Module:	7PAVREPR Research Project (2021/2022)
Candidate no.:	AC16635
Title:	Project Proposal
Project topic:	Predicting depression from wearable-derived sleep data: subject-specific approach with domain knowledge-driven and learned features
Date of submission:	4th February 2022
Supervisors:	Dr. Shaoxiong Sun (primary supervisor), Prof. Richard Dobson, Dr. Amos Folarin

Title:
Predicting depression from wearable-derived sleep data: Subject-specific approach with domain knowledge-driven and learned features

1. Introduction

Major depressive disorder (MDD) is a major public health burden affecting more than 300 million people worldwide.[1] Its symptoms show high variability over time. As a significant proportion of people with MMD undergo relapse, the continuous monitoring and early detection of disease trajectory and relapse are under active research. Despite high heterogeneity in symptoms and severity, sleep disturbance is a common symptom reported by over 90% of depressive patients.[2]

Changes in sleep were shown to exhibit a bidirectional relationship with MDD by several longitudinal studies.[2,3] Sleep disturbance including hypersomnia and insomnia is a comorbidity to depression that can be predicted by worsening depressive symptoms, but is concurrently a prodromal syndrome to depression.[4] On the other hand, MDD onset had predictive value to self-reported incidence, persistence and worsening of sleep disturbances.[5]

Sleep could be assessed subjectively, by which self-reporting may introduce bias, or objectively with the gold standard being polysomnography (PSG), also known as sleep electroencephalogram (EEG).[6] The PSG is greatly limited by the unnatural setting as well as time and financial costs. Thus, there is increasing interest in remote monitoring technology (RMT) for its potential for objective but passive measurement of physiological and behavioural characteristics. Consumer wearables, owing to their non-intrusiveness, are designed to be worn by the user regularly for an extended time, thereby enabling the long-term collection of data of substantial volume. The Fitbit wristband, for example, estimates sleep stages at 30-second epochs.

Wearables-derived sleep characteristics were shown to be related to depression. Initial findings show that sleep features were selected among other remote sensing variables in predicting depression.[7] The study was limited by the small number of subjects or sleeping hours measured, but these findings were recently supported by a large observational study. Features engineered from Fitbit sleep data in the areas of sleep architecture, sleep quality, sleep stability, hypersomnia and insomnia, were shown to be related to disease severity.[2] The potential for Fitbit sleep data to classify the depressive statuses of individuals at a specific time point was also established, in which features were mined by an automated algorithm and then modelled with machine learning and neural networks.[8] However, it must be noted that the discriminative performance is limited.

These studies have their limitations. In the data pre-processing steps, data missingness was reported but its informativeness on depressive status is poorly understood, with the common approach being discarding periods of sleep data where the overall within-period coverage is low or when there are days with extremely low coverage.[2,7] Secondly, the resolution of the data was reduced when features were summarised as aggregates at the night or hour level. Information on neither the cyclical nature nor transitions between sleep stages could be

captured, and their potential association with depressive status was not investigated. Furthermore, current predictive models only consider the general effect but fail to account for between-subject differences and within-subject dependencies.

This work proposed to extend the research on sleep stage data derived from the Fitbit wristband, a consumer wearable, to classify the depressive status of individual patients. The main aim of the proposed study is to design biomarkers and a machine learning approach for individual predictions. Specifically, the research questions are as follows:

(i) What are the key features in terms of predictive ability — those summarised on clinical grounds, generated by automated means or a combination of them? What are their respective discriminative powers, and do they provide complementary predictive value to each other?

(ii) Would a subject-specific model outperform a generalised model?

2. Methodological Considerations

2.1 Data source

Data have been collected from the Remote Assessment of Disease and Relapse – Major Depressive Disorder (RADAR-MDD) project, which is a multi-centre, prospective observational cohort study with the aims of exploring the potential of RMT in depression monitoring.[9] It recruited 623 individuals with a history of MDD from three sites in London, Amsterdam and Barcelona respectively, from 30th November 2017 to 3rd June 2020. Subjects were then followed up for a maximum period of 2 years or until the end of data collection in April 2021. This work's focus is data collected during the pre-pandemic period due to reasons to be discussed in section 2.3.

Two data streams are proposed to be sourced. Sleep records were passively generated by a Fitbit Charge 2/3 wristband, which was given to and required to be worn by each participant. Depressive scores were collected through subject self-reporting via the 'active RMT' mobile application developed for RADAR-MDD. Participants were required to complete the 8-item Patient Health Questionnaire (PHQ-8) every 14 days. Current depression was defined to be scoring ≥10.[10]

Data request would be necessary but not a research passport.

2.2 Data reliability

The data validity of the sleep stage data from consumer wearables has been discussed in the literature. Studies comparing Fitbit-derived labels to ground truth confirmed the correctness of sleep-wake times but showed limitations in the devices' sensitivity and specificity in sleep stage discrimination.[9,11,12] Therefore, this proposed work aims to compare the predictive ability of sleep-wake times, binned aggregates of sleep stage information and the higher-resolution time dynamics. Feature generation from sleep records would be detailed in later sections.

2.3 Data pre-processing

The data collection period overlapped with the COVID-19 pandemic. Therefore, only data collected before the pandemic would be subsetted, with the cut-off date would be set to 30th January when the WHO issued Global Health Emergency.[13] The effects of the pandemic on sleep behaviour and mood is still under active research, but there is evidence that in the RADAR-MDD dataset alone there are noticeable changes to sleeping behaviour in all three sites. For example in total sleep duration, bedtime before and after social distancing measures,[14] with statistically significant changes in mean sleep duration between during- and post-lockdown that interacts with depression.[15]

As the PHQ-8 questionnaire asks about symptom severity for the last 14 days, the sleep records 14-days for the participant before questionnaire completion would be matched to the depressive score during data pre-processing.

2.4 Feature extraction

Table 1 described the features to be extracted. Two main collections of features would be generated according to two rationales from the literature. To attempt to capture more information, this work also proposes additional feature generation directions.

Perspective	Features	Reference
Driven by domain knowledge	● 18 summary statistics across each night's sleep	Zhang et al.[2]
Automated feature learning of hourly sleep stage ratios	● Features extracted by tsfresh package	Gao[8]
Capturing daytime sleep	● Total duration of daytime sleep ● Number of episodes of daytime sleep ● Mean episode duration of daytime sleep	Proposed in this work
Encoding missingness	● Number of days with >50% missing data ● Proportion of missing per night	Proposed in this work
Unsupervised feature learning at finer resolution	● Lower-dimensional vector from convolutional autoencoder	Proposed in this work

Table 1. Features to be extracted from every 2-week period of sleep records

The first existing feature set is based on clinical grounds by Zhang et al., in which 18 features are summarised for each night's sleep in five areas including sleep architecture, sleep quality, sleep stability, insomnia and hypersomnia.[2] Secondly, Gao expanded features from multivariate time series of hourly sleep stage ratios using an automated algorithm (from the `tsfresh` package).[8]

This work proposed 3 new feature extraction perspectives. Firstly, sleep behaviour out of 'normal' sleeping time (that was arbitrarily defined by the researcher) was not considered previously. Therefore, features on the daytime sleep behaviour would be engineered, including the total duration, number of episodes and mean episode duration.

Furthermore, missingness is hypothesised to be informative in predicting depression symptom severity. In RADAR-MDD, the average participant wear-time of the Fitbit device across the entire follow-up duration was 62.5% (σ=9.1%) and thus missingness for sleep records is expected. It is proposed that, within each period, the number of days with more than 50% missing data (used in the literature as cut-off discarding periods of data),[7] as well as the average and variance of the proportion of missing data per night would be computed as input features.

Another proposed perspective is unsupervised feature learning that attempts to capture sleep stage information at a resolution finer than hourly aggregates. A convolutional autoencoder would be used (from the `PyTorch` package), as supported by evidence for its potential to capture variations by daily and seasonal variations in yearly profiles.[16,17] Similarly, high-dimensional sleep data from each period (consisting of records from multiple nights) may possibly be represented by a lower-dimensional vector, which could be used as inputs to classifiers.

2.5 Feature selection

Constant or quasi-constant features would be discarded. Correlation feature selection (CFS) would be employed, which is a filter approach that aims to find a feature subset with high feature-target correlation and low feature-feature correlation.[18]

2.6 Prediction models and performance checking

This proposed work attempted to make two sets of comparisons. The first one compares the performance of different models, where they would be initially run on individual sets of features, and then on the combined sets. Models include logistic regression and random forest from the `sklearn` package and the XGBoost model from the `xgboost` package. XGBoost is a highly effective and relatively fast algorithm that was found to outperform long short-term memory recurrent neural network (LSTM RNN) in this dataset.[8]

The second comparison is between subject-specific and across-subject effects, in consideration of the clustered nature of the data consisting of repeated measurements from subjects. The feature set with the optimal performance from the first comparison would be used. Multilevel models include mixed effects logistic regression and 'mixed effects random forest for clustered data' from the `merf` package.[19,20] Gaussian process boosting from the `GPBoost` package, which combines boosting and mixed effects modelling, would be used to compare results with that from XGboost.[21]

Model performance would be assessed by the F1 score, sensitivity, specificity, and the area under the receiver operating curve (AUROC).[22,23]

2.7 Software and computing resources

Data pre-processing, cleaning, feature extraction, feature selection, model fitting and model performance checking would be conducted using Python 3. Analysis would be conducted on the student's laptop, and on the research group's GPU resources if necessary.

3. Timetable

Time period	*Tasks*
Nov – Jan	● Get guidance on relevant background literature and further details of study materials from supervisor ● Request for data requisition
Jan	● Proposal first draft ● Data quality check and cleaning
Feb – Mar	● Data cleaning ● Write-up of introduction and literature review
Apr	● Data preprocessing and feature extraction
May	● Model fitting ● Write-up of methods
Jun – Aug	● Model fitting and refine feature extraction ● Write-up of results and discussion
Aug	● Project report first draft ● Poster creation ● Poster presentation
Sep	● Project report revision ● Project report submission

References

1. The World Health Organization. Depression. https://www.who.int/news-room/fact-sheets/detail/depression (2021).
2. Zhang, Y. *et al.* Relationship Between Major Depression Symptom Severity and Sleep Collected Using a Wristband Wearable Device: Multicenter Longitudinal Observational Study. *JMIR MHealth UHealth* **9**, e24604 (2021).
3. Bao, Y.-P. *et al.* Cooccurrence and bidirectional prediction of sleep disturbances and depression in older adults: Meta-analysis and systematic review. *Neurosci. Biobehav. Rev.* **75**, 257–273 (2017).
4. Fang, H., Tu, S., Sheng, J. & Shao, A. Depression in sleep disturbance: A review on a bidirectional relationship, mechanisms and treatment. *J. Cell. Mol. Med.* **23**, 2324–2332 (2019).
5. Steiger, A. & Pawlowski, M. Depression and Sleep. *Int. J. Mol. Sci.* **20**, 607 (2019).
6. de Zambotti, M., Goldstone, A., Claudatos, S., Colrain, I. M. & Baker, F. C. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiol. Int.* **35**, 465–476 (2018).
7. Lu, J. *et al.* Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2**, 1–21 (2018).
8. Gao, H. A deep learning approach to Infer depressive status from data collected through wearable devices. (King's College London, 2021).
9. Matcham, F. *et al.* Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): Recruitment, retention, and data availability in a longitudinal remote measurement study. (2021) doi:10.21203/rs.3.rs-612374/v1.
10. Kroenke, K. *et al.* The PHQ-8 as a measure of current depression in the general population. *J. Affect. Disord.* **114**, 163–173 (2009).
11. Haghayegh, S., Khoshnevis, S., Smolensky, M. H., Diller, K. R. & Castriotta, R. J. Accuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and

Meta-Analysis. *J. Med. Internet Res.* **21**, e16273 (2019).

12. Haghayegh, S., Khoshnevis, S., Smolensky, M. H., Diller, K. R. & Castriotta, R. J. Performance assessment of new-generation Fitbit technology in deriving sleep parameters and stages. *Chronobiol. Int.* **37**, 47–59 (2020).

13. Sohrabi, C. *et al.* World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *Int. J. Surg. Lond. Engl.* **76**, 71–76 (2020).

14. Sun, S. *et al.* Using Smartphones and Wearable Devices to Monitor Behavioral Changes During COVID-19. *J. Med. Internet Res.* **22**, e19992 (2020).

15. Leightley, D. *et al.* Investigating the impact of COVID-19 lockdown on adults with a recent history of recurrent major depressive disorder: a multi-Centre study using remote measurement technology. *BMC Psychiatry* **21**, 435 (2021).

16. Ryu, S., Choi, H., Lee, H. & Kim, H. Convolutional Autoencoder Based Feature Extraction and Clustering for Customer Load Analysis. *IEEE Trans. Power Syst.* **35**, 1048–1060 (2020).

17. Fitbeat: COVID-19 estimation based on wristband heart rate using a contrastive convolutional auto-encoder - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0031320321005793.

18. Hall, M. A. Correlation-based Feature Selection for Machine Learning. (The University of Waikato, 1999).

19. Hajjem, A., Bellavance, F. & Larocque, D. Mixed-effects random forest for clustered data. *J. Stat. Comput. Simul.* **84**, 1313–1328 (2014).

20. Wang, P. & Puterman, M. L. Mixed Logistic Regression Models. *J. Agric. Biol. Environ. Stat.* **3**, 175 (1998).

21. Sigrist, F. Gaussian Process Boosting. *ArXiv200402653 Cs Stat* (2021).

22. Zijdenbos, A. P., Dawant, B. M., Margolin, R. A. & Palmer, A. C. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans. Med. Imaging* **13**, 716–724 (1994).

23. Hanley, J. A. & McNeil, B. J. A method of comparing the areas under receiver

operating characteristic curves derived from the same cases. *Radiology* 839–843 (1983).