**Title**

Predicting major depression symptom severity from sleep stages collected using a wristband wearable device with machine learning

**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.3–5 Sleep disturbance is a comorbidity to depression that can be predicted by worsening depressive symptoms, but it is simultaneously a prodromal syndrome to depression. There was evidence that the MDD onset offers predictive value to self-reported incidence, persistence and worsening of sleep disturbances.6

Contrary to subjective reporting which may introduce bias, sleep could also be assessed objectively with the gold standard being polysomnography (PSG), also known as sleep electroencephalogram (EEG).7 However, they are 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 passive and objective 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.

Sleep characteristics measured objectively were shown to be related to depression. Variables readily available from the Fitbit wristband were depressive scores.8 These findings were limited by the small number of subjects or the short sleep duration recorded but 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.7 The potential for Fitbit sleep data to classify the depressive statuses of individuals at a specific time point was also established, with features mined by a data-driven approach and then modelled with machine learning and neural network approaches. However, it must be noted that the discriminative performance is limited.9

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.2,8 Secondly, the resolution of the data was reduced when features are 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. Furthermore, current predictive models only consider the general effect but fail to account for between-subject differences.

This work proposed to extend the investigation on sleep stage data derived from consumer wearables (Fitbit wristband) to classify depression symptom severity of individual patients at a specific time. 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:

1. Which are the key features in the predictive ability — features summarised on clinical grounds, time series features engineered by data-driven means or a combination of them? What are their respective discriminative powers, and do they provide complementary predictive value to each other?
2. Would subject-specific outperform the generalised model?

|  |
| --- |
| 1. *How do subject-specific effects compared to across-subject effects affect the predictive performance?* 2. *May the incorporation of within-night temporal dynamics at higher resolution improve predictive ability?* 3. *Which are the key features in the predictive ability — features summarized on clinical grounds, time series features engineered by data-driven means, or a combination of them? Are some of them potential markers for the early detection of depression relapse?* |

**Methodology**

*2.1 Data source*

Data would be collected from the Remote Assessment of Disease and Relapse – Major Depressive Disorder (RADAR-MDD) study, which is a multi-centre, prospective observational cohort study published in 2021 with the aims of exploring the potential of RMT in depression monitoring.10 In this study, 623 individuals with a history of MDD were recruited in 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.

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 symptom severity was 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.

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.12,13 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 to avoid its effects on sleep behaviour and mood, which are still incompletely understood. The cut-off date would be set to 30th January when the WHO issued Global Health Emergency.11

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.

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

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

Furthermore, missingness is hypothesized 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, as well as the average and variance of the proportion of missing data per night would be computed as input features.

Another proposed approach is unsupervised feature learning that attempts to capture sleep stage information at a resolution finer than hourly aggregates. A convolutional autoencoder would be used, as supported by evidence for its potential to capture variations by daily and seasonal variations in yearly profiles.12 Similarly, high-dimensional data from each period (consisting of records from multiple nights) may be represented by a lower-dimensional vector.

*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.13

*2.6 Prediction models and performance checking*

This proposed work attempted to make two sets of comparisons. The first one compares the discriminative values of the different collections of features generated from different approaches respectively. The models would be run on individual sets of features, and then on the combined sets. The choices of models are described as follows. Baseline models would include a regression-based and a tree-based model, namely logistic regression and random forest (from the sklearn package). XGBoost model (from the xgboost package) would be employed, which is a highly effective and relatively fast algorithm that was found to outperform long short-term memory recurrence neural network (LSTM RNN) in this dataset.9

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. Mixed effects models including logistic regression, along with ‘mixed effects random forest for clustered data’ (from the merf package).14 Gaussian process boosting (from the GPBoost package), which combines boosting and mixed effects modelling, would be used to compare results with that from XGboost.15

Model performance would be assessed by the F1 score, sensitivity, specificity, and the area under the receiver operating curve (AUROC).16,17

*2.7 Software*

Data pre-processing, cleaning, feature extraction, feature selection, model fitting and model performance checking would be conducted using Python.

|  |  |  |
| --- | --- | --- |
| *Approach* | *Features* | *Reference* |
| Driven by domain knowledge | • 18 summary statistics across each night’s sleep | Zhang et al.2 |
| Unsupervised feature learning of hourly sleep stage ratios | • Features extracted by tsfresh package | Haotian9 |
| 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 | • Features extracted by convolutional autoencoder | Proposed in this work |

*Table 1. Features to be extracted from each 14-day period of sleep records*

**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. Zhang, J.-Y. *et al.* Single-cell landscape of immunological responses in patients with COVID-19. *Nat. Immunol.* **21**, 1107–1118 (2020).

6. Steiger, A. & Pawlowski, M. Depression and Sleep. *Int. J. Mol. Sci.* **20**, 607 (2019).

7. Zhang, Y. *et al.* Predicting Depressive Symptom Severity through Individuals’ Nearby Bluetooth Devices Count Data Collected by Mobile Phones: A Preliminary Longitudinal Study. *ArXiv210412407 Cs Stat* (2021).

8. 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).

9. Gao, H. A deep learning approach to Infer depressive status from data collected through wearable devices. (King’s College London, 2021).

10. 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.* https://www.researchsquare.com/article/rs-612374/v1 (2021) doi:10.21203/rs.3.rs-612374/v1.

11. 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).

12. 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).

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

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

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

16. 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).

17. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. | Radiology. https://pubs.rsna.org/doi/abs/10.1148/radiology.148.3.6878708.

**Timetable**

|  |  |
| --- | --- |
| *Time period* | *Targets* |
| November – January | -        Get guidance on relevant background literature and further details of study materials from supervisor  - Request for data requisition |
| Jan | - Proposal first draft |
|  | -        Data cleaning |
|  | -        Project proposal first draft |
| Feb | -        Submit |
|  | -        Data cleaning  -        Write-up of introduction and literature review |
| Mar – May | - Write-up of methods  -        Feature extraction and model fitting |
| Jun – Aug | - Final write-up of results and discussion  - Model fitting and refine feature extraction |
| Aug | - Final write-up first draft |
|  | - Poster creation |
|  | - Poster presentation |
| Sep | - Final write-up revision |
|  | - Final write-up submission |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Experiment* | *Model* | *Features* | *Outcome* |  |
| 1 | Mixed random random forest | Features from Table 1 |  |  |
|  |  | Sleep stage ratio time series (hourly binned) |  |  |
| 2 | Random forest |  |  |  |
|  | Logistic regression with random effects |  |  |  |
|  | Logistic regression without random effects |  |  |  |
|  |  |  |  |  |

*Table 2. Models to be adopted*