion	0.70	0.96	0.45	0.48	0.81	0.49	0.49	0.96
Ensemble	0.78	0.96	0.43	0.48	0.86	0.50	0.52	0.96
Microsleep onsets								
Single Feature	0.74	0.89	0.03	0.06	0.81	0.11	0.09	0.91
Fusion	0.73	0.89	0.02	0.04	0.79	0.10	0.09	0.92
Ensemble	0.76	0.90	0.03	0.05	0.82	0.12	0.10	0.93

Table 7.5 Summary of the prediction (τ = 1 s) performance. All performance metrics are the mean values of 8 test subjects.

	Sn	Sp	Pr	Fm	GM	Phi	AUC_{PR}	AUC_{ROC}
Microsleep states								
Single Feature	0.70	0.95	0.40	0.43	0.80	0.44	0.46	0.94
Fusion	0.66	0.95	0.40	0.44	0.78	0.45	0.49	0.94
Ensemble	0.73	0.95	0.41	0.45	0.83	0.47	0.47	0.94
Microsleep onsets								
Single Feature	0.71	0.87	0.02	0.04	0.78	0.09	0.05	0.89
Fusion	0.70	0.87	0.02	0.03	0.77	0.09	0.06	0.89
Ensemble	0.73	0.88	0.02	0.04	0.80	0.10	0.07	0.91

As can be seen from Table 7.4, all of the approaches resulted in moderately high level of sensitivity and GM at detecting both microsleep states and onsets, whereas, precision, Fm, and phi of all of the approaches, due to class imbalance, were low.

Comparing Table 7.4 and Table 7.5, it is evident that with increasing prediction time τ, all metrics of performance exhibited drop. However, metrics representing minority class (i.e., microsleep) and sensitive to class-imbalance ratio (i.e., sensitivity, precision, phi, Fm, AUC_{PR}) showed faster drop compared to metrics representing majority class (responsive) and insensitive to class-imbalance ratio (i.e., specificity, GM, AUC_{ROC}).

Chapter 8

DISCUSSION

The prediction (τ = 0.25 s) performance metrics of all 21 individual features (refer Figure 6.1) on microsleep state were evaluated in Section 7.1. The performance of microsleep prediction system with the best performing single feature set, multiple features (dat fusion), and proposed ensemble classification techniques were presented in Chapter 7. The aim of this chapter is to discuss/comment on the results.

8.1 EFFECT OF DEMEANING

Compared to recent work of Shoorangiz et al. (2016) on the same data, prediction performances of spectral power features with classifier-dependent feature-selection and an LDA classifier, on microsleep states, were higher, i.e., phi (0.43 vs 0.33), AUC_{PR} (0.44 vs 0.38), and AUC_{ROC} (0.93 vs 0.90). Similar performances were achieved with both feature-selection techniques and classifiers. Similarly, compared to baseline log-power spectral features with multi-subject factor analysis-based feature reduction method and an LDA classifier (Shoorangiz et al. 2017), performances of wavelet spectral power features with classifier-independent feature-selection and an LDA classifier were higher, i.e., phi (0.39 vs 0.34), AUC_{PR} (0.41 vs 0.38), and AUC_{ROC} (0.91 vs 0.90). Irrespective of feature-selection and classifier, improvements in all metrics of performances indicate that the baseline correction (demeaning) addressed inter-session and inter-subject variability in the data (features).

The standard normalization process requires the mean and variance of the full data and, therefore, is not practical in real-time implementations (Peiris et al. 2011). The effect of noise (outliers) on the training data is global and the test data normalized with respect to the noisy training data also become noisy. This means, a portion of noisy data can have an effect on both training and test data. In contrast, demeaning the data with respect to the first 2 min of each session is local with respect to that session. In the case, noise presented in the first 2-min can only affect data of the respective session. Furthermore, with this technique, training and test data were preprocessed independently.

8.2 PERFORMANCES OF INDIVIDUAL FEATURES

For clarity and ease of reading, features were categorized into different groups (i.e., non-normalized and normalized, univariate, bivariate, and multivariate) as shown in Figure 6.1. The performances between and within groups discussed in this section are based on prediction of microsleep states at $\tau = 0.25$ s.

8.2.1 Non-normalized and normalized features

As shown in Table 7.1 and Table 7.2, prediction performances of both bivariate and multivariate normalized features on microsleep states, were poor. The poorer performances of normalized features, are considered to be due to the inherent property of being scaled (amplitude)-invariant and, consequently, a loss of classification-related information. Furthermore, irrespective of task, brain regions are likely to be synchronized at times. A change in the level of such synchronization may only occur or associated with some rarely occurring or abrupt events (e.g., eye blink). In such a scenario, brain regions can be considered as cognitive transmitter and receivers while the communication can be asynchronous, simplex, or half-duplex. In addition, different brain regions are functionally connected in a non-homogeneous way, and specific connectivity patterns emerge during the wakeful resting state (Rocca et al. 2014).

Conversely, non-normalized features are scale-variant, where changes in EEG amplitude and frequency are directly correlated with behavioural performances and circadian rhythms (Huang et al. 2008, Melia et al. 2015, Rocca et al. 2014). Non-normalized features, being sensitive to such variations, can consequently have better microsleep-related classification information. This is further confirmed by similar detection/prediction performances of corresponding univariate features extracted from individual channels of EEG.

Despite being widely used in off-line characterization of sleep-related states of unconsciousness, such as drowsiness (Awais et al. 2017, Dissanayaka et al. 2015), microsleeps (Toppi et al. 2016), and fatigue (Dimitrakopoulos et al. 2018), normalized inter-channel relationships (both functional and effective connectivity) were found to be inappropriate in on-line prediction of microsleep states. In contrast, non-normalized (both bivariate and multivariate) inter-channel relationships resulted in similar performances to their corresponding univariate features extracted from individual EEG channels.

8.2.2 Univariate and bivariate features

With classifier-dependent feature-selection and across both classifiers, except for covariance, prediction performances of both bivariate features (i.e., cross-spectral power and joint entropy) on microsleep states were slightly higher than their corresponding univariate features (i.e., spectral power and entropy) as shown in Table 7.1 and Table 7.2. Such performance differences can be considered as bivariate features, compared to univariate features, have more microsleep-related information. In addition, from machine learning perspectives, compared to univariate, bivariate

features have higher dimensions, which can be considered to have resulted in less biased models (Le Borgne 2005) and consequently, comparatively higher performances.

8.2.3 Bivariate and multivariate features

The bivariate approach does not discard the common effect of a third simultaneously acquired signal and, consequently, compared to multivariate approach, results in lower precision (Blinowska et al. 2004) and is less accurate (Kuś et al. 2004). This can be seen from higher mean performance metrics of multivariate cross-spectral power across the feature-selection techniques and classifiers as shown in Table 7.1 and Table 7.2. Similarly, compared to bivariate, mean performance metrics of multivariate coherence, irrespective of feature-selection technique and classifier, were higher. However, mean prediction performances of bivariate PSI, across both feature-selection techniques and classifiers were better than its multivariate counterpart. Nonlinear phase response of MVAR affects multivariate PSI, which, compared to bivariate PSI, might have resulted in inferior performances.

8.2.4 Individual features

Irrespective of feature-selection technique and classifier, both temporal and spectral features (variance and spectral power) in the univariate group resulted in similar performances at predicting microsleep states. Due to computational ease, the wavelet transform in practice is performed as the inverse Fourier transform of a time series multiplied by a scaled-normalized mother wavelet in Fourier space. As a result, the expectation value of power in wavelet domain is equal to the length of a window multiplied by expectation value of power in the Fourier domain, i.e., $\langle |W^X(n,f)|^2 \rangle = N \langle |X_k|^2 \rangle$ (Torrence and Compo 1998). As per Parseval's theorem, the power of a time series in the time domain is the mean power in frequency domain, i.e., $\sum_{n=1}^{N} |x[n]|^2 = \frac{1}{N} \sum_{k=1}^{N} |X_k|^2$. In addition, for a white-noise time series, wavelet spectral power is equal to variance of the time series.

In the non-normalized bivariate subgroup, spectral features (i.e., cross-spectral power) were better than temporal features (i.e., covariance). While in the normalized bivariate subgroup, temporal features (i.e., correlation) were superior to spectral features (i.e., coherence and PSI).

Similar performances of both variance and spectral power features, and superior performances of correlations to both coherence and PSI (refer Table 7.1 and Table 7.2) contradicted the notion that important information in EEG are often encoded in frequency not time (Astolfi et al. 2006). However, inferior performances of covariances to cross-spectral power supported the notion.

Irrespective of feature-selection technique and classifier, highest performances in both univariate and bivariate groups were achieved with entropy and joint entropy features, respectively, as shown in Table 7.1 and Table 7.2. These performances can be considered due to their robustness against noise and ability to capture microsleep-related information from the EEG time series.

Low performances of mutual information, however, is, at least in part, a consequence of normalization. Joint entropy, in terms of AUC_{ROC} and AUC_{PR}, has been reported to outperform both cross-correlation and mutual information at estimating functional connectivity (Garofalo et al. 2009).

iCOH between two EEG time series depends upon their time-lag. A very small or zero time-lag, as can occur in many experiments, will result in zero iCOH indicating missed interactions in the brain (Nolte et al. 2004). The communication between two closely located EEG electrodes can be instantaneous and can therefore be missed by iCOH that consequently can result in loss of information and subsequently, poor performances at predicting microsleep states.

pCOH between two EEG time series X and Y is the fraction of coherence not shared with another source or time series Z. In the case where three or more time series are fully coherent with each other, partialization of coherence between two time series with the remaining time series can lead to a zero coherence. Due to the assumption of linearity, pCOH becomes insensitive to nonlinear interactions between the time series. In addition, it is very sensitive to noise contamination and tends to identify the signal with highest SNR, irrespective of the underlying connectivity, as the driver (Pereda et al. 2005). A low pCOH will be obtained between a time series, that is a linear combination of a set of other time series, and the set (Blinowska 2011). This explains the low performances of pCOH across both the feature selection techniques and classifiers.

Experimental time series, such as EEG, are often noisy and exhibit random phase slips of 2π (Pereda et al. 2005) that result in an incorrect signed phase difference ($\Delta\theta$) (Bastos and Schoffelen 2016, Blinowska 2011) where phases cancel in time average (refer Equation (6.7)) and subsequently result in a small value of PSI (Quiroga et al. 2002). On the other hand, noisy or chaotic time series may synchronize even if their amplitudes remain uncorrelated (Kreuz et al. 2007, Pereda et al. 2005).

Unlike iCOH, pCOH, and PSI, coherence does not discard information and is sensitive to both amplitude/power and phase of the time series and, in multivariate normalized and non-causal (functional connectivity) group, gave highest mean performances in classifying microsleep states from responsive as shown in Table 7.1 and Table 7.2.

Causal inter-channel relationships (effective connectivity) are very sensitive to preprocessing steps. Transformation of MVAR coefficients into frequency domain inherently involves the sampling time ($\Delta t = \frac{1}{f_s}$) of the time series, which is affected by decimation and interpolation that can eventually influence the connectivity leading to wrong conclusions. Compared to coherence, both PDC and DTF have been shown to adversely affected by decimation at detecting non-causal connections Silfverhuth et al. (2012). Filtering disturbs the information contents of the time series and the coefficients of MVAR are therefore generally changed, which eventually result in spurious (false positives) and missed (low sensitivity) causal connections (Barnett and Seth 2016, Florin et al. 2010, Silfverhuth et al. 2012). In addition, it is to be noted that MVAR is an all-pole infinite impulse response (IIR) filter. In the presence of time series containing slow

modes that are ubiquitous in biological systems, MVAR can be unstable and ultimately result in unreliable measures of causality (Friston et al. 2014). Apart from this, lowest performances of ffDTF and dDTF indicate that both feature sets might have not captured microsleep-related connectivity patterns/information. Furthermore, causal connections are asymmetric and the corresponding features, in the presence of non-homogeneously connected brain regions (Rocca et al. 2014), exacerbate the inherent inter-session and inter-subject variabilities in the EEG time series and therefore poor classification performances.

8.2.5 Overall performance of feature-selection techniques and classifiers

On most of the normalized feature types, microsleep state prediction performance of classifier-dependent feature-selection, compared to classifier-independent, was better as shown in Table 7.1 and Table 7.2. Classifier-dependent feature-selection used AUC_{ROC} as the cost function, whereas, classifier-independent feature-selection used mutual information as the cost function (relevance). Poor performance of classifier-independent feature-selection can be considered due to skewed class distributions, which can have affected the overall selection of informative features.

As mentioned in Section 6.5, due to large amount of data and computational complexity of nonlinear classifiers, two linear classifiers were used to validate the efficacy of different channel-wise and inter-channel features of EEG at detecting/predicting microsleep states and onsets. As shown in Table 7.1 and Table 7.2, overall performance of an LDA with the classifier-dependent feature-selection technique were higher than LSVM. Classifier-dependent feature-selection involved Fisher score-based ranking and an LDA-based wrapper (refer Section 6.4.1), which can be considered to have favoured the LDA classifier and consequently, overall higher performances than LSVM classifier. However, with the classifier-independent feature-selection, both classifiers performed similarly across various feature sets.

Compared to LSVM, overall similar and/or superior performances of LDA classifier, support the argument that simple-decision boundaries and estimates via Gaussian models are stable (Hastie et al. 2008) and that this is even more evident in imbalanced datasets.

8.2.6 Discriminatory features

As mentioned in Section 7.1.8, most of the discriminatory features for prediction of microsleep states were from the theta band and from occipital, parietal, and temporal regions of the brain.

Selection of the top non-normalized discriminatory features from the theta sub-band is in accordance with findings that theta activity is associated with microsleeps (Davidson et al. 2007, Harrison and Horne 1996, Jonmohamadi et al. 2016, Toppi et al. 2016) . Similarly, parietal and occipital regions correspond to changes in connectivity during, and prior, to the onset of microsleeps (Toppi et al. 2016) and microsleep-related eye closures (Jonmohamadi et al. 2016). During a continuous compensatory tracking task (CTT), power spectra in the occipital region have been reported to fluctuate in both low- and high-error epochs (Huang et al. 2008). Parietal

cortex, involved in the integration of sensory-motor information and cognition, has been reported to be impaired by sleep (Thomas et al. 2000). Neural activity during microsleeps increases in the parietal, occipital, frontal regions (Poudel et al. 2018).

Importantly, irrespective of the feature selection technique, the top discriminatory normalized features were the connections from frontal regions and different frequency bands. The frontal brain region is reported to have associated with motor processing task, suggesting a change in coordination between regions during the transition from an alert to a drowsy state (Awais et al. 2017).

8.3 DATA FUSION

The best prediction performance metrics of feature-level and decision-level fusion on microsleep states were (phi = 0.48, AUC_{PR} = 0.48, AUC_{ROC} = 0.95) and (phi = 0.46, AUC_{PR} = 0.43, AUC_{ROC} = 0.93), respectively. Compared to the best performance metrics (phi = 0.47, AUC_{PR} = 0.50, AUC_{ROC} = 0.95) of single features (i.e., joint entropy), the overall performances of both fusion techniques were lower.

Poor performances of feature-level fusion indicate that different feature sets were not orthogonal (complementary) to each other. As mentioned in Section 6.6.1, to address numerical unbalances, all individual training features were normalized with respect to their min-max values, which generally are affected by outliers. Normalization of test data with respect to min-max values of noisy training data, eventually results in degraded test performances. Poor performances of feature-level fusion can also be considered due to presence of outliers in the training data.

Feature-level fusion of power spectral features and delay vector variances of 7 EEG channels did not improve the detection accuracy of microsleep events (Golz et al. 2007). Similarly, Peiris et al. (2011), contrary to feature-level fusion of nonlinear and power spectral features, achieved the best lapse detection performance with power spectral features.

As mentioned in Section 6.6.2, the final decision in decision-level data fusion was based on mean of posterior probabilities of contributing classifiers, where each classifier was fed with individual feature sets. Therefore, irrespective of their performances, each classifier of the corresponding feature set got equal weights in making the final decision. Poor performances of decision-level fusion can be considered due to uniform weights assigned to each of the contributing classifier.

As shown in Figure 6.3 and described in Section 6.6, the single classifier in feature-level fusion was fed with whole concatenated training data, which comprised features selected from all 21 types of features. Similarly, each of the multiple classifiers in decision-level fusion was fed with each of the concatenated training feature set. In either technique, one classifier was required to handle the data. Contrary to our expectations, the performance metrics achieved

with data fusion were no better than single feature set. This indicates that there is essentially no orthogonal information on microsleeps to that contained in joint entropy.

8.4 ENSEMBLE TECHNIQUES

Compared to a single classifiers, performances of all ensemble techniques on microsleep states and onsets were higher as show in Figure 7.6 and Figure 7.7, respectively. Compared to a single classifier and data fusion, the higher performances achieved with ensemble classifications confirm that multiple classifiers using different portions of the same data (features) can better handle variability in the data (Moacir 2011, Valdovinos and Sanchez 2009). In addition, these improvements in the performances validate our approach of treating the data of each individual subject as a cluster.

Overlapping clusters of training subjects resulted in the overall highest performances on microsleep states. In overlapping clusters, some portions of data are fed to multiple classifiers (as mentioned in Section 6.7.4), where each classifier, instead of concatenated long data, learns from different permutations of the data. In addition, data distributions of overlapping clusters contain some degree of similarity. As a result, distribution of unseen data are more likely to match with multiple distributions of the training data. Compared to two proposed ensemble techniques (i.e., RMV and DMV), performances of overlapping clusters were slightly higher.

8.5 DETECTION AND PREDICTION OF MICROSLEEPS

The overall best detection and prediction (τ = 1 s) performances of three different approaches (i.e., single feature set, fusion, and ensemble classification) are summarized in Table 7.4 and Table 7.5, respectively on microsleep states and onsets. Joint entropy was the best single feature set, while feature-level fusion and overlapping clusters were the best fusion technique and ensemble classification, techniques, respectively.

At τ = 1 s, the overall best performances in single feature sets, fusion techniques, and ensemble classification techniques were achieved with joint entropy features (phi = 0.44, AUC_{PR} = 0.46, AUC_{ROC} = 0.94), feature-level fusion (phi = 0.45, AUC_{PR} = 0.44, AUC_{ROC} = 0.94), and overlapping clusters (phi = 0.47, AUC_{PR} = 0.47, AUC_{ROC} = 0.94), respectively, on microsleep states. Similarly, on microsleep onsets, the overall best prediction (τ = 1 s) performances of single joint entropy features, feature-level fusion, and overlapping clusters were (phi = 0.09, AUC_{PR} = 0.05, AUC_{ROC} = 0.89), (phi = 0.09, AUC_{PR} = 0.05, AUC_{ROC} = 0.89), and (phi = 0.10, AUC_{PR} = 0.07, AUC_{ROC} = 0.91), respectively.

It is evident that detection and prediction performances (on both microsleep states and onsets) of all approaches in terms of AUC_{ROC} were reasonably good but were low in terms of AUC_{PR} and phi. As shown in Table 4.1, class-imbalance ratio between microsleep states and responsive states, across the test subjects, was 1:813.40–1:2.26, which between microsleep onsets and responsive states was even higher (i.e., 1:4067–1:157). Relatively high values of

AUC_{ROC} were achieved due to its insensitivy to class-imbalance ratio (Fawcett 2006). Whereas both phi and AUC_{PR} being sensitive to class-imbalance data (Saito and Rehmsmeier 2015) were very low.

An AUC_{ROC} of 0.5 represents a random classifier (Fawcett 2006, Saito and Rehmsmeier 2015). Whereas a value of AUC_{PR} representing a random classifier is not fixed and depends on the class imbalance ratio. An AUC_{PR} of $\frac{P}{(P+N)}$ has been suggested to be used to represent a random classifier, where P is the number of positive class instances (i.e., microsleeps) and N is the number of negative class instances (i.e., responsive states) (Saito and Rehmsmeier 2015). Detection performance of all non-causal and normalized single features, feature-level fusion, and all ensemble techniques on both microsleep states and onsets, in terms of AUC_{PR} and AUC_{ROC}, were substantially higher than a random classifier.

Notwithstanding high AUC_{ROC}, the performance in terms of phi was very low across all of the methods due to highly imbalanced property of the dataset, which in addition varied substantially among different sessions and subjects.

8.6 COMPARISON WITH PREVIOUS STUDIES

Shoorangiz et al. (2016, 2017) used the same data to predict microsleep states at τ = 0.25 s and with a temporal resolution of 0.25 s. Using a 5-s window, they extracted 192 log-power spectral features from various frequency bands of EEG. A mutual information based greedy-forward-feature-selection algorithm was used to select informative features. With an LDA classifier, they achieved best performance metrics (phi, AUC_{PR}, AUC_{ROC}) of (0.33, 0.38, 0.90). Using the same features with Bayesian multi-subject factor analysis and an LDA classifier, a slightly higher phi of 0.34 was achieved. Except for covariance, all of non-normalized and non-causal single features across both feature-selection techniques and classifiers resulted in substantially higher performances compared to those of Shoorangiz et al. (2016, 2017).

Shoorangiz (2018) extracted various features from 2, 5, and 10 s window of EEG. He proposed four Bayesian models to reduce dimensionality of the feature sets and investigated four linear classifiers (i.e., LDA, LSVM, tree augmented naïve Bayes (TAN), variational Bayesian logistic regressionl (VBLR)) to discriminate microsleeps. The best detection performance metrics (phi = 0.47, AUC_{PR} = 0.49, AUC_{ROC} = 0.95) on microsleep states were achieved with an LDA classifier and meta-features of wavelet log mean squared features (WLMSF). In this research, slightly higher performance metrics (phi = 0.48, AUC_{PR} = 0.51, AUC_{ROC} = 0.95) were achieved with single joint entropy features, classifier-dependent feature-selection, and an LDA classifier. As shown in Table 7.4 and Figure 7.6, feature feature-level fusion and all of the proposed ensemble techniques resulted in higher performance metrics compared to those of Shoorangiz (2018) at detecting microsleep states. The best detection performance metrics (phi = 0.08, AUC_{PR} = 0.09, AUC_{ROC} = 0.91) on microsleep onsets were achieved with power spectral features and an LSVM classifier. As shown in Table 7.4, the overall best performances achieved in this research were slightly higher than those of Shoorangiz (2018).

The revised gold standard (Shoorangiz 2018) was used in this study and, consequently, the results presented in this thesis are not directly comparable to earlier microsleep studies (Ayyagari 2017, Davidson et al. 2007, Peiris 2008). However, the following is an attempt to provide an unbiased comparison between this research and the literature. It should be noted that the current project aimed to identify behavioural microsleeps, whereas lapses of responsiveness were the main focus of the previous works (Ayyagari 2017, Ayyagari et al. 2015, Davidson et al. 2007, Peiris 2008, Peiris et al. 2011). As mentioned in Section 2.3, behavioural microsleeps are episodes of sleep-related suspension of performance, whereas lapses of responsiveness include all short episodes of failure to respond in goal-directed tasks. As described in Section 4.3, previous works defined lapses as occurrence of tracking flat-spots and/or lapse video-ratings. Hence, a lapse could have been due to a tracking flat-spot, while the subject was awake. However, Shoorangiz (2018) defined microsleeps as non-tracking episodes accompanied with video ratings of deep-drowsy or lapse.

Davidson et al. (2007) extracted log-power spectral features from EEG and utilized an LSTM recurrent neural network as classifier to detect lapses with a temporal resolution of 1 s. They achieved the highest performance metrics (phi, AUC_{PR}, AUC_{ROC}) of (0.38, 0.41, 0.84). In a similar study, Ayyagari (2017) used the same features with a stack generalization of leaky echo-state neural networks to detect lapses. His highest performance metrics (phi, AUC_{PR}, AUC_{ROC}) were (0.44, 0.45, 0.88). Both of these studies used the whole dataset without data pruning. Evidently, the performances of all of the non-normalized and non-causal single features, fusion techniques, and the proposed ensemble techniques were higher than those of Davidson et al. and Ayyagari. However, a true comparison is not possible as they detected lapses, whereas detection and prediction of microsleeps were performed in this thesis. In addition, as mentioned in Section 6.8, phi and AUC_{PR} are sensitive to the class distribution (i.e., class-imbalance ratio) of data. The imbalance ratio of the refined and original gold standards have been reported as (0.001–0.44) and (0.01–0.36), respectively (Shoorangiz 2018). Despite similar means, the lower extremes of both imbalance ratios were substantially different. Consequently, comparison between performances in terms of phi and AUC_{PR} is strictly not valid. Nevertheless, the highest AUC_{ROC} of this study for microsleep state detection with a temporal resolution of 0.25 s was 0.96 (cf. 0.88 from Ayyagari (2017) and 0.84 from Davidson et al. (2007) for lapse detection).

Peiris et al. (2011) and Ayyagari et al. (2015) also performed lapse detection with a temporal resolution of 1 s. They, however, pruned the EEG-data by discarding the noisy epochs and performed the lapse detection task on the pruned EEG-epochs only. Peiris et al. (2011) used a stack generalization of LDA classifiers and achieved performances metrics (phi, AUC_{PR}, AUC_{ROC}) of (0.39, 0.43, 0.84). Ayyagari et al. (2015) used a stack generalization of leaky echo-state neural networks and achieved performances metrics (phi, AUC_{PR}, AUC_{ROC}) of (0.51, 0.47, 0.91). To compare the effect of pruning on the class-imbalance ratio, the epoch rejection process of these studies was replicated. After rejection of noisy epochs, the average imbalance ratio of the remaining data across all subjects was (0.01–0.38). As mentioned earlier, comparison of phi and AUC_{PR} between data sets of different imbalance ratios can be misleading.

Notwithstanding, compared to the unpruned dataset, these studies were able to achieve higher AUC_{ROC}. In this thesis, the highest microsleep detection AUC_{ROC} was 0.96, which is higher than all previous works.

Peiris et al. (2011) investigated detection of lapse events with a temporal resolution of 1 s. Lapse events were marked as true positive (correctly detected) if any point within an event was successfully identified as the lapse state. Additionally, lapse events without recognizing any points as lapse, were called missed events. To identify lapses, they used a stack generalization of multiple LDA classifiers and log-power spectral features. They achieved an average sensitivity and specificity of 0.74 and 0.59, respectively. As shown in Table 7.4, we achieved microsleep onset detection sensitivity of 0.73–0.76 and specificity of 0.89–0.90.

Golz et al. (2007) fused spectral and delay vector variances of 7 EEG, 2 EOG and 3 eye-tracking signals per eye (pupil size, x and y gaze coordinates) and achieved an accuracy of 0.91 with RBF-SVM on classification of microsleep and alert events. This promising but erroneous accuracy was achieved by balancing the test data and cross-validation was performed on concatenated data from all of the subjects. In doing so, independence of the test and training data, and hence generalization accuracy of the system, were lost. Similarly, with spectral features and a claimed accuracy of 0.88 on the prediction of microsleep events (Golz et al. 2016).

8.7 OVERALL PERFORMANCE OF MICROSLEEP PREDICTION SYSTEM

In terms of phi, AUC_{PR}, and AUC_{ROC}, compared to earlier work on this data, we were able to achieve the overall best microsleep (state and onset) detection and prediction performance. As mentioned in Table 7.4 and Table 7.5, the overall microsleep system with all approaches (i.e., single feature set with single classifier, multiple features, ensemble classifications) resulted in reasonably high sensitivity, specificity, GM, and AUC_{ROC}. Notwithstanding the best performance in terms of Fm, phi and AUC_{PR}, our system is still far away from real-life implementation.

As mentioned in Section 6.5, to account for class imbalance ratio in the training data sets, an algorithm-based approach referred to as cost-sensitive learning (Lin and Chen 2013) was employed. Otherwise, the accuracy in predicting majority class will be high and low for minority class. Considering both classes (i.e., microsleep and responsive) equally important (or equal misclassification cost), equal prior probabilities (0.5) were assigned to both classes in the decision. This can be considered to have resulted in high sensitivity, low precision, and minimal effect on specificity.

To account for class-imbalance, two data-based approaches of oversampling – synthetic minority oversampling technique (SMOTE) and adaptive synthetic sampling (ADASYN) – have been investigated to predict microsleep states. The difference between threshold-free metrics of performance (i.e., AUC_{PR}, AUC_{ROC}) achieved with and without both oversampling methods has been reported to be minimal (Shoorangiz et al. 2016).

Ineffectiveness of data-based approaches and bias of algorithm-based approach towards sensitivity indicate high level of data complexity (overlapping/non-separability of majority and minority class) and variability (i.e., inter-session and inter-subject). Addressing both data complexity and variability followed by class-imbalance ratio can result in improved metrics of performance.

8.8 SUMMARY

This chapter commented on various types of features, feature demeaning, and classification techniques. Non-normalized and non-causal features were found to have the most microsleep-related information. In addition to being computationally inexpensive, time-domain univariate (i.e., variance) features were found to have comparable results to that of corresponding univariate spectral features. Except for mutual information, information-theoretic features gave the highest performances within their respective groups. Performances of multivariate cross-spectral power and coherence were higher than that of their corresponding bivariate counterparts.

With unbiased (i.e., classifier-independent) feature-selection technique, both classifiers resulted in similar performance. Comparatively, higher performance metrics were achieved with classifier-dependent feature-selection and an LDA classifier.

The performances of data fusion techniques were no better than the single best performing feature set. Compared to a single classifier, all proposed ensemble techniques resulted in the overall higher detection and prediction performance on both microsleep states and onsets.

Chapter 9

CONCLUSION AND FEATURE RESEARCH

9.1 SUMMARY

The motivations for this project were (1) to explore and various inter-channel relationships of EEG as features to detect/predict microsleeps and (2) to investigate data fusion (including ensemble learning) techniques to improve detection/prediction accuracies of microsleeps. Accurate and noninvasive prediction of imminent microsleeps can potentially avoid sleep-related fatal accidents by continuously monitoring an individual's responsiveness, especially in occupations requiring extended visuomotor performance (e.g., transportation).

Data of 15 non-sleep-deprived healthy individuals, performing a 1-D continuous tracking task, were collected in a previous study (Peiris 2008). The data comprised EEG, tracking performance, and eye video. Participants took part in two 1-h long sessions. However, only 8 out of 15 subjects had a one or more microsleeps during the two 1-hour sessions. In this study, the data of those 8 subjects were used to investigate the efficacy of various inter-channel relationships at detecting/predicting microsleeps and to evaluate the performances of overall microsleep detection/prediction system. Due to shortcomings of the original gold standard, a refined gold standard (Shoorangiz 2018) was used. Irrespective of the expert's video ratings, at least 5 s of coherent tracking was marked (labeled) as responsive. The conjunction of a non-tracking episode and a video rating of deep-drowsy or lapse was labeled as microsleep. Due to a lack of information to accurately identify the state of responsiveness, the remainder of the gold-standard was labeled as uncertain. In this study, detection/prediction of microsleeps was done with a temporal resolution of 0.25 s (i.e., 4 Hz) and uncertain labels were discarded.

Seven bivariate, symmetric, and data-driven inter-channel feature sets were extracted from each pair of EEG channels. Three of the 7 were non-normalized (i.e., covariance, wavelet cross-spectral power, joint entropy), whereas the remaining 4 were normalized (i.e., correlation, coherence, mutual information, PSI). Eleven multivariate feature sets were extracted from coefficients of an MVAR model fitted to each epoch of EEG. Five of these were non-causal and symmetric (i.e., cross-spectral power, coherence, iCOH, pCOH, PSI), whereas the remaining 6 were causal (i.e., PDC, GPDC, DTF, nDTF, ffDTF, dDTF). In addition, three univariate feature sets (i.e., variance, spectral power, and entropy) were extracted from individual channels of the

EEG. All features were extracted from an EEG window length of 5 s and discriminatory features were used as input of the classifiers. To account for inter-session and inter-subject variability in the data, features of each session were demeaned with respect to the mean of their first 2-min.

Two supervised feature-selection techniques were used to select discriminatory and relevant features from each feature set. The first technique was classifier-dependent feature-selection, which comprised Pearson's correlation coefficient-based filter followed by Fisher's score-based ranking and an LDA-based wrapper. The second technique was classifier-independent, which was a combination of relevance-based ranking (i.e., mutual information between a feature and class labels) and relevance-based SFS method.

Two linear classifiers, i.e., LDA and LSVM, were used to validate the efficacy of a feature set at detecting/predicting microsleeps. Priors for both classifiers were incorporated to address the class imbalance in the training data sets.

To improve detection/prediction accuracies of microsleeps, all 21 feature sets were fused at feature-level and decision-level. Two-stage feature-selection was performed in feature-level fusion and homogeneous (same type) classifiers were used in decision-level fusion. In addition, 3 ensemble classification techniques were proposed, which together with soft majority voting were used on joint entropy features. In three of these techniques, data from each subject were treated as separate cluster, whereas in the 4-th technique individual clusters were combined, using divergence, to form overlapping clusters.

Like recent work of Shoorangiz (2018), this study focused on detection/prediction of microsleeps states and microsleep onsets (see Figure 4.4 and Figure 4.5, respectively). The microsleep state detection/prediction system aims to predict an imminent microsleep τ s prior to its occurrence and continuously identify the state of responsiveness. If the prediction of one microsleep state is missed, it will try to predict/detect the following microsleep state. On the other hand, microsleep onset detection/prediction system continuously attempts to predict the onset of an imminent microsleep. However, if an onset prediction is missed, the entire event of microsleep is missed.

9.2 KEY FINDINGS

Non-normalized and non-causal (both bivariate and multivariate) inter-channel features performed better than their corresponding normalized features, indicating that microsleep-related information is sensitive to both amplitude and frequency. The performances of univariate temporal (i.e., variance) and spectral features were similar. Correlation feature set performed better than its bivariate and normalized counterparts (i.e., coherence, mutual information, PSI). Bivariate cross-spectral power features were better than covariance features. Multivariate features were often better than bivariate counterparts. Both entropy and joint entropy were the best single feature sets in their respective univariate and bivariate groups.

Baseline correction (demeaning) of non-normalized features substantially improved the

performances on microsleep states and onsets. These improvement can be considered as baseline correction addressed inter-session and inter-subject variability in the features. Overall performances of both feature-level and decision-level fusion were lower than single best performing feature set (i.e., joint entropy). Compared to single classifiers, overall performances of all of the proposed ensemble techniques were higher.

The overall highest detection performance metrics (phi = 0.48, AUC_{PR} = 0.51, AUC_{ROC} = 0.95) and (phi = 0.11, AUC_{PR} =0.09, AUC_{ROC} = 0.91) on microsleep states and onsets, respectively using single feature sets were achieved with joint entropy features indicating its capability of extracting microsleep-related information.

Feature-level fusion, compared to decision-level, resulted in the overall higher detection performance metrics of (phi = 0.49, AUC_{PR} = 0.49, AUC_{ROC} = 0.96) and (phi = 0.10, AUC_{PR} = 0.09, AUC_{ROC} = 0.92) on microsleep states and onsets, respectively.

Performances of all of the proposed ensemble techniques were higher than that of single classifiers, indicating their capability in handling data variability. Overlapping clusters resulted in the overall highest detection performance metrics of (phi = 0.50, AUC_{PR} = 0.52, AUC_{ROC} = 0.96) and (phi = 0.12, AUC_{PR} =0.10, AUC_{ROC} = 0.93) on microsleep states and onsets, respectively.

Although with single joint entropy features and proposed ensemble techniques, we were able to achieve higher results, the overall performances are still insufficient for real-life applications. Our results in terms of phi, AUC_{PR}, and AUC_{ROC} indicate moderately high sensitivity but low precision (i.e., too many false positives) of microsleep detection/prediction.

9.3 REVIEW OF HYPOTHESES

•**Hypothesis 1** — Inter-channel relationships can improve the detection/prediction accuracy of microsleep states and onsets over spectral features extracted from independent channels.

The accuracy of both detection and prediction of microsleep states from cross-spectral power (both bivariate and multivariate) and joint entropy was superior to spectral features extracted from independent channels, thus supporting this hypothesis. Except for bivariate cross-spectral power, the hypothesis was also true for microsleep onsets as the performances of both multivariate cross-spectral power and joint entropy were higher than that of univariate spectral power features.

•**Hypothesis 2** — Fusion of multiple feature sets can improve detection/prediction accuracy of microsleep states and onsets over single feature set.

Compared to the single best performing feature set (joint entropy), the performances of both feature-level and decision-level fusion on both microsleep states and onsets were similar, which do not support this hypothesis and suggest not to use data fusion in detection/prediction of microsleeps.

Mean detection/prediction performances of all the proposed ensemble classification techniques irrespective of the feature-selection technique and classifier type, compared to the best performing feature set and with a single classifier, were higher on both microsleep states and onsets. Therefore, considering ensemble learning as decision-level fusion supports this hypothesis.

9.4 CRITIQUE

A small number of participants limits the generalization of our findings and, is therefore, a limitation of the this research.

The velocity of the 1D target was zero whenever it changed direction (i.e., at positive or negative peak), as shown in upper panel of Figure 4.3. Shoorangiz (2018) reported that the velocity of the pseudo-random target in each 128-s cycle fell to zero 34 times. These episodes were marked uncertain and were subsequently discarded. As a result, all of the microsleep might have not been identified. Such limitation, however, can be prevented in future studies by using a 2-D continuous tracking task (Poudel et al. 2014).

On a positive note, single joint entropy features and all of the proposed ensemble techniques, resulted in overall higher performances at detecting/predicting microsleep states and onset. These improvements indicate that joint entropy features have captured microsleep-related information and all of the proposed ensemble techniques have addressed variability in the data.

Our results in terms of phi, AUC_{PR}, and AUC_{ROC} indicate that both microsleep states and onsets can be predicted prior to their occurrence with a moderately high sensitivity, especially with shorter prediction times but that low phi and AUC_{PR}, indicating low precision (i.e., too many false positives) make our system impractical for real-life applications.

9.5 FUTURE WORK RECOMMENDATIONS

To assess the generalization of microsleep detection/prediction method on unseen data, the number of participants in future studies needs to be larger. In addition, from a machine learning perspective, a large number of participants could provide a better estimate of bias-variance trade-off. For a fixed-time on a task, a large number of participants would result in a large number of training examples and, consequently, a better learning/training of a classifier. This, in the loss function, will reduce the variance term, which is superior to increase in the bias (Le Borgne 2005). Decrease in variance leads to a more simple but generalized (i.e., less sensitive to a specific training set) classifier.

The subjectivity involved in video ratings is likely to have introduced some errors in the gold standard (i.e., incorrect labels). Supervised machine learning algorithms learn and explore the relationship between features and labels. In addition to variable imbalanced data across subjects, an imperfect gold standard can therefore have a substantial impact on overall detection/prediction

performances. In future studies, video recordings and subsequently the gold standard need to be algorithmically (automatically) developed and any uncertainty therein needs to be discarded at the time of training.

Our results showed that non-normalized features performed better than normalized features. On the other hand, normalized features did not show microsleep-related information, but being scale-invariant can handle large inter-session and inter-subject EEG variability (Rocca et al. 2014). Contrary to our expectations, both feature-level and decision-level fusions could not improve microsleep detection/prediction accuracies. For a single-trial EEG brain-computer interfaces, Heger et al. (2014) using effective connectivity features as filter to spectral power features, achieved significant performance improvements. Normalized inter-channel relationships as FIR filter to non-normalized features needs to be investigated. Filtering of non-normalized features with respect to normalized ones can potentially address inter-subject variability in data and preserve microsleep-related information contained in EEG amplitudes. Addressing variabilities in EEG while maintaining signal shapes could well improve microsleep detection/prediction accuracies. In addition, portion and type (e.g., microsleep, responsive) of training data and the type of normalized inter-channel relationship (i.e., bivariate, multivariate, causal or non-causal) need to be investigated.

In the current study, all inter-channel relationships were extracted from fixed length EEG windows and the corresponding features did not have microsleep-related temporal information. Therefore, exploiting methods such as sub-segmentation of bivariate and incorporation of Kalman filter in multivariate inter-channel relationships to maintain temporal information might provide higher accuracies of microsleep detection/prediction.

Compared to a single classifier, higher performances on microsleep detection/prediction were achieved with all of the proposed ensemble techniques. Our results suggest further exploration of different ways to form overlapping clusters and fuse decisions of individual classifiers.

In many applications nonlinear classifiers perform better than linear counterparts. However, for a large and high dimensional data, nonlinear classifiers are computationally expensive (see Section 6.5). Data length can be handled by avoiding concatenation and treating data of each subject as a cluster. Subjects with large data can further be clustered with respect to sessions and potential over-fitting can be avoided by properly regularizing nonlinear classifiers. Finally, fusion (i.e., soft majority voting) of properly regularized nonlinear base classifiers, each trained to specific data portion (subject), may improve microsleep detection/prediction accuracies. However, the number of clusters and whether each cluster contains minority class, need to be investigated.

REFERENCES

Albalawi, H. and Li, X. (2018), 'Single-channel real-time drowsiness detection based on electroencephalography', In *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 99–101.

Alhaddad, M.J. (2012), 'Common average reference (CAR) improves P300 speller', *International Journal of Engineering and Technology*, Vol. 2, No. 3, p. 21.

Ali, S., Tirumala, S.S. and Sarrafzadeh, A. (2015), 'Ensemble learning methods for decision making: Status and future prospects', In *International Conference on Machine Learning and Cybernetics*, Vol. 1, pp. 211–216.

Astolfi, L., Cincotti, F., Mattia, D., Marciani, M.G., Baccala, L.A., Fallani, F.D.V., Salinari, S., Ursino, M., Zavaglia, M. and Babiloni, F. (2006), 'Assessing cortical functional connectivity by partial directed coherence: Simulations and application to real data', *IEEE Transactions on Biomedical Engineering*, Vol. 53, No. 9, pp. 1802–1812.

Awais, M., Badruddin, N. and Drieberg, M. (2017), 'EEG brain connectivity analysis to detect driver drowsiness using coherence', In *International Conference on Frontiers of Information Technology*, pp. 110–114.

Ayyagari, S. (2017), *Reservoir Computing approaches to EEG-based Detection of Microsleeps*, PhD thesis, University of Canterbury.

Ayyagari, S., Jones, R.D. and Weddell, S.J. (2015), 'Optimized echo state networks with leaky integrator neurons for EEG-based microsleep detection', In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3775–3778.

Baccalá, L.A. and Sameshima, K. (2001), 'Partial directed coherence: a new concept in neural structure determination', *Biological Cybernetics*, Vol. 84, No. 6, pp. 463–474.

Baccalá, L.A., Sameshima, K. and Takahashi, D.Y. (2007), 'Generalized partial directed coherence', In *15th International Conference on Digital Signal Processing*, pp. 163–166.

Bajaj, V. and Pachori, R.B. (2013), 'Automatic classification of sleep stages based on the time-frequency image of EEG signals', *Computer Methods and Programs in Biomedicine*, Vol. 112, No. 3, pp. 320–328.

Barnett, L. and Seth, A.K. (2016), 'Detectability of Granger causality for subsampled continuous-time neurophysiological processes', *Journal of Neuroscience Methods*, Vol. 275, pp. 93–121.

Bastos, A.M. and Schoffelen, J.M. (2016), 'A tutorial review of functional connectivity analysis methods and their interpretational pitfalls', *Frontiers in Systems Neuroscience*, Vol. 9, No. 175, p. 23 pp.

Blinowska, K.J. (2011), 'Review of the methods of determination of directed connectivity from multichannel data', *Medical and Biological Engineering*, Vol. 49, No. 5, pp. 521–529.

Blinowska, K.J., Kuś, R. and Kamiński, M. (2004), 'Granger causality and information flow in multivariate processes', *Physical Review E*, Vol. 70, No. 5, p. 050902(4pp).

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D. and Babiloni, F. (2014), 'Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness', *Neuroscience and Biobehavioral Reviews*, Vol. 44, pp. 58–75.

Boughorbel, S., Jarray, F. and El-Anbari, M. (2017), 'Optimal classifier for imbalanced data using Matthews correlation coefficient metric', *PLOS ONE*, Vol. 12, No. 6, pp. 1–17.

Brittain, J., Halliday, D.M., Conway, B.A. and Nielsen, J.B. (2007), 'Single-trial multi-wavelet coherence in application to neurophysiological time series', *IEEE Transactions on Biomedical Engineering*, Vol. 54, No. 5, pp. 854–862.

Broersen, P.M.T. (2009), 'Vector autoregressive order selection in practice', *IEEE Transactions on Instrumentation and Measurement*, Vol. 58, No. 8, pp. 2565–2573.

Bruns, A. (2004), 'Fourier-, Hilbert- and wavelet-based signal analysis: are they really different approaches?', *Journal of Neuroscience Methods*, Vol. 137, No. 2, pp. 321–332.

Buckley, R., J., Helton, S., W., Innes, C.R.H., Dalrymple-Alford, J.C. and Jones, R.D. (2016), 'Attention lapses and behavioural microsleeps during tracking, psychomotor vigilance, and dual tasks', *Consciousness and Cognition*, Vol. 45, pp. 174–183.

Burgess, A.P. (2012), 'Towards a unified understanding of event-related changes in the EEG: The firefly model of synchronization through cross-frequency phase modulation', *PLOS ONE*, Vol. 7, No. 9, pp. 1–21.

Castanedo, F. (2013), 'A review of data fusion techniques', *The Scientific World Journal*, Vol. 2013, p. 19 pp.

Chapelle, O. (2007), 'Training a support vector machine in the primal', *Neural Computation*, Vol. 19, No. 5, pp. 1155–1178.

Chapotot, F. and Becq, G. (2010), 'Automated sleep–wake staging combining robust feature extraction, artificial neural network classification, and flexible decision rules', *International Journal of Adaptive Control and Signal Processing*, Vol. 24, No. 5, pp. 409–423.

Chella, F., Pizzella, V., Zappasodi, F. and Marzetti, L. (2016), 'Impact of the reference choice on scalp EEG connectivity estimation', *Journal of Neural Engineering*, Vol. 13, No. 3, p. 036016 (21pp).

Chen, L., Zhao, Y., Zhang, J. and Zou, J. (2015), 'Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning', *Expert Systems with Applications*, Vol. 42, pp. 7344–7355.

Chen, S., Luo, Z. and Gan, H. (2018), 'An entropy fusion method for feature extraction of EEG', *Neural Computing and Applications*, Vol. 29, No. 10, pp. 857–863.

Clercq, W.D., Vergult, A., Vanrumste, B., Paesschen, W.V. and Huffel, S.V. (2006), 'Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram', *IEEE Transactions on Biomedical Engineering*, Vol. 53, No. 12, pp. 2583–2587.

Correa, A.G., Orosco, L. and Laciar, E. (2014), 'Automatic detection of drowsiness in EEG records based on multimodal analysis', *Medical Engineering & Physics*, Vol. 36, No. 2, pp. 244–249.

Şen, B., Peker, M., Çavuşoğlu, A. and Çelebi, F.V. (2014), 'A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms', *Journal of Medical Systems*, Vol. 38, No. 3, p. 21 pp.

Dasarathy, B.V. (1997), 'Sensor fusion potential exploitation–Innovative architectures and illustrative applications', *Proceedings of the IEEE*, Vol. 85, No. 1, pp. 24–38.

Davidson, P.R., Jones, R.D. and Malik, P.T. (2007), 'EEG-based lapse detection with high temporal resolution', *IEEE Transactions on Biomedical Engineering*, Vol. 54, No. 5, pp. 832–839.

Dietterich, T.G. (2000), 'Ensemble methods in machine learning', In *Multiple Classifier Systems*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–15.

Dimitrakopoulos, G.N., Kakkos, I., Dai, Z., Wang, H., Sgarbas, K., Thakor, N., Bezerianos, A. and Sun, Y. (2018), 'Functional connectivity analysis of mental fatigue reveals different network topological alterations between driving and vigilance tasks', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 26, No. 4, pp. 740–749.

Dissanayaka, C., Ben-Simon, E., Gruberger, M., Maron-Katz, A., Sharon, H., Hendler, T. and Cvetkovic, D. (2015), 'Comparison between human awake, meditation and drowsiness EEG activities based on directed transfer function and MVDR coherence methods', *Medical and Biological Engineering*, Vol. 53, No. 7, pp. 599–607.

Duchi, J. (2007), 'Derivations for linear algebra and optimization', *Berkeley, California*, Vol. 3, p. 13.

Duda, R.O., Hart, P.E. and Stork, D.G. (2000), *Pattern Classification*, Wiley-Interscience, 2nd ed.

Duffy, F.H., Iyer, V.G. and Surwillo, W.W. (1989), 'Brain electrical activity: An introduction to EEG recording', In *Clinical Electroencephalography and Topographic Brain Mapping*, pp. 1–10, Springer.

Fawcett, T. (2006), 'An introduction to ROC analysis', *Pattern Recognition letters*, Vol. 27, No. 8, pp. 861–874.

Fell, J. and Nikolai (2011), 'The role of phase synchronization in memory processes', *Nature Neuroscience*, Vol. 12, No. 2, pp. 105–118.

Finkbeiner, K.M., Wilson, K.M., Russell, P.N. and Helton, W.S. (2015), 'The effects of warning cues and attention-capturing on the sustained attention to response task', *Experimental Brain Research*, Vol. 233, No. 4, pp. 10161–1068.

Florin, E., Gross, J., Pfeifer, J., Fink, G.R. and Timmermann, L. (2010), 'The effect of filtering on Granger causality based multivariate causality measures', *NeuroImage*, Vol. 50, No. 2, pp. 577–588.

Friston, K.J. (2011), 'Functional and effective connectivity: A review', *Brain connectivity*, Vol. 1, No. 1, pp. 13–36.

Friston, K.J., Moran, R. and Seth, A.K. (2013), 'Analysing connectivity with Granger causality and dynamic causal modelling', *Current Opinion in Neurobiology*, Vol. 23, No. 2, pp. 172 –178.

Friston, K.J., Bastos, A.M., Oswal, A., van Wijk, B., Richter, C. and Litvak, V. (2014), 'Granger causality revisited', *NeuroImage*, Vol. 101, pp. 796–808.

Garofalo, M., Nieus, T., Massobrio, P. and Martinoia, S. (2009), 'Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks', *PLOS ONE*, Vol. 4, No. 8, pp. 1–14.

Garrett, D., Peterson, D.A., Anderson, C.W. and Thaut, M.H. (2003), 'Comparison of linear, nonlinear, and feature selection methods for EEG signal classification', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 11, No. 2, pp. 141–144.

Gibbons, J.D. and Chakraborti, S. (2003), *Nonparametric Statistical Inference*, Marcel Dekker Inc., 4th ed.

Golz, M., Sommer, D., Chen, M., Mandic, D. and Trutschel, U. (2007), 'Feature fusion for the detection of microsleep events', *Journal of VLSI Signal Processing*, Vol. 49, pp. 329–342.

Golz, M., Sommer, D. and Krajewski, J. (2016), 'Prediction of immediately occurring microsleep events from brain electrical signals', *Current Directions in Biomedical Engineering*, Vol. 2, No. 1, pp. 149–153.

Gong, J. and Kim, H. (2017), 'RHSBoost: Improving classification performance in imbalance data', *Computational Statistics & Data Analysis*, Vol. 111, pp. 1–13.

Grinsted, A., Moore, J.C. and Jevrejeva, S. (2004), 'Application of the cross wavelet transform and wavelet coherence to geophysical time series', *Nonlinear Processes in Geophysics*, Vol. 11, pp. 561–566.

Gu, Q., Li, Z. and Han, J. (2011), 'Generalized fisher score for feature selection', In *27th Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 266–273.

Harrison, Y. and Horne, J.A. (1996), 'Occurrence of microsleeps during daytime sleep onset in normal subjects', *Electroencephalography and Clinical Neurophysiology*, Vol. 98, No. 5, pp. 411–416.

Harvy, J., Sigalas, E., Thakor, N., Bezerianos, A. and Li, J. (2018), 'Performance improvement of driving fatigue identification based on power spectra and connectivity using feature level and decision level fusion', In *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, .

Hastie, T., Tibshirani, R. and Friedman, J. (2008), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, New York, 2nd ed.

Heger, D., Terziyska, E. and Schultz, T. (2014), 'Connectivity based feature-level filtering for single-trial EEG BCIs', In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2064–2068.

Higgins, J.S., Michael, J., Austin, R., Åkerstedt, T., Van Dongen, H.P.A., Watson, N., Czeisler, C., Pack, A.I. and Rosekind, M.R. (2017), 'Asleep at the wheel–the road to addressing drowsy driving', *Sleep*, Vol. 40, No. 2, pp. (zsx001) 1–9 pp.

Hu, S., Lai, Y., Valdes-Sosa, P.A., Bringas-Vega, M.L. and Yao, D. (2018), 'How do reference montage and electrodes setup affect the measured scalp EEG potentials?', *Journal of Neural Engineering*, Vol. 15, No. 2, p. 026013 (13pp).

Huang, R.S., Jung, T.P., Delorme, A. and Makeig, S. (2008), 'Tonic and phasic electroencephalographic dynamics during continuous compensatory tracking', *NeuroImage*, Vol. 39, No. 4, pp. 1896–1909.

Innes, C.R.H., Poudel, G.R. and Jones, R.D. (2013), 'Efficient and regular patterns of nighttime sleep are related to increased vulnerability to microsleeps following a single night of sleep restriction', *Chronobiology International*, Vol. 30, No. 9, pp. 1187–1196.

Jones, R.D., Poudel, G.R., Innes, C.R.H., Davidson, P.R., Peiris, M.T.R., Malla, A.M., Signal, T.L., Carroll, G.J., Watts, R. and Bones, P.J. (2010), 'Lapses of responsiveness: Characteristics, detection, and underlying mechanisms', In *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 1788–1791.

Jonmohamadi, Y., Poudel, G.R., Innes, C.C.R.H. and Jones, R.D. (2016), 'Microsleeps are associated with stage-2 sleep spindles from hippocampal-temporal network', *International Journal of Neural Systems*, Vol. 26, No. 04, p. 1650015 (12 pp).

Kabbara, A., Khalil, M., El-Falou, W., Eid, H. and Hassan, M. (2016), 'Functional brain connectivity as a new feature for P300 speller', *PLOS ONE*, Vol. 11, No. 1, p. e0146282 (18 pp).

Kamiński, M. and Blinowska, K.J. (1991), 'A new method of the description of the information flow in brain structures', *Biological Cybernetics*, Vol. 65, pp. 203–210.

Kayser, J. and Tenke, C.E. (2015), 'Issues and considerations for using the scalp surface laplacian in EEG/ERP research: A tutorial review', *International Journal of Psychophysiology*, Vol. 97, No. 3, pp. 189–209.

Khan, S., Bandyopadhyay, S., Ganguly, A.R., Saigal, S., III, D.J.E., Protopopescu, V. and Ostrouchov, G. (2007), 'Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data', *Physical Review E*, Vol. 76, No. 2, p. 026209 (15 pp).

Klein, A., Sauer, T., Jedynak, A. and Skrandies, W. (2006), 'Conventional and wavelet coherence applied to sensory-evoked electrical brain activity', *IEEE Transactions on Biomedical Engineering*, Vol. 53, No. 2, pp. 266–272.

Kong, W., Zhou, Z., Jiang, B., Babiloni, F. and Borghini, G. (2017), 'Assessment of driving fatigue based on intra/inter-region phase synchronization', *Neurocomputing*, Vol. 219, pp. 474–482.

Korzeniewska, A., gorzata Mańczak, M., Kamiński, M., Blinowska, K.J. and Kasicki, S. (2003), 'Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method', *Journal of Neuroscience Methods*, Vol. 125, pp. 195–207.

Korzeniewska, A., Crainiceanu, C.M., Kuś, R., Franaszczuk, P.J. and Crone, N.E. (2008), 'Dynamics of event-related causality in brain electrical activity', *Human Brain Mapping*, Vol. 29, No. 10, pp. 1170–1192.

Krajewski, J., Batliner, A. and Wieland, R. (2008), 'Multiple classifier applied on predicting microsleep from speech', In *19th International Conference on Pattern Recognition*, pp. 1–4.

Krajewski, J., GolzG, M., Sommer, D. and Wieland, R. (2009), 'Genetic algorithm based feature selection applied on predicting microsleep from speech', In *4th European Conference of the International Federation for Medical and Biological Engineering*, Springer, pp. 184–187.

Kraskov, A., Stögbauer, H. and Grassberger, P. (2004), 'Estimating mutual information', *Physical Review E*, Vol. 69, No. 6, p. 066138 (16 pp).

Kreuz, T., Mormann, F., Andrzejak, R.G., Kraskov, A., Lehnertz, K. and Grassberger, P. (2007), 'Measuring synchronization in coupled model systems: A comparison of different approaches', *Physica D: Nonlinear Phenomena*, Vol. 225, No. 1, pp. 29 –42.

Kumar, Y., Dewal, M. and Anand, R. (2014), 'Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine', *Neurocomputing*, Vol. 133, pp. 271–279.

Kuncheva, L.I. and Whitaker, C.J. (2003), 'Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy', *Machine Learning*, Vol. 51, No. 2, pp. 181–207.

Kuś, R., Kamiński, M. and Blinowska, K.J. (2004), 'Determination of EEG activity propagation: pair-wise versus multichannel estimate', *IEEE Transactions on Biomedical Engineering*, Vol. 51, No. 9, pp. 1501–1510.

Kwak, N. and Choi, C.H. (2002), 'Input feature selection by mutual information based on Parzen window', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 12, pp. 1667–1671.

Lahat, D., Adali, T. and Jutten, C. (2015), 'Multimodal data fusion: An overview of methods, challenges, and prospects', *Proceedings of the IEEE*, Vol. 103, No. 9, pp. 1449–1477.

Lawrence, N.D. and Schölkopf, B. (2001), 'Estimating a kernel Fisher discriminant in the presence of label noise', In *18th International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 306–313.

Le Borgne, Y. (2005), 'Bias-variance trade-off characterization in a classification problem: What differences with regression?', *Machine Learning Group, University of Libre de Bruxelles, Belgium*.

Lee, Y.Y. and Hsieh, S. (2014), 'Classifying different emotional states by means of EEG-based functional connectivity patterns', *PLOS ONE*, Vol. 9, No. 4, pp. 1–13.

Léger, D., Bayon, V., Ohayon, M.M., Philip, P., Ement, P., Metlaine, A., Chennaoui, M. and Faraut, B. (2014), 'Insomnia and accidents: cross-sectional study (equinox) on sleep-related home, work and car accidents in 5293 subjects with insomnia from 10 countries', *Journal of Sleep Research*, Vol. 23, No. 2, pp. 143–152.

Lemm, S., Blankertz, B., Dickhaus, T. and Müller, K.R. (2011), 'Introduction to machine learning for brain imaging', *NeuroImage*, Vol. 56, No. 2, pp. 387–399.

Lin, C.T., Huang, K.C., Chuang, C.H., Ko, L.W. and Jung, T.P. (2013), 'Can arousing feedback rectify lapses in driving? Prediction from EEG power spectra', *Journal of Neural Engineering*, Vol. 10, p. 056024 (10pp).

Lin, F.C., Ko, L.W., Chuang, C.H., Su, T.P. and Lin, C.T. (2012), 'Generalized EEG-based drowsiness prediction system by using a self-organizing neural fuzzy system', *IEEE Transactions on Circuits and Systems–I*, Vol. 59, No. 9, pp. 2044–2055.

Lin, W.J. and Chen, J.J. (2013), 'Class-imbalanced classifiers for high-dimensional data', *Briefings in Bioinformatics*, Vol. 14, No. 1, pp. 13–26.

Liu, Y.T., Lin, Y.Y., Wu, S.L., Chuang, C.H. and Lin, C.T. (2016), 'Brain dynamics in predicting driving fatigue using a recurrent self-evolving fuzzy neural network', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 27, No. 2, pp. 347–360.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F. and Arnaldi, B. (2007), 'A review of classification algorithms for EEG-based brain–computer interfaces', *Journal of Neural Engineering*, Vol. 4, No. 2, pp. R1–R13.

López, V., Fernández, A., García, S., Palade, V. and Herrera, F. (2013), 'An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics', *Information Sciences*, Vol. 250, pp. 113–141.

Lütkepohl, H. (2007), *New Introduction to Multiple Time Series Analysis*, Springer Publishing Company, Incorporated.

Malla, A.M., Davidson, P.R., Bones, P.J., Green, R. and Jones, R.D. (2010), 'Automated video-based measurement of eye closure for detecting behavioral microsleep', In *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 6741–6744.

Mangai, U.G., Samanta, S., Das, S. and Chowdhury, P.R. (2010), 'A survey of decision fusion and feature fusion strategies for pattern classification', *IETE Technical Review*, Vol. 27, No. 4, pp. 293–307.

Maraun, D. and Kurths, J. (2004), 'Cross wavelet analysis: significance testing and pitfalls', *Nonlinear Processes in Geophysics*, Vol. 11, No. 4, pp. 505–514.

Marple, S.L. (1999), 'Computing the discrete-time "analytic" signal via FFT', *IEEE Transactions on Signal Processing*, Vol. 47, No. 9, pp. 2600–2603.

May, R., Dandy, G. and Maier, H. (2011), 'Review of input variable selection methods for artificial neural networks', In *Artificial Neural Networks - Methodological Advances and Biomedical Applications*, InTech.

McFarland, D.J., McCane, L.M., David, S.V. and Wolpaw, J.R. (1997), 'Spatial filter selection for EEG-based communication', *Electroencephalography and clinical Neurophysiology*, Vol. 103, No. 3, pp. 386–394.

Melia, U., Guaita, M., Vallverd/'u, M., Embid, C., Vilaseca, I., Salamero, M. and JoanSantamaria (2015), 'Mutual information measures applied to EEG signals for sleepiness characterization', *Medical Engineering and Physics*, Vol. 37, pp. 297–308.

Memar, P. and Faradji, F. (2018), 'A novel multi-class EEG-Based sleep stage classification system', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 26, No. 1, pp. 84–95.

Moacir, P.P.J. (2011), 'Combining classifiers: From the creation of ensembles to the decision fusion', In *24th IEEE SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials*, pp. 1–10.

Mullen, T.R., Kothe, C.A.E., Chi, Y.M., Ojeda, A., Kerth, T., Makeig, S., Jung, T.P. and Cauwenberghs, G. (2015), 'Real-time neuroimaging and cognitive monitoring using wearable dry EEG', *IEEE Transactions on Biomedical Engineering*, Vol. 62, No. 11, pp. 2553–2567.

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S. and Hallett, M. (2004), 'Identifying true brain interaction from EEG data using the imaginary part of coherency', *Clinical Neurophysiology*, Vol. 115, No. 10, pp. 2292–2307.

Oostenveld, R. and Praamstra, P. (2001), 'The five percent electrode system for high-resolution EEG and ERP measurements', *Clinical Neurophysiology*, Vol. 112, No. 4, pp. 713–719.

Pageaux, B., Marcora, S.M., Rozand, V. and Lepers, R. (2015), 'Mental fatigue induced by prolonged self-regulation does not exacerbate central fatigue during subsequent whole-body endurance exercise', *Frontiers in Human Neuroscience*, Vol. 9, No. 67, p. 12pp.

Parker, C. (2011), 'An analysis of performance measures for binary classifiers', In *IEEE International Conference on Data Mining*, pp. 517–526.

Peiris, M.T.R. (2008), *Lapses in Responsiveness: Characteristics and Detection from the EEG*, PhD thesis, University of Canterbury.

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', *Journal of Sleep Research*, Vol. 15, No. 3, pp. 291–300.

Peiris, M.T.R., Davidson, P.R., Bones, P.J. and Jones, R.D. (2011), 'Detection of lapses in responsiveness from the EEG', *Journal of Neural Engineering*, Vol. 8, p. 016003 (15pp).

Pereda, E., Quiroga, R.Q. and Bhattacharya, J. (2005), 'Nonlinear multivariate analysis of neurophysiological signals', *Progress in Neurobiology*, Vol. 77, No. 1-2, pp. 1–37.

Planet, S. and Iriondo, I. (2012), 'Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition', In *7th Iberian Conference on Information Systems and Technologies*, pp. 1–6.

Plett, M.I. (2007), 'Transient detection with cross wavelet transforms and wavelet coherence', *IEEE Transactions on Signal Processing*, Vol. 55, No. 5, pp. 1605–1611.

Poudel, G.R., Jones, R.D., Innes, C.R.H., Watts, R., Davidson, P.R. and Bones, P.J. (2010), 'Measurement of BOLD changes due to cued eye-closure and stopping during a continuous visuomotor task via model-based and model-free approaches', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 18, No. 5, pp. 479–488.

Poudel, G.R. (2010), *Functional magnetic resonance imaging of lapses of responsiveness during Visuomotor tracking*, PhD thesis, The University of Otago.

Poudel, G.R., Innes, C.R., Bones, P.J., Watts, R. and Jones, R.D. (2014), 'Losing the struggle to stay awake: Divergent thalamic and cortical activity during microsleeps', *Human brain mapping*, Vol. 35, No. 1, pp. 257–269.

Poudel, G.R., Innes, C.R. and Jones, R.D. (2018), 'Temporal evolution of neural activity and connectivity during microsleeps when rested and following sleep restriction', *NeuroImage*, Vol. 174, pp. 263–273.

Powers, D.M.W. (2011), 'Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation', *Journal of Machine Learning Technologies*, Vol. 2, No. 1, pp. 37–63.

Quiroga, R.Q., Kraskov, A., Kreuz, T. and Grassberger, P. (2002), 'Performance of different synchronization measures in real data: A case study on electroencephalographic signals', *Physical Review E*, Vol. 65, No. 4, p. 041903(14pp).

Quitadamo, L.R., Cavrini, F., Sbernini, L., Riillo, F., Bianchi, L., Seri, S. and Saggio, G. (2017), 'Support vector machines to detect physiological patterns for EEG and EMG-based human–computer interaction: a review', *Journal of Neural Engineering*, Vol. 14, No. 1, pp. 011001 (pp. 1–27).

Quyen, M.L.V., Foucher, J., Lachaux, J.P., Rodriguez, E., Lutz, A., Martinerie, J. and Varela, F.J. (2001), 'Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony', *Journal of Neuroscience Methods*, Vol. 111, No. 2, pp. 83–98.

Rahman, A. and Verma, B. (2010), 'A novel ensemble classifier approach using weak classifier learning on overlapping clusters', In *The International Joint Conference on Neural Networks*, pp. 1–7.

Rocca, D.L., Campisi, P., Vegso, B., Cserti, P., Kozmann, G., Babiloni, F. and Fallani, F.D.V. (2014), 'Human brain distinctiveness based on EEG spectral coherence connectivity', *IEEE Transactions on Biomedical Engineering*, Vol. 61, No. 9, pp. 2406–2412.

Rojas, G., Alvarez, C., Montoya Moya, C., de la Iglesia Vaya, M., Cisternas, J. and Gálvez, M. (2018), 'Study of resting-state functional connectivity networks using EEG electrodes position as seed', *Frontiers in Neuroscience*, Vol. 12, No. 235, p. 12.

Saeys, Y., naki Inza, I. and naga, P.L. (2007), 'A review of feature selection techniques in bioinformatics', *Bioinformatics*, Vol. 23, No. 19, pp. 2507–2517.

Saito, T. and Rehmsmeier, M. (2015), 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', *PLOS ONE*, Vol. 10, No. 3, pp. 1–21.

Sanei, S. and Chambers, J.A. (2007), *EEG Signal Processing*, John Wiley & Sons.

Schlögl, A. (2006), 'A comparison of multivariate autoregressive estimators', *Signal Processing*, Vol. 86, No. 9, pp. 2426–2429.

Schlögl, A. and Supp, G. (2006), 'Analyzing event-related EEG data with multivariate autoregressive parameters', In C. Neuper and W. Klimesch (editors), *Event-Related Dynamics of Brain Oscillations*, Vol. 159 of *Progress in Brain Research*, pp. 135 –147, Elsevier.

Shoorangiz, R. (2018), *Prediction of Microsleeps from EEG using Bayesian Approaches*, PhD thesis, University of Canterbury.

Shoorangiz, R., Weddell, S.J. and Jones, R.D. (2016), 'Prediction of microsleeps from EEG: Preliminary results', In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4650–4653.

Shoorangiz, R., Weddell, S.J. and Jones, R.D. (2017), 'Bayesian multi-subject factor analysis to predict microsleeps from EEG power spectral features', In *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4183–4186.

Silfverhuth, M.J., Hintsala, H., Kortelainen, J. and Seppänen, T. (2012), 'Experimental comparison of connectivity measures with simulated EEG signals', *Medical and Biological Engineering*, Vol. 50, No. 7, pp. 683–688.

Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A. and Demchuk, E. (2003), 'Nearest neighbor estimates of entropy', *American Journal of Mathematical and Management Sciences*, Vol. 23, No. 3-4, pp. 301–321.

Sirois, B., Trutschel, U., Edwards, D., Sommer, D. and Golz, M. (2010), 'Predicting accident probability from frequency of microsleep events', *World Congress on Medical Physics and Biomedical Engineering*, Vol. 25, pp. 2284–2286.

Strand, O.N. (1977), 'Multichannel complex maximum entropy (autoregressive) spectral analysis', *IEEE Transactions on Automatic Control*, Vol. 22, No. 4, pp. 634–640.

Straube, S. and Krell, M.M. (2014), 'How to evaluate an agent's behavior to infrequent events? – Reliable performance estimation insensitive to class distribution', *Frontiers in Computational Neuroscience*, Vol. 8, p. 43.

Sun, J., Hong, X. and Tong, S. (2012), 'Phase synchronization analysis of EEG signals: An evaluation based on surrogate tests', *IEEE Transactions on Biomedical Engineering*, Vol. 59, No. 8, pp. 2254–2263.

Tefft, B.C. (2010), 'Asleep at the wheel: The prevalence and impact of drowsy driving', https://aaafoundation.org/wp-content/uploads/2018/02/2010DrowsyDrivingReport.pdf, accessed: August 2018.

Tefft, B.C. (2012), 'Prevalence of motor vehicle crashes involving drowsy drivers, United States, 1999–2008', *Accident Analysis & Prevention*, Vol. 45, pp. 180–86.

Thomas, M., Sing, H., Belenky, G., Holcomb, H., Mayberg, H., Dannals, R., Wagner JR., H., Thorne, D., Popp, K., Rowland, L., Welsh, A., Balwinski, S. and Redmond, D. (2000), 'Neural basis of alertness and cognitive performance impairments during sleepiness. I. Effects of 24 h of sleep deprivation on waking human regional brain activity', *Journal of Sleep Research*, Vol. 9, No. 4, pp. 335–352.

Thomas, P., Morris, A., Talbot, R. and Fagerlind, H. (2013), 'Identifying the causes of road crashes in Europe', In *57th Annual Conference on Annals of Advances in Automotive Medicine*, pp. 13–22.

Tong, S. and Thakor, N.V. (2009), *Quantitative EEG analysis methods and clinical applications*, Artech House.

Toppi, J., Astolfi, L., Poudel, G.R., Innes, C.R., Babiloni, F. and Jones, R.D. (2016), 'Time-varying effective connectivity of the cortical neuroelectric activity associated with behavioural microsleeps', *NeuroImage*, Vol. 124, pp. 421–432.

Torrence, C. and Webster, P. (1999), 'Interdecadal changes in the ENSO-monsoon system', *Journal of Climate*, Vol. 12, No. 8, pp. 2679–2690.

Torrence, C. and Compo, G.P. (1998), 'A practical guide to wavelet analysis', *Bulletin of the American Meteorological society*, Vol. 79, No. 1, pp. 61–78.

Torres, G., Cinelli, M.P., Hynes, A.T., Kaplan, I.S. and Leheste, J.R. (2014), 'Electroencephalogram mapping of brain states', *Journal of Neuroscience and Neuroengineering*, Vol. 3, No. 2, pp. 1–5.

Valdovinos, R.M. and Sanchez, J.S. (2009), 'Combining multiple classifiers with dynamic weighted voting', In Corchado, E and Wu, X and Oja, E and Herrero, A and Baruque, B (editor), *Hybrid Artifcial Intelligence System*, Vol. 5572 of *Lecture Notes in Artificial Intelligence*, pp. 510–516.

Vanlaar, W., Simpson, H., Mayhew, D. and Robertson, R. (2008), 'Fatigued and drowsy driving: A survey of attitudes, opinions and behaviors', *Journal of Safety Research*, Vol. 39, No. 3, pp. 303–309.

Walters-Williams, J. and Li, Y. (2009), 'Estimation of mutual information: A survey', In *International Conference on Rough Sets and Knowledge Technology*, Springer, pp. 389–396.

Wang, G., Sun, Z., Tao, R., Li, K., Bao, G. and Yan, X. (2016), 'Epileptic seizure detection based on partial directed coherence analysis', *IEEE Journal of Biomedical and Health Informatics*, Vol. 20, No. 3, pp. 873–879.

Wang, H.E., Bénar, C.G., Quilichini, P.P., Friston, K.J., Jirsa, V.K. and Bernard, C. (2014a), 'A systematic framework for functional connectivity measures', *Frontiers in Neuroscience*, Vol. 8, p. 405.

Wang, Y.T., Huang, K.C., Wei, C.S., Huang, T.Y., Ko, L.W., Lin, C.T., Cheng, C.K. and Jung, T.P. (2014b), 'Developing an EEG-based on-line closed-loop lapse detection and mitigation system', *Frontiers in Neuroscience*, Vol. 8, No. 321, p. 11pp.

World Health Organization (2015), 'Global status report on road safety', http://apps.who.int/iris/bitstream/10665/189242/1/9789241565066_eng.pdf, accessed: August 2018.

Yang, J., yu Yang, J., Zhang, D. and feng Lu, J. (2003), 'Feature fusion: parallel strategy vs. serial strategy', *Pattern Recognition*, Vol. 36, pp. 1369–1381.

Yu, X., Chum, P. and Sim, K.B. (2014), 'Analysis the effect of PCA for feature reduction in non-stationary EEG based motor imagery of BCI system', *Optik - International Journal for Light and Electron Optics*, Vol. 125, No. 3, pp. 1498–1502.

Yuan, G.X., Ho, C.H. and Lin, C.J. (2012), 'Recent advances of large-scale linear classification', *Proceedings of the IEEE*, Vol. 100, No. 9, pp. 2584–2603.

Yule, G.U. (1900), 'On the association of attributes in statistics', *Philosophical Transactions of the Royal Society of London*, Vol. 194, pp. 257–319.

Zhou, Z.H. (2012), *Ensemble Methods: Foundations and Algorithms*, Chapman and Hall/CRC.

Appendix A

INDIVIDUAL NON-NORMALIZED AND NON-CAUSAL FEATURES

A.1 DETECTION AND PREDICTION OF MICROSLEEP STATES

A decline in performances of non-normalized and non-causal features with classifier-independent feature-selection technique and an LDA classifier are shown in Figure A.1. The overall highest mean detection and prediction (at τ = 1 s) performance metrics (phi = 0.48, AUC_{PR} = 0.51, AUC_{ROC} = 0.95) and (phi = 0.44, AUC_{PR} = 0.46, AUC_{ROC} = 0.94) were achieved with joint entropy features.

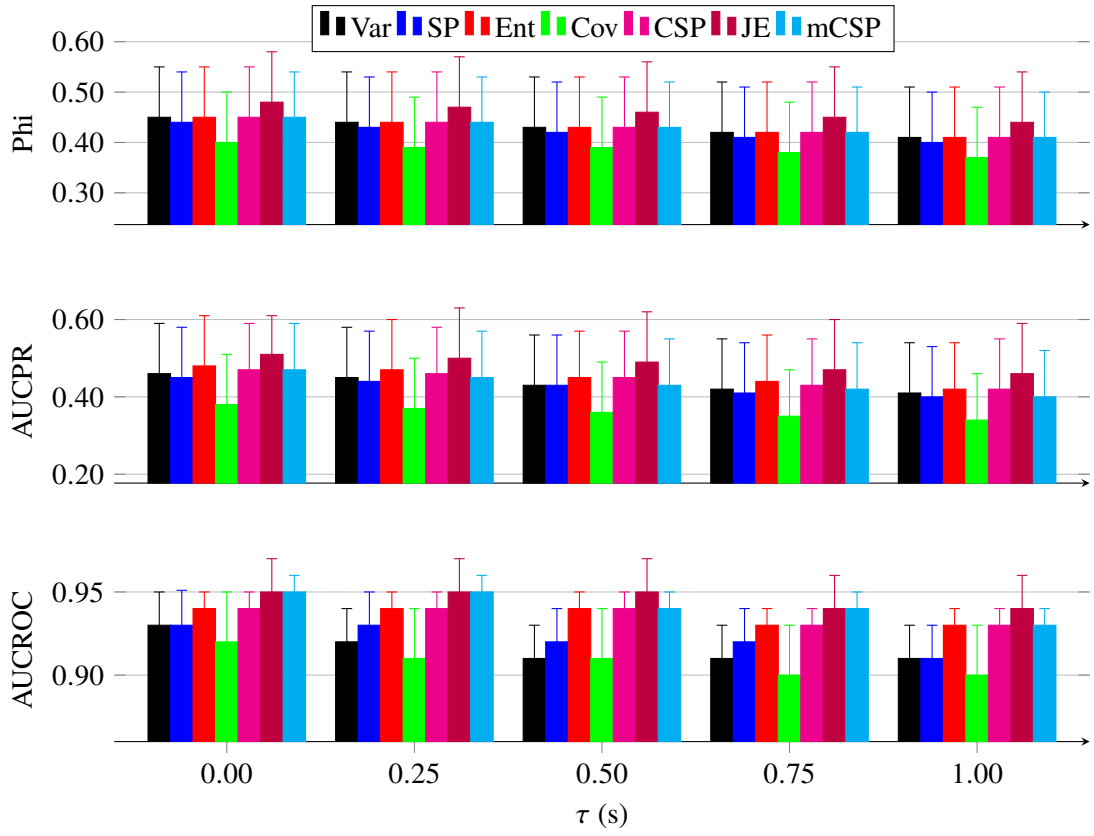

Figure A.1 Performance (mean ± SE) of microsleep state prediction using non-normalized features with classifier-dependent feature-selection and a single LDA classifier for τ = 0.00 –1.00 s. mCSP represents multivariate cross-spectral power features.

A decline in performances of non-normalized and non-causal features with classifier-independent feature-selection technique and an LSVM classifier are shown in Figure A.2. The highest mean detection phi of 0.45 was achieved with both joint entropy and multivariate cross-spectral power features. Whereas, highest mean detection AUC_{PR} of 0.48 and $AUC_{ROC}$of 0.95 were achieved with joint entropy and multivariate cross-spectral power features respectively. Similarly, highest mean prediction (at τ = 1 s) phi of 0.42 and AUC_{PR} of 0.43 were achieved with joint entropy features. Whereas, both joint entropy and multivariate cross-spectral power features gave the highest mean prediction AUC_{ROC} of 0.93.

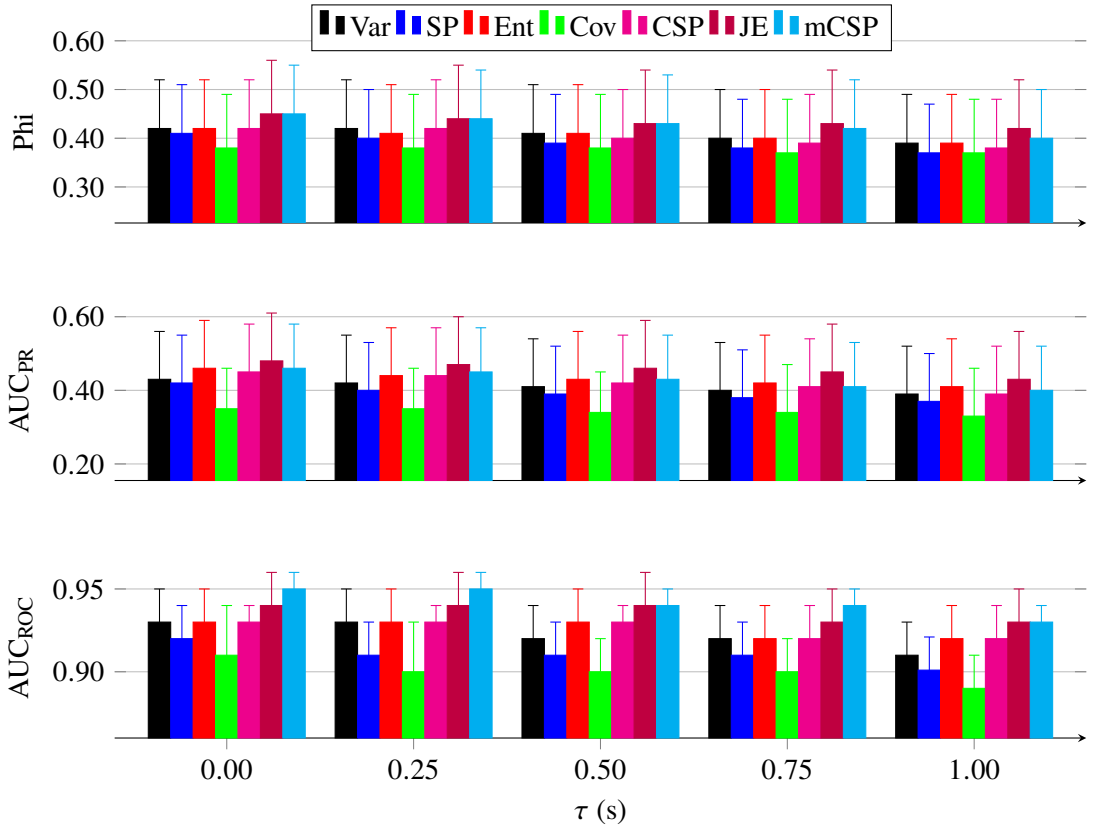

Figure A.2 Performance (mean ± SE) of microsleep state prediction using non-normalized features with classifier-dependent feature-selection and a single LSVM classifier for τ = 0.00 –1.00 s.

With classifier-independent feature-selection technique and an LDA classifier, performances of non-normalized and non-causal features with respect to prediction time τ are shown in Figure A.3. Multivariate cross-spectral power features gave the highest mean detection phi of 0.45, whereas, prediction (at τ = 1 s) phi of 0.41 was achieved with both entropy and multivariate cross-spectral power features. Similarly, the highest mean detection AUC_{PR} of 0.51 was achieved with multivariate cross-spectral power features, whereas, the highest mean prediction AUC_{PR} of 0.45, detection and prediction AUC_{ROC}s of 0.95 and 0.93, respectively were achieved with both entropy and joint entropy features.

Figure A.4 shows the performances of non-normalized and non-causal features with classifier-independent feature-selection technique and an LSVM classifier.The highest mean detection and prediction (at τ = 1 s) phis of 0.43 and 40, respectively were achieved with entropy and multivariate cross-spectral power features. In addition joint entropy gave highest mean detection phi of 0.43. Highest mean detection AUC_{PR} of 0.50 was achieved with joint entropy,

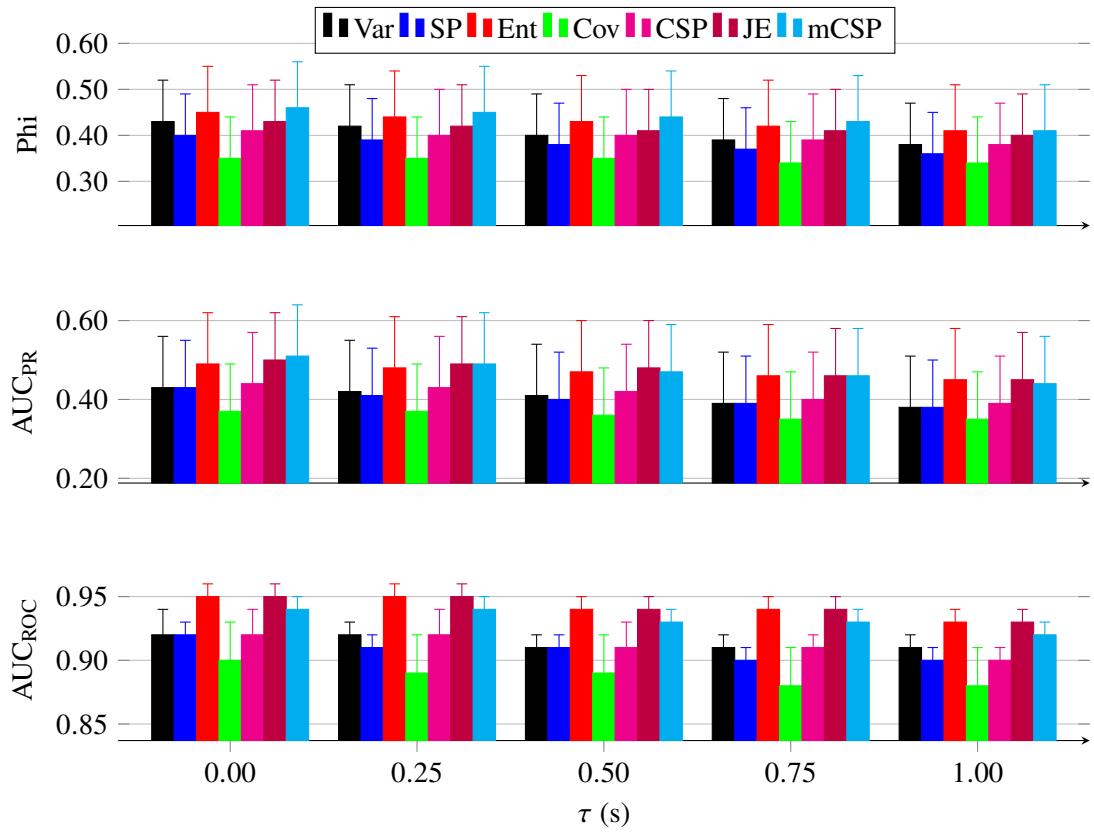

Figure A.3 Performance (mean ± SE) of microsleep state prediction using non-normalized features with classifier-independent feature selection and single LDA classifier for τ = 0.00 –1.00 s.

whereas, highest mean prediction AUC_{PR} of 0.45 was achieved with entropy features. Both information-theoretic features resulted in highest mean detection AUC_{ROC} of 0.95, whereas, highest mean prediction AUC_{ROC} of 0.94 was achieved with entropy features.

A.2 DETECTION AND PREDICTION OF MICROSLEEP ONSETS

Microsleep onset prediction performances, with respect to prediction time τ, of all non-normalized and non-causal feature sets with both feature-selection techniques and an LDA classifier are shown in Figure A.5.

Compared to state predictions, microsleep onset predictions exhibit faster performance drops across the feature sets. With increasing prediction time τ, all feature-types shown similar performance trends. However, joint entropy features with classifier-dependent feature-selection technique gave the highest mean phis and AUC_{PR}s for τ = 0–3 s, and the highest AUC_{ROC} across the prediction times. However, with classifier-independent feature-selection technique, highest phis for τ = 0–2 s were achieved with entropy, joint entropy, and multivariate cross-spectral power features. Entropy features, across the prediction times, gave the highest mean AUC_{ROC}. Microsleep onset prediction performances with LSVM classifier and across both feature selection techniques are shown in Figure A.6. Joint entropy features with classifier-dependent feature selection resulted in the highest mean phis and AUC_{PR}s for τ = 0–2 s and 0–3 s, respectively, and the highest AUC_{ROC} across the prediction times. However, with classifier-independent feature-selection technique, both joint entropy and multivariate cross spectral power features,

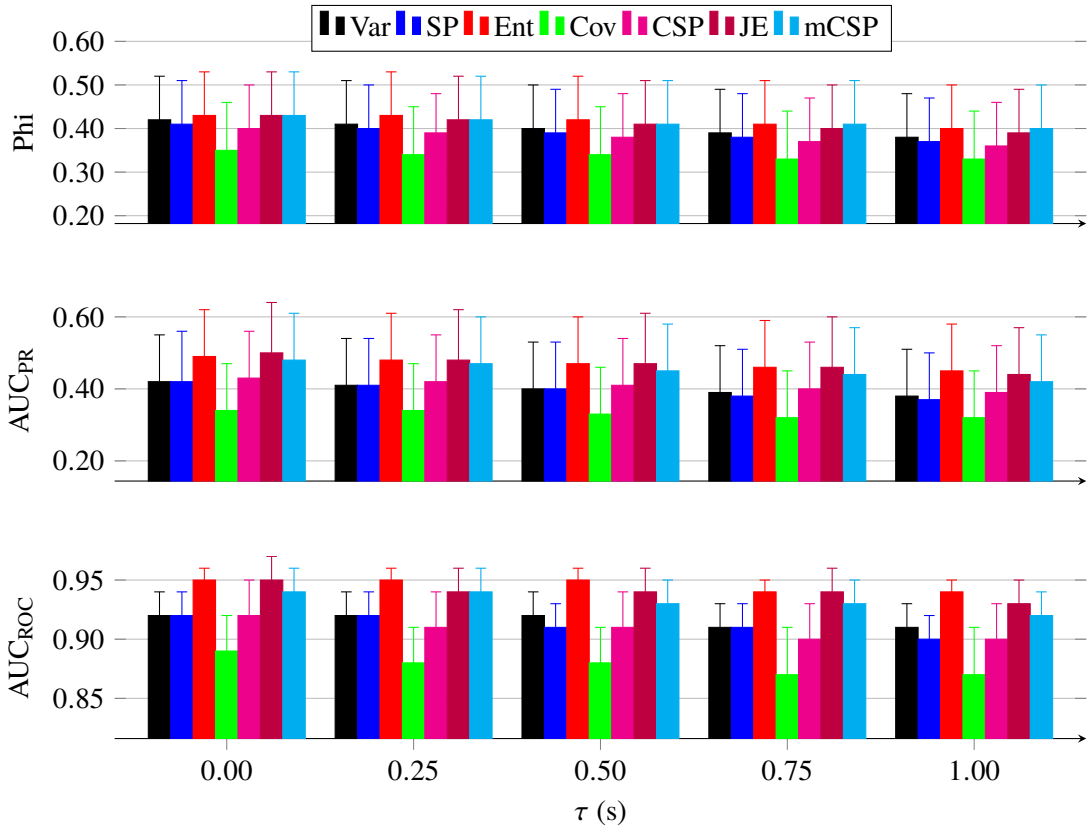

Figure A.4 Performance (mean ± SE) of microsleep state prediction using non-normalized features with classifier-independent feature selection and single LSVM classifier for τ = 0.00 –1.00 s.

except for τ = 3 s, gave the highest mean phis. Whereas, except for detection, prediction AUC_{PR}s of all feature-types were the same. The highest AUC_{ROC}, across the prediction times, was achieved with entropy features.

Joint entropy features with classifier-dependent feature selection resulted in the highest mean phis and AUC_{PR}s for τ = 0–2 s and 0–3 s, respectively, and the highest AUC_{ROC} across the prediction times. However, with classifier-independent feature-selection technique, both joint entropy and multivariate cross spectral power features, except for τ = 3 s, gave the highest mean phis. Whereas, except for detection, prediction AUC_{PR}s of all feature-types were the same. The highest AUC_{ROC}, across the prediction times, was achieved with entropy features.

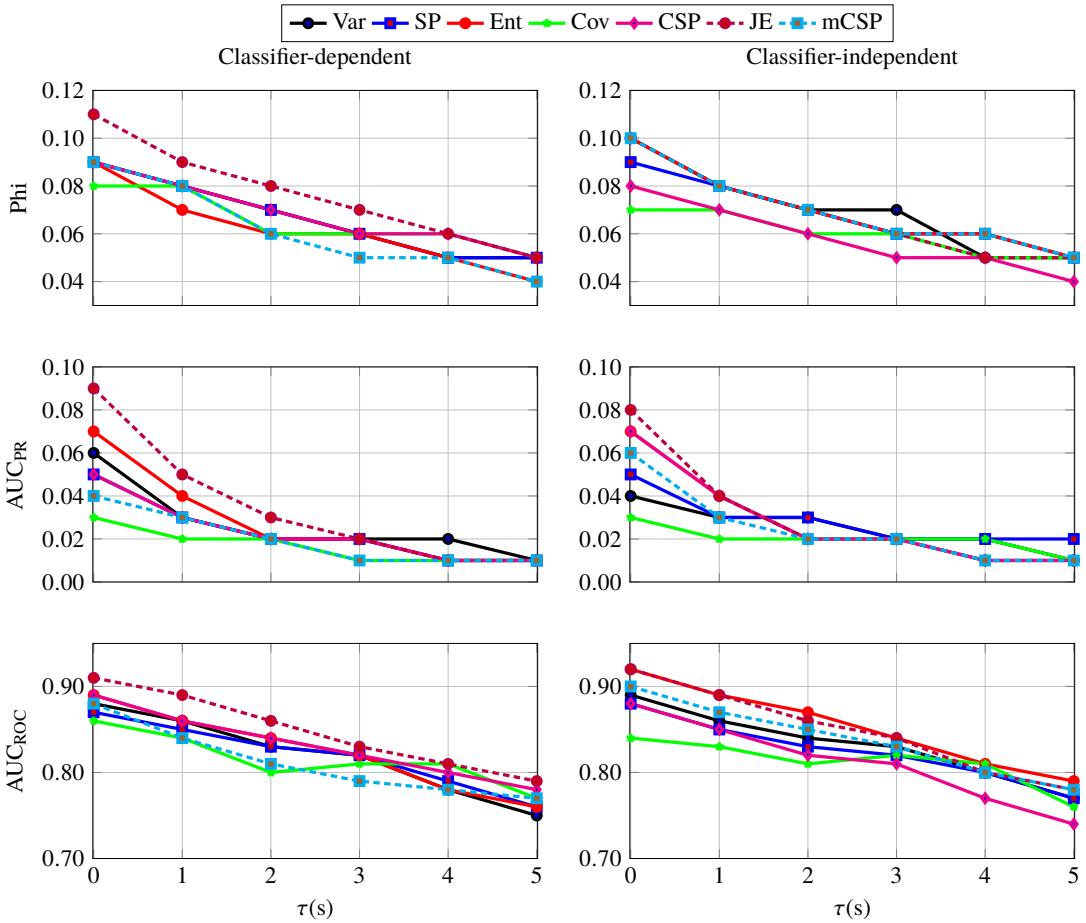

Figure A.5 Mean performance microsleep onset prediction with both feature selection techniques and LDA classifier on non-normalized and non-causal feature sets, for $\tau = 0$ –5s.

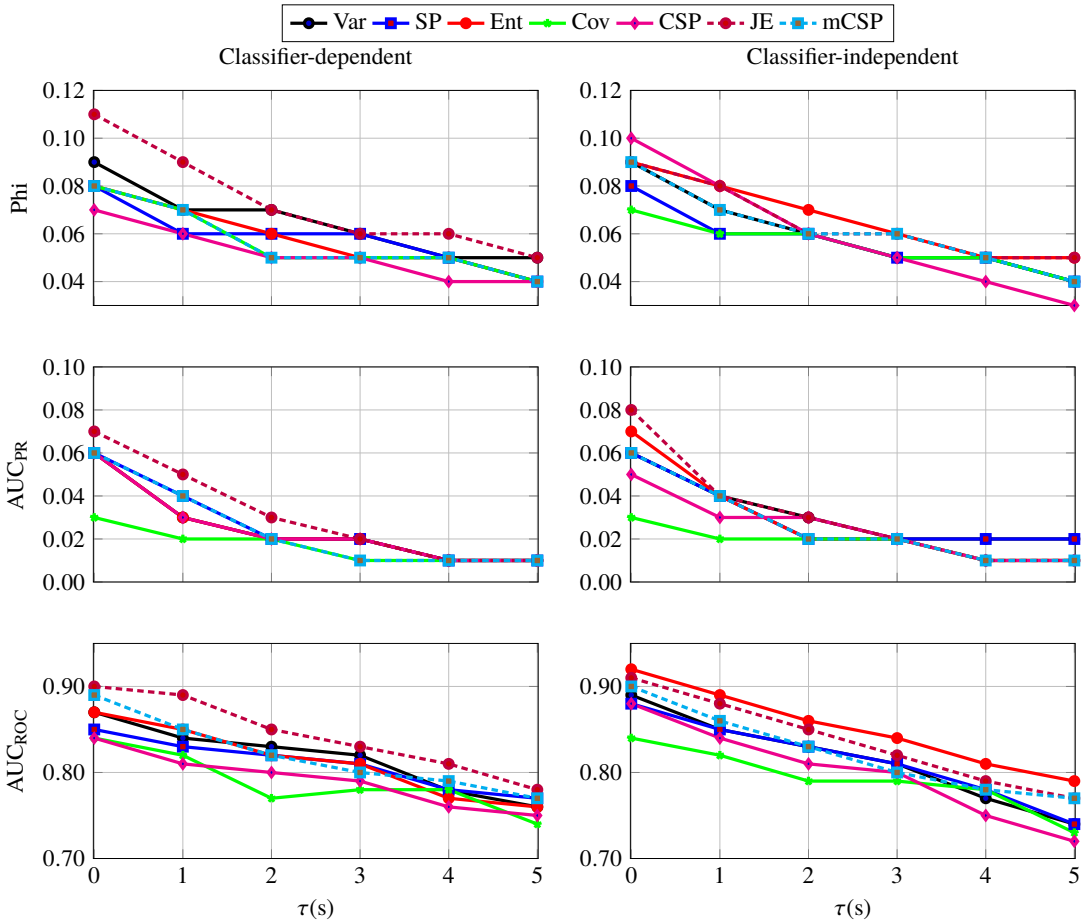

Figure A.6 Mean performance microsleep onset prediction with both feature selection techniques and LSVM classifier on non-normalized and non-causal feature sets, for $\tau = 0$ –5s.

Appendix B

DATA FUSION

B.1 DETECTION AND PREDICTION OF MICROSLEEP STATES

The performances of feature-level fusion with both feature-selection techniques and classifiers with increasing prediction time τ are shown in Figure B.1.

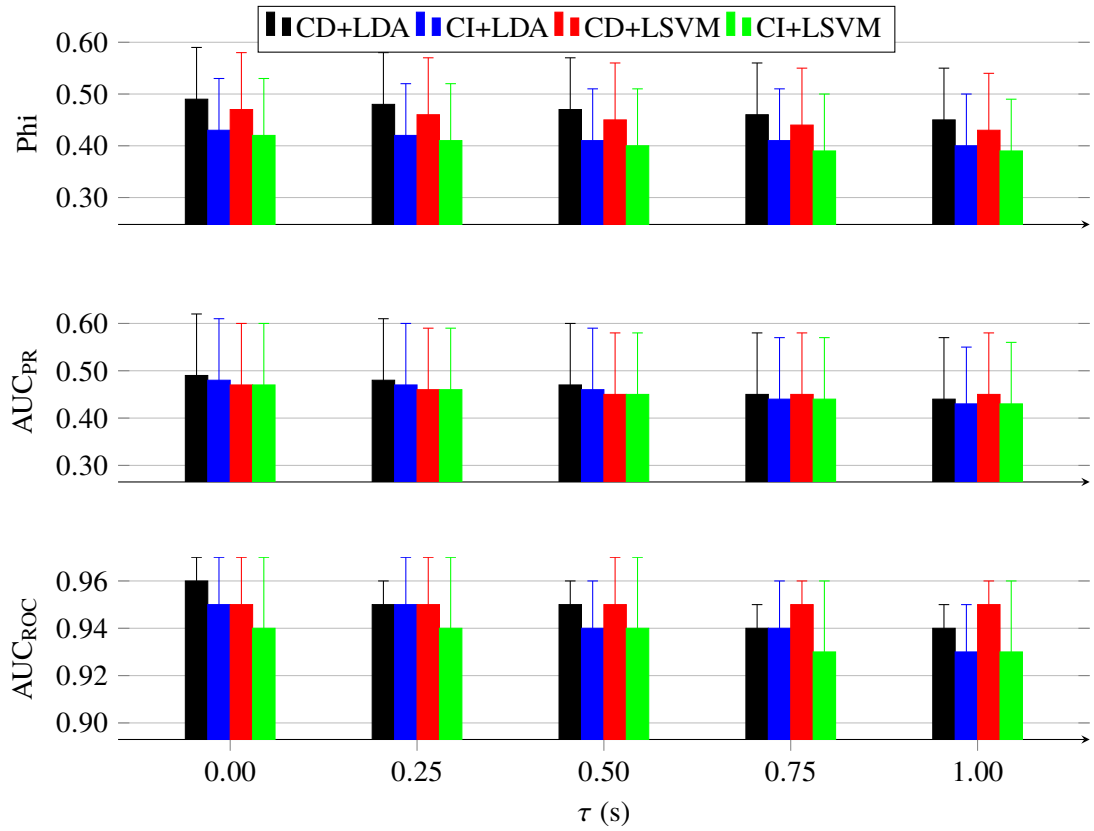

Figure B.1 Performance (mean ± SE) of microsleep state prediction using feature-level fused features with both feature selection techniques and classifiers for τ = 0.00 –1.00 s. Classifier-dependent and classifier-independent techniques were abbreviated as CD and CI, respectively.

Across the prediction times τ, compared to classifier-independent, performances of classifier-dependent feature-selection with both classifiers were higher. The highest mean detection and prediction (at τ = 1 s) phis of 0.49 and 0.45, respectively were achieved with an LDA classifier. Similarly, the highest detection AUC_{PR} and AUC_{ROC} of 0.49 and 0.96, respectively

were achieved with classifier-dependent feature-selection and an LDA. However, prediction AUC_{PR} and AUC_{ROC} of 0.45 and 0.95, respectively, were achieved with classifier-independent feature-selection and an LSVM classifier.

Figure B.2 shows the performances (for τ = 0–1 s) of decision-level fusion with both feature-selection techniques and both types of classifiers on microsleep states. On decision-level

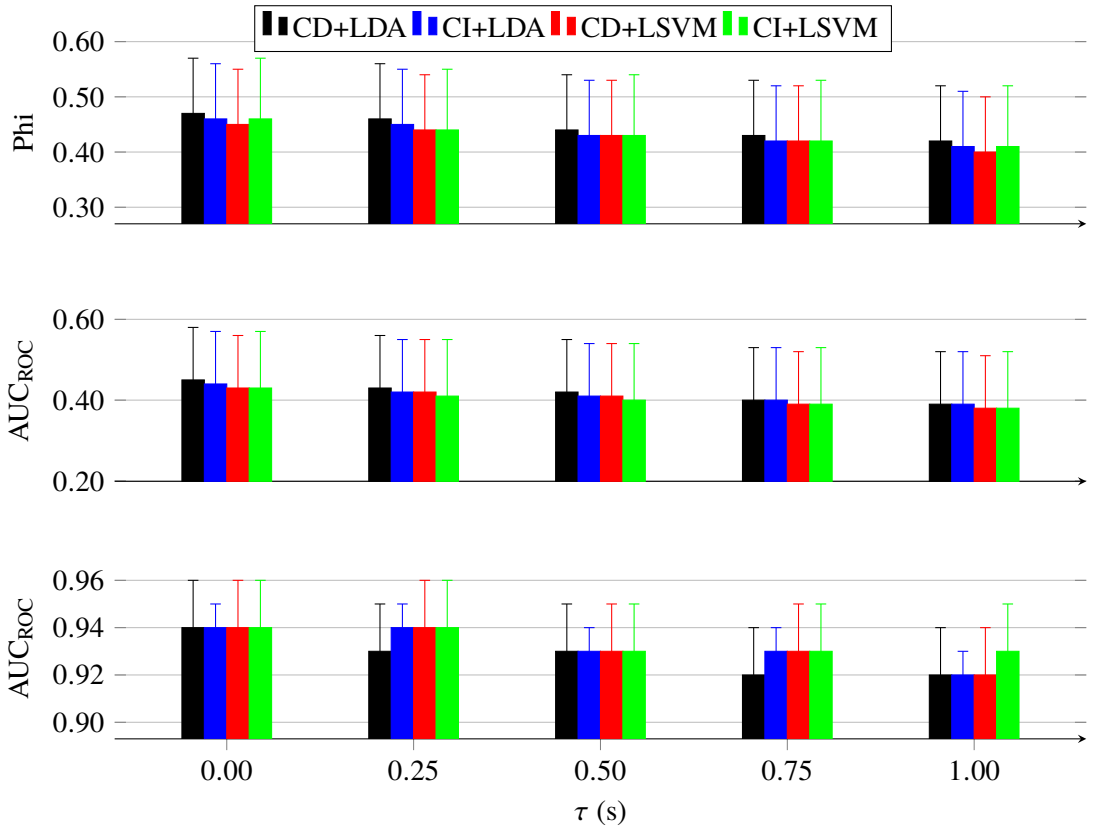

Figure B.2 Performance (mean ± SE) of microsleep state prediction using decision-level fused features with both feature selection techniques and classifiers for τ = 0.00 –1.00 s. Classifier-dependent and classifier-independent techniques were abbreviated as CD and CI, respectively.

fusion, classifier-dependent feature-selection technique and LDA classifiers, across the prediction times τ, gave the highest phi and AUC_{PR}. The highest detection phi and AUC_{PR} were 0.47 and 0.45, respectively. Whereas, the highest prediction (at τ = 1 s) phi and AUC_{PR} were 0.42 and 0.39 respectively. All feature-selection and classifiers combinations gave the same detection AUC_{ROC} of 0.94, whease, the highest prediction (at τ = 1 s) AUC_{ROC} of 0.93 was achieved with classifier-independent feature-selection and LSVM classifiers.

B.2 DETECTION AND PREDICTION OF MICROSLEEP ONSETS

Using feature-level fusion, the highest mean detection phi of 0.10 and AUC_{PR} of 0.09 were achieved with classifier-dependent feature-selection technique and both classifiers. Same phi of 0.10 was also achieved with classifier-independent feature-selection and an LSVM classifier. However, all combinations of feature-selection techniques and classifiers, resulted in same prediction (at τ = 5 s) phi and AUC_{PR} of 0.05 and 0.01 respectively. Except for classifier-dependent feature-selection and an LSVM classifier, all combinations resulted in detection and

prediction AUC_{ROC} of 0.91 and 0.77, respectively.

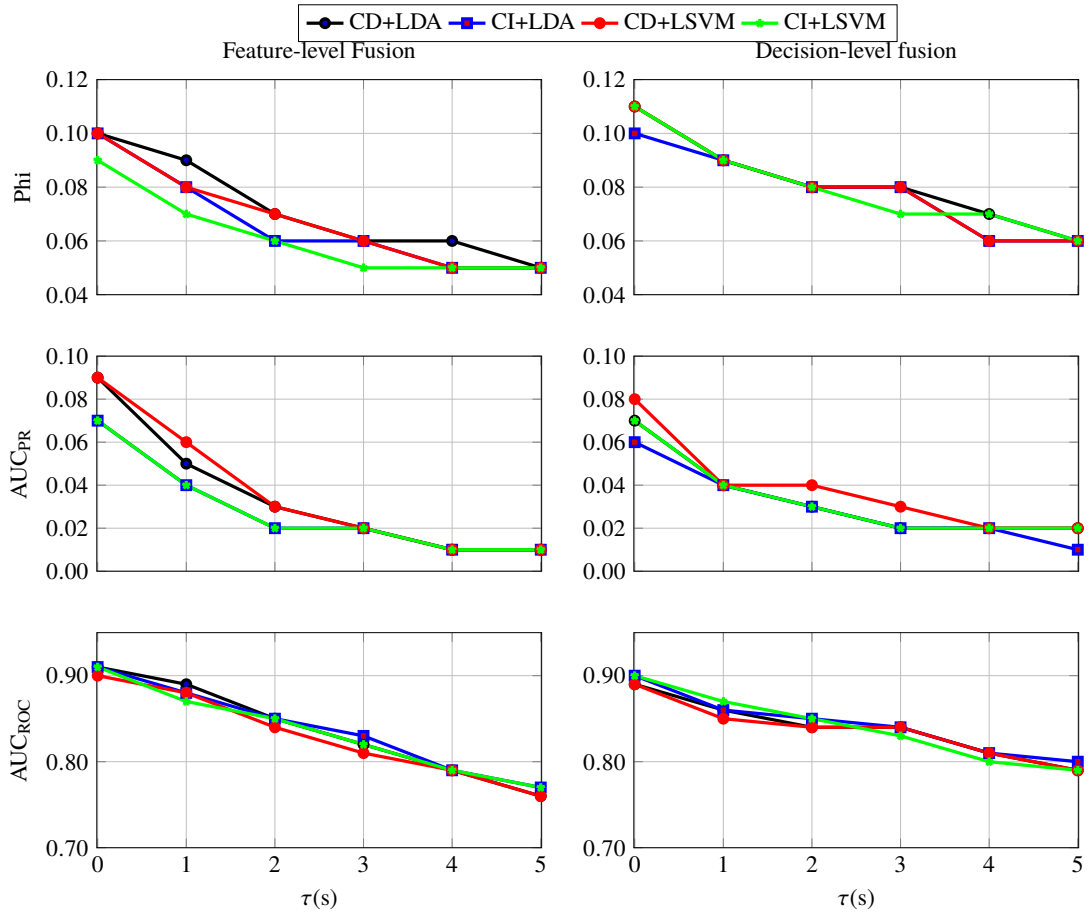

Figure B.3 Mean performance microsleep onset prediction with both feature-selection techniques and classifiers on feature-level and decision-level fused data, for τ = 0 –5 s. Classifier-dependent and classifier-independent techniques were abbreviated as CD and CI respectively.

Using decision-level fusion, mean detection phi of 0.11 was achieved with classifier-dependent feature-selection and both types of classifiers, whereas, all combinations resulted in the same prediction (at τ = 5 s) phi of 0.06. The highest mean detection AUC_{PR} of 0.08 was achieved with classifier-dependent feature-selection and LSVM classifiers. The highest mean detection and prediction AUC_{ROC} of 0.90 and 0.80, respectively, were achieved with classifier-independent feature-selection and LDA classifiers.

Appendix C

ENSEMBLE CLASSIFICATION TECHNIQUES

C.1 DECTECTION AND PREDICTION OF MICROSLEEP STATES

In addition to a single LSVM classifier, mean performances of four LSVM-based ensemble techniques with classifier-dependent feature selection at predicting microsleep states for increasing prediction time τ = 0–1 s are shown in Figure C.1.

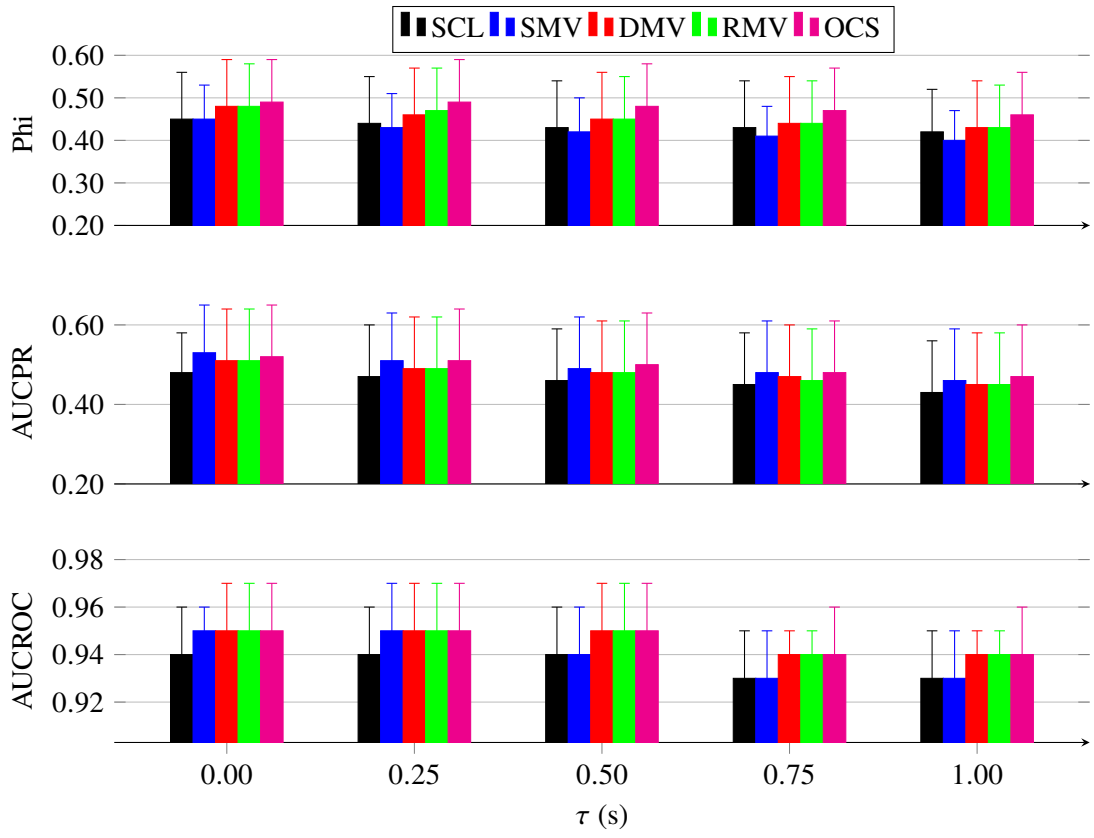

Figure C.1 Performances (mean±SE) of microsleep state prediction using joint entropy features with classifier-dependent feature selection and LSVM classifier, for τ = 0 –1 s. Single classifier, standard majority voting, and overlapping clusters were abbreviated as SCL, SMV, and OCS, respectively.

Compared to a single LSVM classifier and except for standard majority voting, all proposed ensemble techniques, and across the prediction times, resulted in the higher mean phis. Highest mean detection and prediction (at τ = 1 s) phis of 0.49 and 0.46, respectively were achieved with

overlapping clusters. The highest mean detection AUC_{PR} of 0.53 was achieved with standard majority voting, whereas, the highest mean prediction (at τ = 1 s) AUC_{PR} of 0.47 was achieved with overlapping clusters. Mean detection AUC_{ROC}s of all ensemble techniques were the same and compared to a single classifier, were slightly higher (i.e., 0.94 vs 0.95). Except for majority voting, all ensemble techniques at τ = 1 s, resulted in mean AUC_{ROC}s of 0.94.

Using classifier-independent features selection, performances of a single LDA classifier and four LDA-based ensemble techniques at predicting microsleep states for increasing prediction time τ = 0–1 s are shown in Figure C.2.

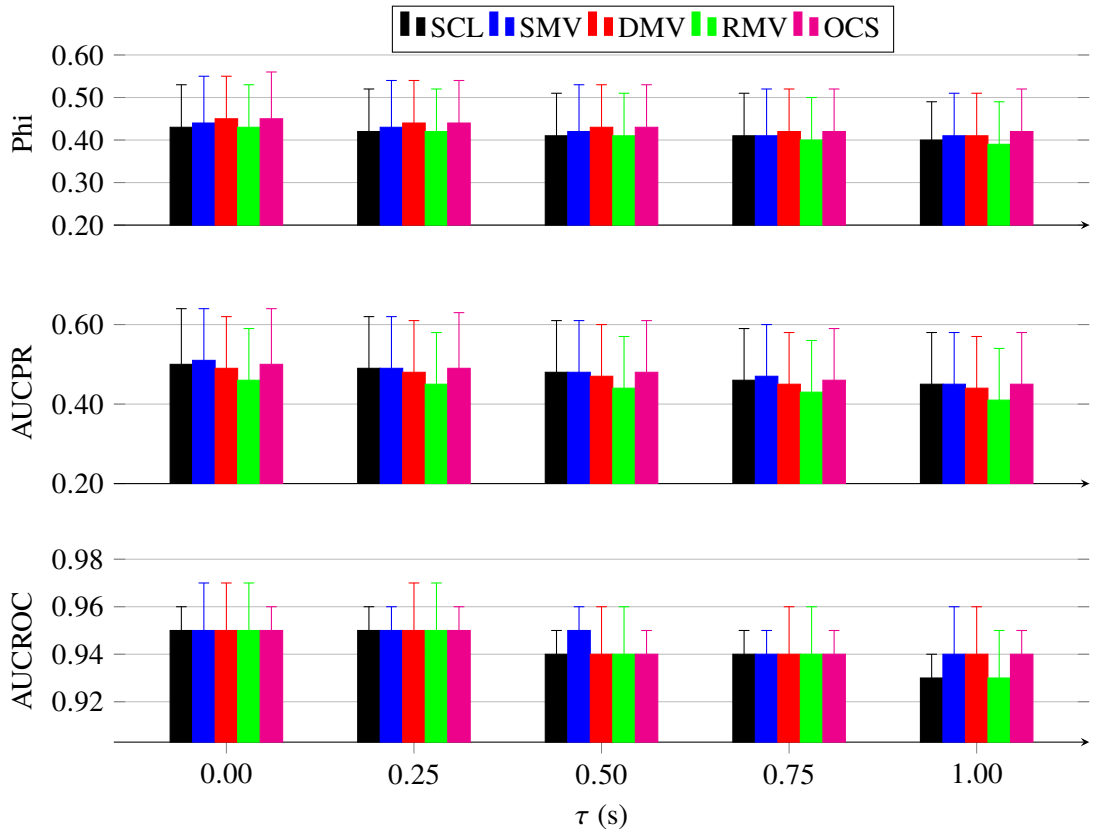

Figure C.2 Performances (mean±SE) of microsleep state prediction using joint entropy features with classifier-independent feature selection and LDA classifier, for τ = 0 –1s. Single classifier, standard majority voting, and overlapping clusters were abbreviated as SCL, SMV, and OCS, respectively.

Except for RMV, all ensemble techniques, compared to a single LDA classifier, resulted in the higher mean phis. Mean detection phi of 0.45 was achieved with DMV and overlapping clusters, whereas, mean prediction (at τ = 1 s) phi of 0.42 was achieved with the overlapping clusters. Mean detection AUC_{PR}s of both RMV and DMV, compared to a single classifier, were lower. The highest mean detection AUC_{PR}s of 0.51 was achieved with standard majority voting, whereas, mean prediction AUC_{ROC} of 0.45 was achieved with single classifier, majority voting, and overlapping clusters. Mean detection AUC_{ROC}s of all of the ensemble techniques and a single classifier were the same (0.95). Comparatively higher prediction AUC_{ROC} of 0.94 was achieved with majority voting, DMV, and overlapping clusters.

Using classifier-independent features selection, mean performances of a single LSVM classifier and four LSVM-based ensemble techniques at predicting microsleep states for increasing prediction time τ = 0–1 s are shown in Figure C.3.

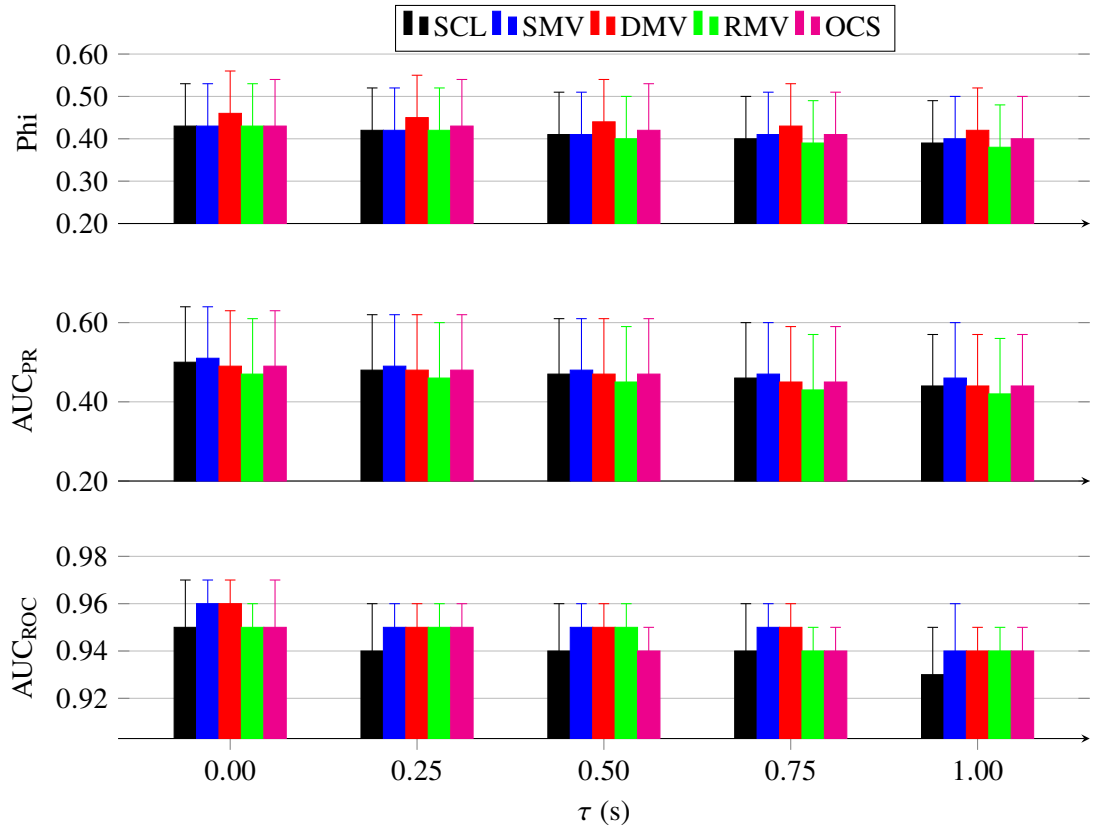

Figure C.3 Performances (mean±SE) of microsleep state prediction using joint entropy features with classifier-independent feature selection and LSVM classifier, for $\tau = 0$ –1s. Single classifier, standard majority voting, and overlapping clusters were abbreviated as SCL, SMV, and OCS, respectively.

The highest mean detection and prediction (at $\tau = 1$ s) phi of 0.46 and 0.42, respectively were achieved with DMV. The highest detection and prediction AUC_{PR}s of 0.51 and 0.46, respectively were achieved with majority voting. The highest detection AUC_{ROC} of 0.96 was achieved with both majority voting and DMV, whereas, all ensemble techniques gave the prediction ($\tau = 1$ s) AUC_{ROC} of 0.94.

C.2 DETECTION AND PREDICTION OF MICROSLEEP ONSETS

With classifier-dependent feature-selection technique, mean performances of both single classifiers (i.e., LDA, LSVM) and four ensemble techniques (using both types of classifiers) at predicting microsleep onsets for increasing prediction time $\tau = 0$–5 s are shown in Figure C.4.

Using LDA as single and base classifiers, standard majority voting resulted in the highest mean detection of phi 0.13, whereas, both DMV and RMV resulted mean prediction (at $\tau = 5$ s) phi of 0.06. The highest mean detection AUC_{PR} of 0.10 was achieved with overlapping clusters, whereas, at $\tau = 5$ s, the same prediction AUC_{PR} of 0.01was achieved with all the ensemble techniques and a single classifier. The highest mean detection AUC_{ROC} of 0.93 was achieved with overlapping clusters, whereas, at $\tau = 5$ s, the same mean AUC_{ROC}s of 0.80 was achieved with the all ensemble techniques except majority voting.

Using LSVM as single and base classifiers, DMV, RMV, and overlapping clusters, all gave the highest mean detection phis of 0.13, whereas, mean prediction (at $\tau = 5$ s) phi of 0.06 was

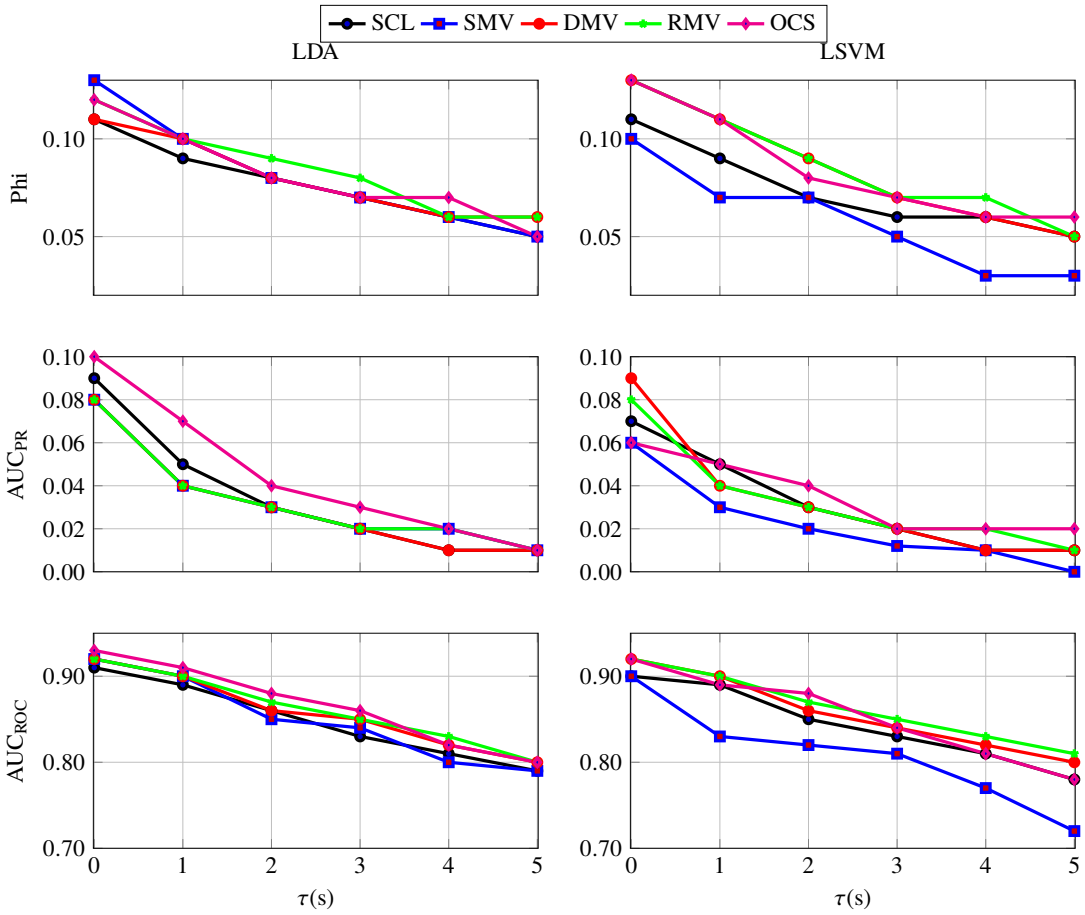

Figure C.4 Mean performances at predicting microsleep onsets for τ = 0 –5 s, using joint entropy features with classifier-dependent feature-selection and both LDA and LSVM classifiers in single and ensemble configurations. Single classifier, standard majority voting, and overlapping clusters were abbreviated as SCL, SMV, and OCS, respectively.

achieved with overlapping clusters. The highest detection AUC_{PR} of 0.09 was achieved with DMV, whereas, at τ = 5 s, overlapping clusters resulted in the highest mean AUC_{PR} of 0.03. Except for majority voting, all the ensemble techniques resulted in the detection AUC_{ROC}s of 0.92, whereas, the highest prediction AUC_{ROC} of 0.81 was achieved with RMV. Majority voting contrary to microsleep states, gave the lowest performances on microsleep onsets.

With classifier-independent feature-selection technique, mean performances of both single classifiers (i.e., LDA, LSVM) and four ensemble techniques (using both types of classifiers) at predicting microsleep onsets for increasing prediction time τ = 0–5 s are shown in Figure C.5.

Using LDA as base classifiers, RMV resulted in the highest detection phi of 0.12 but with fastest drop. Both DMV and overlapping clusters gave the highest prediction (τ = 5 s) phi of 0.06. Detection AUC_{PR} of both a single classifier and majority voting was the higher than all of the ensemble techniques, i.e., 0.08 vs 0.07, whereas, all of the techniques resulted in the same prediction mean AUC_{PR} of 0.01. All of the ensemble techniques resulted in the same mean detection AUC_{ROC} of 0.92, whereas, at τ = 5 s, AUC_{ROC} of 0.80 was achieved with both RMV and DMV.

Using LSVM as base classifiers, the highest mean detection phi of 0.12 was achieved with both RMV and DMV, whereas, the highest prediction (at τ = 5 s) phi of 0.06 was achieved with

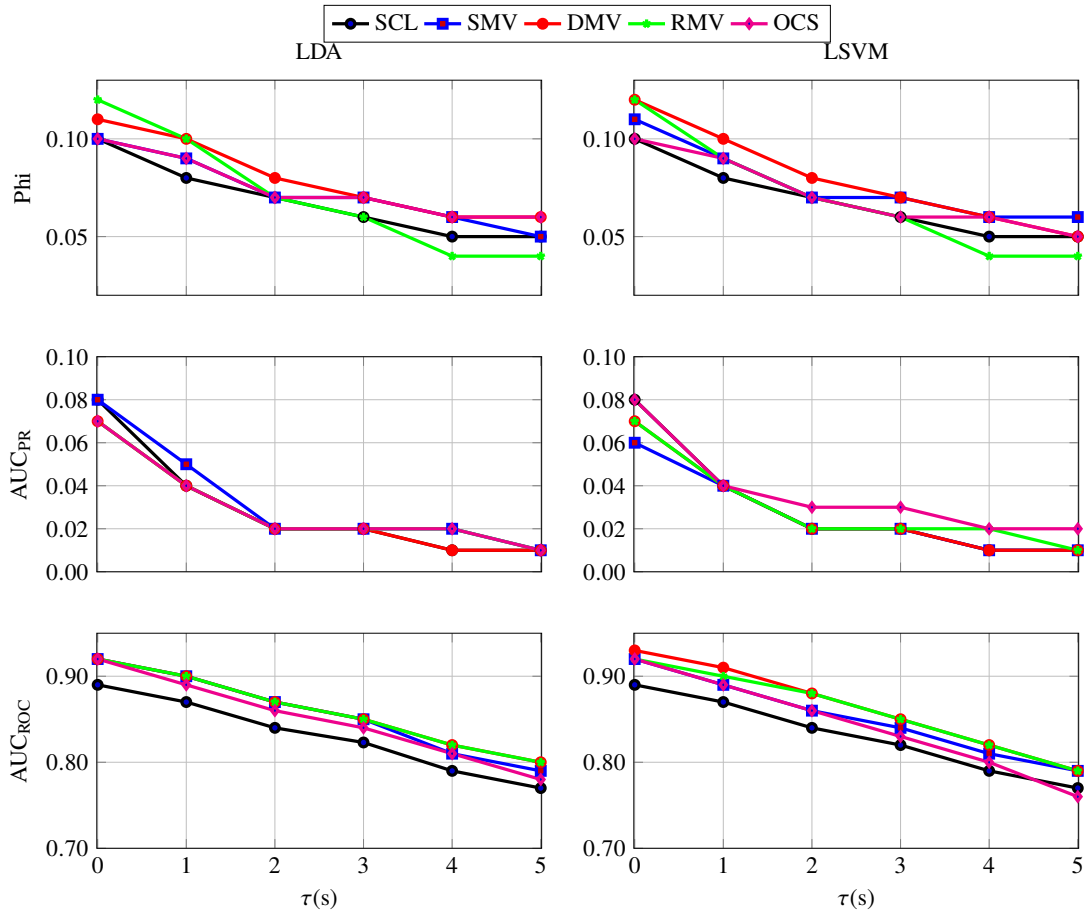

Figure C.5 Mean performances at predicting microsleep onsets for $\tau = 0 - 5$ s, using joint entropy features with classifier-independent feature selection and both LDA and LSVM classifiers in single and ensemble configurations. Single classifier, standard majority voting, and overlapping clusters were abbreviated as SCL, SMV, and OCS, respectively.

majority voting. Mean detection AUC_{PR}s of a single classifier and overlapping clusters were the same, i.e., 0.08, whereas, mean prediction AUC_{PR} of 0.02 was achieved with overlapping clusters. The highest mean detection AUC_{ROC} of 0.93 was achieved with DMV, whereas, except for overlapping clusters, all of the ensemble techniques gave the same mean prediction AUC_{ROC}s of 0.79.