

Better Diagnosing Narcolepsy

Cibele Halasz and Preeyanka Shah
 Department of Electrical Engineering
 Stanford University
 Stanford, CA 94305

Email: cibelemh@stanford.edu and psk3@stanford.edu



Abstract—Narcolepsy is a sleeping disorder that affects approximately 1 in 2000 people in the United States. [1] It is difficult to diagnose because its symptoms resemble those of other diseases and because signals collected during sleep studies such as EEGs, ECGs, EMGs have to be manually inspected by humans during the diagnosis process. In this project, we set out to determine what features can be used to classify narcolepsy based on hypnograms and what is the overall error rate based on a requirement of no greater than 5% Type I error. We determined that useful features that can be incorporated into a machine learning based model include the presence of sleep onset REM (SOREM), the percentage of transitions from stage 2 to REM sleep, awake to REM sleep, awake to NREM stage 2 as well as percentage of total sleep time in NREM stage 1 sleep. For the majority of the model development we used support vector machines for classification. Using 10-fold cross validation, we report 8.86% overall mean error and additionally Type I mean error of 3.00% and Type II mean error of 30.4%.

1 INTRODUCTION

Narcolepsy is a neurological disorder in which the brain is unable to regulate sleep-wake cycles normally. People with this condition experience disturbed nocturnal sleep and abnormal daytime sleeping patterns. Such symptoms manifest themselves as mental cloudiness, memory problems, lack of energy, depression, cataplexy and sleep paralysis. [1] Narcolepsy is a fairly rare disease and it is difficult to diagnose because its symptoms can be mistaken for symptoms of other diseases. [2]

Patients with suspected sleep disorders undergo sleep studies or nocturnal polysomnograms where they are monitored using several different types of signals. Information about a person's sleep behavior can be obtained by from recording brainwaves (electroencephalogram or EEG), eye movements (electrooculography or EOG), muscle tone (electromyography or EMG) and heart rate (ECG). [3] By visually scoring the EEG, EOG, and EMG, a hypnogram can be manually produced from these signals. Hypnograms are graphs that represent the stages of sleep as a function of time usually with 30 second precision. Certain frequencies displayed by EEGs, EOGs and EMGs are characteristic and determine

what stage of sleep or wake the subject is in.

Given that the most common method for diagnosing narcolepsy is through manual analysis of the sleep signals discussed above, it is a fairly labor intensive process. One question that has arose in the sleep disorder field is whether machine learning techniques can be used to automate the diagnosis process to some extent. For the scope of this project, we focused on examining whether hypnogram data can be used to better predict narcolepsy using classification algorithms.

2 REQUIREMENTS OF PROJECT

The main goal of this project is to develop an algorithm that predict whether a patient has narcolepsy based on a hypnogram collected during a sleep study. The specific requirements are as follows

- The features used in model fitting have a clinical basis
- The algorithm diagnoses narcolepsy with high specificity. There should be less than 5% false positives.
- Specificity should be prioritized over sensitivity. It is acceptable to miss narcoleptic patients.

3 DATASET

Our data set comprised of 137 narcoleptic patients and 512 control (non-narcoleptic) patients. The 137 narcoleptic patients are from 3 cohorts: a Stanford Sleep Study and 2 Xyrem (a drug that treats narcolepsy) studies. The 512 non-narcoleptic patients from the a Wisconsin Sleep study cohort.

All data was provided by the Stanford Center for Sleep Sciences and Medicine and was fully deidentified before being worked with.

Datasets	Data Size
Winsconsin(non-narcoleptic)	512
SXB22(narcoleptic)	5
SXB15(narcoleptic)	122
Stanford Clinic(narcoleptic)	11

The hypnogram signal traditionally has 30 second precision and 6 different stages which is seen in figure 1.

Sleep Stages	Value
Artifact(NaN)	-1
Awake	0
NREM1	1
NREM2	2
NREM3 (deep sleep)	3
REM	5

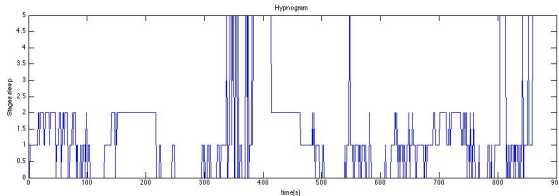


Fig. 1. Raw Hypnogram

3.1 Feature Development

One of the major areas of this project was feature development and selection. Given the ultimate goal of enabling physicians to use this work in clinically diagnosing narcolepsy, all features had to be understandable by clinicians.

The approach to feature development that we took was to extract ideas from papers discussing the impact of various sleep disorders on sleep patterns and converting them into features. [4] [5] [6]

In addition to sleep onset REM (SOREM) which is a binary signal denoting whether REM sleep occurred in the first 15 minutes of falling asleep, other features included transitions between various states, average duration of a REM stage in the first and second half of sleep and percentage of sleep time in each stage.

In order to better understand the importance of each feature we plotted histograms of the distribution of each feature as well as the distribution conditional on whether a patient had narcolepsy was present or not. In Figure 2, the specificity of SOREM in diagnosing narcolepsy is clear.

Based on conversations with Emmanuel Mignot, a leading narcolepsy researcher, to determine features other than SOREM, we removed the first 15 minutes of data.

Generally speaking, the features fell into several different categories:

- Total Sleep
- Presence of SOREM
- Percentage of time spent in a stage

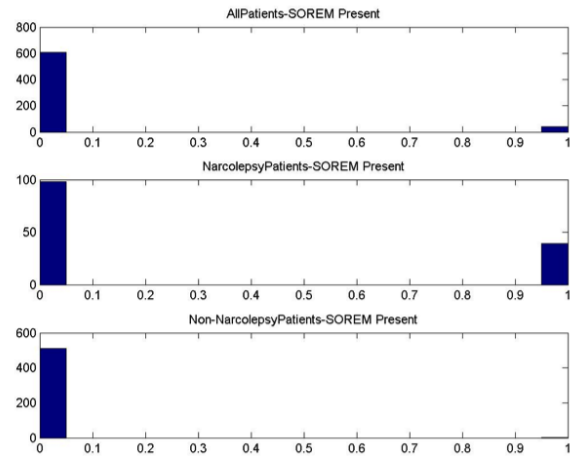


Fig. 2. Presence of SOREM in a patient.

- Number of transitions from one stage to another
- Length of a segment within a specific stage

One issue of consideration was how to clean up the data. In the process of analyzing the data, three outliers were removed from the data set.

Although a spike in stage for one or two cycles was likely to be noise, it is also known that narcoleptic patients display fragmented sleep including REM sleep. [7] Hence in the process of denoising, meaningful information could be removed.

Ultimately a second set of features was developed based on a mode-filtered signal. A 5-window mode filter was selected because it would remove 1 and 2 cycle outliers but keep longer term fluctuations that may represent information. A mode filter was selected over a mean or median filter is that each sleep stage can be considered a different category that are not perfectly ordered.

From that we were able to extract 42 features. They

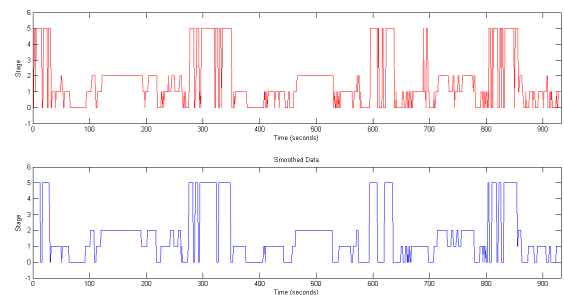


Fig. 3. Raw and Smoothed Hypnogram Waveform

account mainly for the number of transitions into REM, time spent in different stages of sleep, the percentage of time spent in those stages, the number of segments for

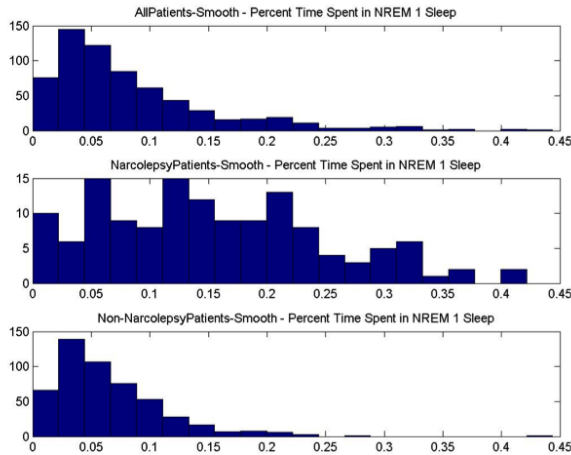


Fig. 4. Percentage of time spent in NREM sleep in smoothed waveform

each of the stages of sleep.

For training, all the features were normalized such that all entries ranged between 0 and 1.

4 METHODS

4.1 Machine Learning Algorithms Used

All development was done in MATLAB.

For most of training, testing and the majority of the time spent in model development, we used the LIBLINEAR library to implement L2-regularized L2-loss support vector classification within MATLAB. [8] This specific implementation minimizes the following objective

$$\min_{\alpha} \frac{1}{2} \left(\alpha^T \left(Q + \frac{I}{2 * C} \right) \alpha \right) - e^T \alpha \text{ s.t. } 0 \leq \alpha_i \quad (1)$$

The first thing that was done was determining appropriate weights for our criterion so that Type I errors would remain under 5% but that overall error would also be minimized.

Using the found weights, feature selection was done using forward sequential feature selection and using the SVM classification method described below. In forward sequential search, a single feature is used to predict the data. Then an additional feature is added to the model and whichever feature reduces the criterion the greatest is added to the model. This process is iterated until the criterion is no longer reduced by the addition of additional features.

After feature selection, we tested the accuracy of models based on SVM, Naive Bayes, Random Trees, Neural Network and logistic regression model using the validation process described below.

4.2 Feature Selection

To select the 5-6 best features from an original list of 40+ features, we used forward sequential search using SVM models with 10-fold cross-validation. We used the following criterion that needed to be further reduced:

$$C = w_1 * e_{typeI} + w_2 * e_{typeII}$$

We consider type I errors as incorrectly predicting a patient as narcoleptic when they are not and type II errors are predicting that a patient is not narcoleptic when in reality, they are. For this specific problem, reducing type I errors was more important.

4.3 Validation

To validate our model, we reported the average error as well as Type I and Type II error across 10-fold cross validation using the selected features. Given the small dataset, we did not leave out data to test on at this time. Although overfitting did exist, it was minimized as all model selection and validation was done using 10-fold cross validation. However, as the work is taken farther, we could test the model on unseen data.

5 RESULTS AND DISCUSSION

In the Methods section of the paper, we discussed using a weighted sum of errors in order to fit the model. The process of finding optimal weights was important in order to ensure that Type I errors remained below 5% during the process of reducing overall errors. In this scenario we found that weighing each type of error equally provided the optimal outcome which is clear in figure 5. However, if the ratios of narcoleptics and non-narcoleptics in the original data set were different, the weights may have needed to be tuned differently.

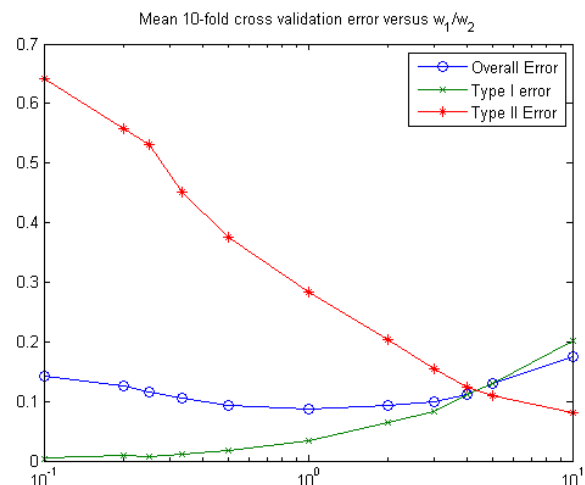


Fig. 5. Error rates versus ratio of w1 versus w2

Using forward sequential feature search, the 6 features that were found to be most relevant in predicting

narcolepsy with high specificity were the following¹. In figure 6, we display the decrease in error during each iteration of the forward sequential feature search. The features below are presented in order of being added to the model.

- SOREM Present
- Percent Transitions from stage 2 to REM
- Percent transitions from stage 0 to REM in the smoothed signal
- Percent Transitions from stage 0 to REM
- Percent Time Spent in NREM 1 Sleep
- Percent Transitions from stage 0 to 2

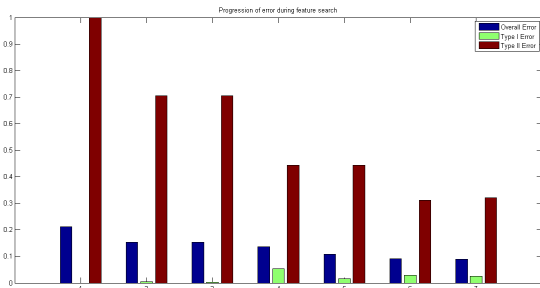


Fig. 6. Feature Search Error - L2-regularized L2-loss SVC

To determine whether the numbers reported were valid and consistent based on the an individual index selection used for k-means cross validation, we ran the algorithm 10 times to better determine the stability of the error values. With the L2-regularized L2-loss support vector machine classification model, we reported a mean error of 8.86% with a standard deviation of 0.51%.

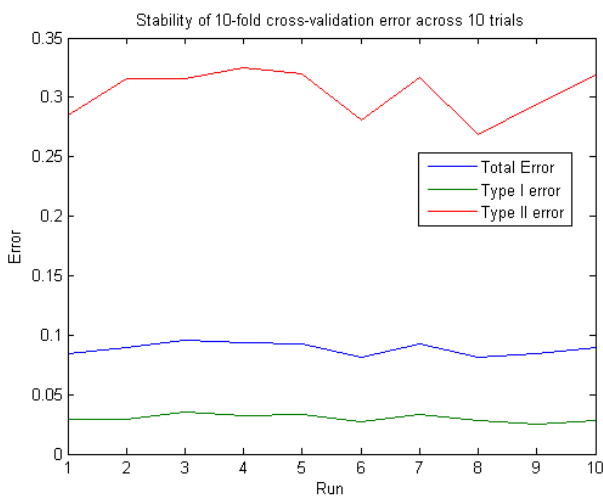


Fig. 7. Stability of 10-fold cross-validation error across ten trials

1. Percent transitions are taken over total number of transitions in hypnogram

5.1 Assessing Different Machine Learning Algorithms

In addition to using a model trained using support vector machines, we also explored other machine learning algorithms. The results for different algorithms are summarized below:

Algorithm	% Type I Error	% Type II Error	% Overall Error
SVM	3.00	30.40	8.86
Naive Bayes	7.77	33.33	13.07
Neural Net.	9.22	37.90	15.87
Rand. Forests	6.53	40.25	14.64
Log. Regress.	2.45	35.67	9.62

Initial trials had suggested that an SVM based model would have the lowest error. After selecting features, we were able to validate that assumption. Although logistic regression shows a lower type II error, its average overall error is slightly higher. However, given that it is significantly easier and faster to implement, it is a viable alternative.

In the future, we plan to test the model against unseen data. However, we foresee limitations in the model in that we may have been working with biased data. All of our controls came from a single cohort. As can be expected, many of them likely had another sleep disorder and furthermore it is known that the Wisconsin Sleep Study cohort is not diverse, age wise. Hypnograms can change significantly with age and hence, it will remain to be seen whether our model can be generalized and scalable. However, we have high hopes that at least some of the features will provide meaningful information in predicting narcolepsy.

It is important to mention that hypnograms are hand-scored by a technician through the analysis of the EEG, EOG, and EMG signals. This means that although the hypnogram requires less preprocessing, it is susceptible to human error.

6 CONCLUSION AND FURTHER WORK

Preliminary testing suggests that we have achieved our goal of using machine learning classification for predicting narcolepsy with high specificity.

The SVM classification based model shows the most promise in correctly classifying narcolepsy within clinical studies. For this method, we were able to achieve a 3.00% type I error which is within the specifications of the Stanford Center for Sleep Sciences and Medicine and an overall error rate of 8.86%. Although there is some over-fitting, these are issues that can be resolved once we have validated our model on new, more diverse clinical data. Nevertheless, our project was able to achieve the goals it had set out on and provides a solid basis for further research in

automating diagnosis of narcolepsy.

ACKNOWLEDGMENT

We would like to acknowledge Emil Munk and Emmanuel Mignot of the Mignot Lab at the Stanford Center for Sleep Sciences and Medicine.

REFERENCES

- [1] What are the signs and symptoms of narcolepsy. [Online]. Available: <http://www.nhlbi.nih.gov/health/health-topics/topics/nar/signs.html>
- [2] How is narcolepsy diagnosed. [Online]. Available: <http://www.nhlbi.nih.gov/health/health-topics/topics/nar/diagnosis.html>
- [3] Overnight, attended sleep study. [Online]. Available: <http://stanfordhospital.org/clinicsmedServices/clinics/sleep/test/overnight-attended-study.html>
- [4] P. Drakatos, A. Suri, S. E. Higgins, I. O. Ebrahim, R. T. Muza, C. A. Kosky, A. J. Williams, and G. D. Leschziner, "Sleep stage sequence analysis of sleep onset rem periods in the hypersomnias," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 84, no. 2, pp. 223–227, 2013.
- [5] J. Harsh, J. Peszka, G. Hartwig, and M. Mitler, "Night-time sleep and daytime sleepiness in narcolepsy," *Journal of sleep research*, vol. 9, no. 3, p. 309, 2000.
- [6] I. Marti, P. O. Valko, R. Khatami, C. L. Bassetti, and C. R. Baumann, "Multiple sleep latency measures in narcolepsy and behaviourally induced insufficient sleep syndrome," *Sleep medicine*, vol. 10, no. 10, pp. 1146–1150, 2009.
- [7] C. H. Schenck and M. W. Mahowald, "Motor dyscontrol in narcolepsy: Rapid-eye-movement (rem) sleep without atonia and rem sleep behavior disorder," *Annals of neurology*, vol. 32, no. 1, pp. 3–10, 1992.
- [8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.