Artificial Intelligence to Aid in Diagnosis of Type I Narcolepsy

Yoav N. Nygate, MS1, Sam Rusk, BS1, Chris R. Fernandez, MS1, Zac Winzurk, BS1, Emerson M. Wickwire, PhD2, Emmanuel Mignot, MD, PhD3, Nathaniel F. Watson, MD, MS4

1 EnsoData Research, Ensodata, Madison, WI, USA
2 Department of Psychology, University of Maryland, Baltimore, MD
3 Center for Sleep Science and Medicine, Stanford University, Stanford, CA, USA
4 Department of Neurology, University of Washington School of Medicine, Seattle, WA

Introduction

What is Narcolepsy?

Narcolepsy is a chronic sleep disorder characterized by excessive daytime sleepiness, sleep paralysis, and sleep related hallucinations. Narcolepsy is divided into two types - narcolepsy type 1 (NT1) and narcolepsy type 2 (NT2). Compared to NT2, NT1 is further characterized by the presence of cataplexy - sudden muscle weakness triggered by emotion.

Prevalence of Narcolepsy:

In a US population that is actively using medical and pharmacy services, the prevalence of NT1 in 2016 was estimated at 44.3 per 100,000 and was observed to be higher in the Midwest at 58.2 per 100,000; however, it is believed that the disorder is substantially undiagnosed and the actual prevalence might be much higher [1]. The onset of narcolepsy symptoms range from 8 to 22 years and contains a bimodal distribution incidence with a peak at age 15 and an additional one in the mid-30s [2].

Diagnosis of Narcolepsy:

Multiple sleep latency tests (MSLT) are central to NT1 diagnosis; however, current medications, sleep schedule, drug use, and testing environment can all compromise the interpretation of an MSLT. Moreover, MSLT suffers from an imperfect sensitivity, requiring repetition when a high suspicion of narcolepsy is present. Furthermore, the test suffers from an imperfect specificity, resulting in false positives, often requiring interpretation in the clinical context [2]. Although being an integral part of the standard of care for NT1 diagnosis, the ordering frequency of MSLTs is declining, with a decline of 20% in frequency observed between the years 2013-2016 [1]. Incorrect MSLT interpretation has devastating consequences for patients, including prolonged suffering from untreated NT1 or adverse effects from unnecessary testing and therapies. An alternative and more reliable diagnostic test for NT1 is done by measuring CSF levels of hypocretin utilizing a lumbar puncture [3]. However, because of the risks and uncomfortable nature of this procedure, this test is seldomly performed in clinical practice.

Study Objective:

Due to the complex nature and various confounders of NT1 diagnosis, simple, accessible, accurate and cost-effective diagnostic solutions are needed. In this study, we explore a range of methods for the automated detection of NT1 based on single night polysomnography (PSG). We evaluate their performance on a publicly available dataset.

The Dataset

- Previously published dataset (MNC dataset) [4]:
  - Polysomnography (PSG) sleep studies collected from 6 different cohorts.
  - N=235 NT1 patients.
  - N=431 negative controls.
- Historical Database:
  - Historical database of over 1 million sleep studies
  - N=3,000 PSG sleep studies used as negative controls sampled from over 300 clinics across the U.S. and 10 different recording devices.
  - N=1,000 PSG sleep studies used for trainining a sleep staging model, sampled from 1 clinic using a specific recording device.
  - N=50,000 PSG sleep studies used for training a sleep staging model, sampled from over 300 clinics across the U.S. and 10 different recording devices.

Methodology

Models:

- Hypnodensity V1 Random Forest (Hypno-RF V1):
  - We trained a sleep staging model using the N=1,000 dataset.
  - We then extracted the hypnodensity [4] using the output probabilities of the trained sleep staging model.
  - We trained a random forest model using 400 hand engineered features (as defined in reference [4]) extracted from the hypnodensity.
- Hypnodenisty V2 Random Forest (Hypno-RF V2):
  - Same as Hypno-RF V1 except, the sleep staging model was trained using the N=50,000 dataset.
- PSG Sleep Report Data Random Forest (Sleep-RF):
  - Trained a random forest model using 15 different PSG-based and sleep-based report data calculated for each subject in the MNC dataset.
- PSG-EEG Based Deep Learning Model (PSG-DL):
  - Trained a deep learning model on raw EEG signals derived from both the MNC dataset, as well as the N=3,000 dataset.
- Ensemble Model:
  - Taking the mean of the probabilities produced by PSG-DL, Sleep-RF, and Hypno-RF V2 models.

Evaluation Methods:

- All models were trained using 10-fold cross-validation
- Performance was evaluated using area under the receiver operating characteristic curve (AUC-ROC), sensitivity, specificity, and F1-Score.
- When possible, we evaluated feature importance using the Gini Index.
- We also performed statistical analysis on the report data variables.

Results: Narcolepsy Type 1 Detection Performance

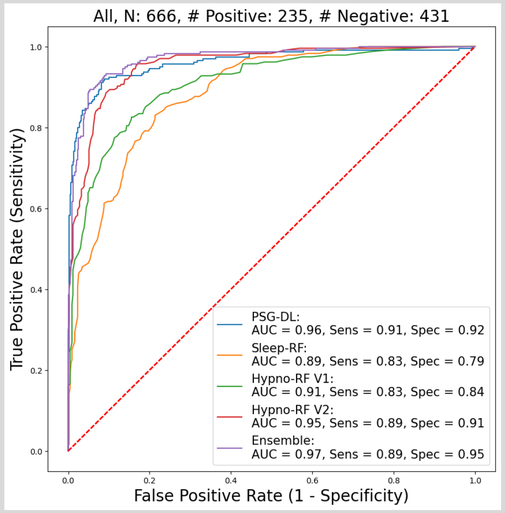

Figure 1. ROC curve comparing the performance of each model. Sensitivity and specificity were calculated by choosing the threshold that maximizes the F1-Score.

Results: Feature Importance Analysis for Sleep-RF

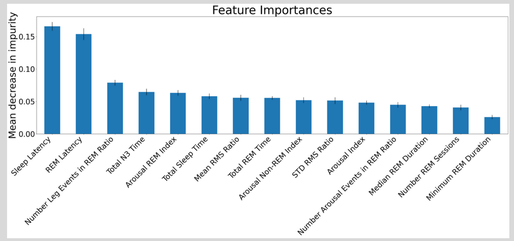

Figure 2. Feature importance analysis for the Sleep-RF model. The values in the y-axis represent the importance of each feature. The higher the mean decrease in impurity is, the more important the feature is to the overall prediction.

PSG Report Data Variables Statistical Analysis

Feature Name	Positive Population	Negative Population	OLS P-Value	OLS Coefficient	Gini
Sleep Latency	0.21 ± 0.604	0.883 ± 1.758	0.0	-0.183669	0.166
Total N3 Time	1.298 ± 1.684	1.044 ± 1.13	3E-05	0.119404	0.065
Median REM Duration	0.205 ± 0.272	0.255 ± 0.296	0.003938	-0.59487	0.043
Minimum REM Duration	0.056 ± 0.208	0.089 ± 0.246	0.014084	0.508941	0.026
REM Latency	1.347 ± 3.304	2.159 ± 2.98	0.04453	-0.026918	0.154
Arousal REM Index	21.896 ± 22.596	18.021 ± 24.7	0.159873	0.003634	0.063
Number REM Sessions	7.919 ± 8.758	5.543 ± 5.896	0.256839	0.008924	0.041
Arousal Non-REM Index	7.144 ± 11.8	4.665 ± 7.088	0.295553	0.017817	0.052
Total Sleep Time	7.31 ± 2.416	6.702 ± 2.298	0.687989	0.008871	0.058
Arousal Index	5.101 ± 6.53	3.594 ± 5.102	0.706944	-0.01157	0.048
Number Leg Events in REM Ratio	0.248 ± 0.402	0.13 ± 0.37	0.754354	0.033496	0.079
Mean RMS Ratio	0.697 ± 0.686	0.599 ± 0.588	0.784755	-0.019252	0.056
Total REM Time	1.748 ± 1.544	1.37 ± 1.11	0.787871	0.01743	0.055
STD RMS Ratio	1.024 ± 1.468	0.862 ± 1.136	0.830914	0.007382	0.051
Number Arousal Events in REM Ratio	0.291 ± 0.286	0.212 ± 0.254	0.945966	-0.014746	0.045

Table 1. Ordinary least squares (OLS) summary. We ran all variables through an OLS model where each time one variable was varied while all other variables were controlled for.

Discussion

ROC Observations:

- Hypno-RF V2 outperforms Hypno-RF V1:
  - The larger increase in sensitivity may suggest that the features used to train Hypno-RF V1 were limited in their capability to detect patients with NT1.
  - Training the sleep staging model using a large and diverse dataset generates better hypnodensity-based features.
  - When utilizing a lower performing sleep staging model, the overlap between sleep stage probabilities (which is hypothesized to be a main feature for the detection of NT1 from PSGs) might be attributed to model artifacts and in-accuracies, rather than narcoleptic behavior during sleep.
- PSG-DL model outperforms both Hypno-RF models:
  - This might be attributed to the fact that the PSG-DL model was trained using 3,000 controls sampled from various clinics and softwares, as well as trained using the raw PSG-EEG signals. This enabled the model to learn its own features rather than depend on hand-engineered ones. Overall, this resulted in a more specific, sensitive, and potentially more generalizable model,
- Ensemble model outperforms all other methods:
  - This highlights the potential of using a combination of multiple approaches to detect NT1.

Feature Importance Observations:

- Sleep latency and REM latency were the top two important features which is expected for NT1 detection.
- The percentage of the number of leg events that occurred in REM is the third most important feature, and the arousal REM index - number of arousal events per hour of REM sleep, is the fifth most important feature. This might suggest that the level of sleep fragmentation during REM varies between patients with NT1 and healthy controls.

Statistical Analysis Observations:

- Sleep latency, total N3 time, median REM duration, minimum REM duration, and REM latency all produced statistically significant differences between the positive and negative populations.
- Sleep latency and REM latency both had a negative association with NT1 patients - patients with NT1 have lower sleep and REM latencies, as expected.
- Total N3 time had a positive association with NT1 patients which might suggest that patients with narcolepsy tend to reach N3 sleep more often during the night.

Conclusions and Future Work

- ML methods automatically detected NT1 in PSG-EEG with promising degrees of accuracy.
- These methods overcome common barriers to accurate diagnosis that can compromise the interpretability of an MSLT.
- Broad implementation of this method has potential to supplement the MSLT, increasing diagnostic detection rates and accuracy for NT1.
- Access to care can be broadened if NT1 diagnosis were to occur anywhere a sleep EEG can be obtained, including the home setting, rather than being limited to in-center testing.
- This work has potential to further the developments of T1N detectors and assist in their widespread clinical adoption.
- Additional research is underway to help improve accuracy and ensure these methods are generalizable across platforms and clinical datasets.

References

1. Acquavella, John, et al. "Prevalence of narcolepsy and other sleep disorders and frequency of diagnostic tests from 2013–2016 in insured patients actively seeking care." Journal of Clinical Sleep Medicine 16.8 (2020): 1255-1263.

2. Golden, Erin C., and Melissa C. Lipford. "Narcolepsy: Diagnosis and management." Cleve Clin J Med 85.12 (2018): 959-969.

3. Sateia, Michael J. "International classification of sleep disorders." Chest 146.5 (2014): 1387-1394.

4. Stephansen, Jens B., et al. "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy." Nature communications 9.1 (2018): 5229.