UC Irvine
UC Irvine Previously Published Works

Title

Categorizing Sleep in Older Adults with Wireless Activity Monitors Using LSTM Neural Networks

Permalink

https://escholarship.org/uc/item/0df7c8cs

ISBN

9781538613115

Authors

Yildiz, Selda
Opel, Ryan A
Elliott, Jonathan E
et al.

Publication Date

2019-07-01

DOI

10.1109/embc.2019.8857453

Peer reviewed

Categorizing Sleep in Older Adults with Wireless Activity Monitors Using LSTM Neural Networks

Selda Yildiz,

Oregon Health and Science University, Portland, OR 97239 and the VA Portland Health Care System, Portland, OR 97239 USA

Ryan A. Opel,

VA Portland Health Care System, Portland, OR 97239 USA

Jonathan E. Elliott,

VA Portland Health Care System, Portland, OR 97239 and Oregon Health and Science University, Portland, OR 97239 USA

Jeffrey Kaye [Member, IEEE],

Oregon Health and Science University, Portland, OR 97239 USA

Hung Cao [Senior Member, IEEE],

School of Engineering, University of California, Irvine, CA, 92679 USA

Miranda M. Lim

VA Portland Health Care System, Portland, OR 97239 and Oregon Health and Science University, Portland, OR 97239 USA

Abstract

Novel approaches are needed to accurately classify and monitor sleep patterns in older adults, particularly those with cognitive impairment and non-normative sleep. Traditional methods ignore underlying sleep architecture in these patient populations, and other modern approaches tend to focus on healthy, normative patient populations. In this paper, we developed a model using a long-short-term memory neural network (LSTM) and trained it on a sample of older, non-normative patients. The 22 nights of data collected were trained on gold-standard polysomnography (PSG) as ground truth and were compared against the clinical standard threshold-based method for sleep detection. The LSTM more than doubled the traditional method's ability to detect clinically-relevant wakefulness during sleep (37.7% vs. 15%) without significantly sacrificing accuracy (67.7% vs. 75%) or precision (90.7% vs. 94%) of sleep classification.

Keywords

yildiz@ohsu.edu.

I. Introduction

Sleep disturbances may be one of the earliest clinical signs of neuropathology in dementia [1]. Emerging literature shows strong evidence for a bidirectional relationship between sleep and dementia. Indeed, an important biological function of sleep may be to facilitate clearance of neuronal wastes from the brain parenchyma (e.g., the "glymphatic function" of cerebrospinal fluid and brain interstitial fluid exchange) [2]. Early screening and detection of a change in sleep patterns may present a valuable window of opportunity during which potential interventions could be tested [1]. However, detecting sleep impairment is challenging, as the gold standard method of overnight polysomnography (PSG) is time-limited, cumbersome, and expensive. At best, PSG provides a random snapshot of one or a handful of nights of sleep in the laboratory. Obtaining overnight PSG on older, cognitively impaired subjects is particularly burdensome and challenging.

Wearable devices are one possible solution to the time-limited vignette provided by PSG. Actigraphy devices – wearable accelerometers embedded into a wristwatch – allow long-term data collection in the home, thus more accurately capturing subjects' sleep patterns. Actigraphy is relatively unobtrusive and well tolerated, even in the older, cognitively impaired population. Actigraphy has been validated against PSG [3], [4] with high accuracy for distinguishing sleep versus wake in certain healthy/normative populations [5], [6], and in populations with sleep disorders including most recently in obstructive sleep apnea patients [7]-[10]. However, actigraphy has two major short-comings: i) sleep staging challenges to interpret in aged and/or cognitively impaired subjects who have irregular sleep and circadian rhythms, and ii) difficulty categorizing wake -with frequently over-estimating the quantity of sleep- when subjects lay still in bed [11]. Thus, novel approaches are needed in this regard.

Traditional threshold-based actigraphy scoring algorithms, including Cole-Kripke [12] and its derivatives, are problematic in non-normative populations due to their minimal assumptions about the underlying transition probabilities between sleep stages, acting as band-pass filtered, but otherwise direct, translations of raw activity data. Due to the tendency to simplify or smooth-over, these algorithms create underlying biases in sleep staging for older and cognitively impaired individuals, especially those with sleep problems.

Several groups [13]-[15] have attempted to stage sleep using a logistic regression model that considers multiple derived activity parameters in a moving window (e.g. moving average, standard deviation of the window). While these implementations improve upon threshold-based algorithms, problems still exist even in young and healthy subjects. All logistic regression models report low specificity (~50%) for detecting wakefulness [15], thus underestimating the amount of true wakefulness and skewing clinically-relevant sleep metrics, including total sleep time and sleep efficiency.

More recently, Fang *et al.* compared the ability of a variety of machine learning models to predict sleep using passive smart home sensors, and illustrated the benefits of a neural network over a hidden Markov model (HMM) or conditional random field (CRF) for the task [16]. While HMMs and CRFs address the underlying transition and emission

probabilities problems inherent in previous methods, both approaches require knowledge of prior probabilities, leading to decreased accuracy and sensitivity.

To address these problems and to score sleep, we proposed employing a neural network machine learning algorithm: a Long Short-Term Memory model (LSTM) [17], which is particularly well suited for longitudinal data. Neural networks have numerous advantages over both traditional methods, including the ability to both dynamically adapt to training data and detect complex, non-linear relationships between independent and dependent variables [18], and HMMs and CRFs, including utilizing backpropagation.

II.   Methods

A.   Actiwatch

The Philips Actiwatch 2 (Philips Respironics, Bend, Oregon) is a wrist-worn activity monitoring device that has been validated against PSG in normative populations [5]. The device uses a solid-state piezoelectric accelerometer to record activity, sampling at 32 Hz and band-pass filtered from 0.35 – 7.5 Hz. The accelerometer's sensitivity is approximately 0.025 G with a range of 0.5 – 2 G peak. The Actiwatch also collects information on ambient light levels using a silicon photodiode, which has a wavelength range of 400 – 900 nm (570 nm peak sensitivity) and a photopic illuminance range of 5 – 100,000 Lux. The device is typically 90% accurate at 3,000 Lux.

While the piezoelectric accelerometer samples at a fixed 32 Hz, Actiwatches can be configured to record data in various fixed intervals ranging from 30 seconds to 5 minutes. With a 30 second sampling configuration, for example, the total measured activity and light levels are summed over this 30 second interval, which is then stored to the device's internal memory. Configuration of this interval does not affect the accelerometer's sampling frequency, but it does affect the device's battery life and the maximum recording length due to memory usage. When Actiwatch data are downloaded, the exported files contain information on the total activity levels and light for each interval, depending on the frequency specified during configuration.

In a typical clinical setting, Actiwatches are used to monitor a patient's sleep patterns most commonly from ten days to two weeks at a time. In order to create a model that would translate to a clinical setting protocol optimized for longer recordings, we chose to configure the Actiwatches to record data at two-minute intervals (e.g. 0.5 Hz). This specific frequency was chosen because it is the longest sampling interval that the Philips' Actiware implementation of Cole-Kripke will stage sleep, despite the watch allowing configuration of five-minutes, and we wanted to compare the classification of our model to the standard sleep staging model chosen by sleep clinics.

B.   Polysomnography

All subjects completed a clinically-indicated, inlaboratory, Type I sleep study recorded using Polysmith® version 9.0 (Nihon Kohden 2012). Sleep staging was performed by an American Academy of Sleep Medicine (AASM)-accredited polysomnographic technician, and interpreted by a board-certified physician in Sleep Medicine. Standard parameters as

specified by the AASM were captured in the PSG recordings, including electroencephalography (EEG), electromyography (EMG) of the mentalis muscle, electrocardiography (EKG), electrooculography (EOG; left and right eyes), peripheral blood-oxygen saturation (SpO_2), respiratory movement/effort (thorax and abdominal), airflow (nasal and oral), auditory (snoring), and body positioning (right side, left side, supine, prone). Following standard AASM guidelines, each PSG was sleep staged in 30-second intervals, or "epochs", beginning from the first second of the PSG recording.

Fig. 1 displays a sample of actigraphy data versus PSG data collected in the same subject overnight (~8.3 hours). Activity levels are high during wake periods (W) and near-zero during PSG-validated sleep stages (REM, N1, N2, N3).

C. Subjects

Written informed consent was obtained from each subject for this prospective study approved by the VA Portland Health Care System Institutional Review Board (IRB #4085). This study was conducted in accordance with the ethical guidelines of the Belmont Report [19]. Thirty-two ($n = 32$) participants were recruited from the VA Portland Health Care System Sleep Disorders Clinic, who were undergoing clinically-indicated sleep studies (PSG). Patients consented to wear an Actiwatch for 4 weeks, including the night of their PSG. Five participants were eliminated from the original cohort for having insufficient sleep quantity or quality resulting in $n = 27$ at this stage of analysis. PSG studies lasted 7.6 hours on average [5.6 – 8.7 hours]. All remaining participants had at least 90 minutes of sleep. An additional five patients were eliminated due to poor data quality with either their PSG recording or Actiwatch data resulting in final sample size of $n = 22$ (mean age: 49.3 ± 17.6 years; age range: 26-72). The study consort diagram is displayed in Fig. 2, and patient characteristics are displayed in Table 1.

All subjects completed a series of surveys before their PSG. Surveys included demographic information and two validated questionnaires: the Patient-Related Outcomes Measurement Information System (PROMIS) Cognitive Function [20] to asses patient-perceived cognitive deficits such as verbal fluency, concentration, verbal and non-verbal memory, and perceived changes in these cognitive subdomains, and the Neurobehavioral Symptom Inventory (NSI) [21] as a measure of post-concussion symptoms after traumatic brain injury (TBI), which contains assessment of cognitive symptoms including concentration, forgetfulness, and ability to make decisions. On average, our cohort self-reports cognitive dysfunction consistent with some symptoms of early mild cognitive impairment.

D. Data Processing and Machine Learning

Sleep studies were exported from the Portland VA Sleep Clinic in Open eXchange Data Format (.xdf) [22], and were cleaned, processed, and exported to SQL using the Python "openxdf" module. Actiwatch files were exported from Philips Actiware software to CSV and were cleaned, processed, and exported to SQL using the Python "actiwatch" module.

Because of the sampling rate disparity between the subjects' PSG (2 Hz) and actigraphy devices (0.5 Hz), actigraphy data needed to be interpolated to match the physician-validated sleep scoring (Fig. 3). Four interpolation methods were used – nearest neighbor, linear,

quadratic, and cubic – and both activity and light data from the Actiwatch underwent interpolation. After all relevant files were processed, each participant's actigraphy data then underwent interpolation using the four interpolation methods before being joined with their PSG data.

Using the merged, interpolated data, numerous LSTM neural networks were trained, testing multiple architectures and hyper-parameter configurations (i.e., variables that are related to properties of the neural network, such as learning rate, the network structure, etc.). These configurations modulated the sequence length (3, 5, 7, 9, 11, 13, or 15 epochs), the dropout rate (0%, 5%, 10%, or 20%), and the interpolation method for both activity and light, amounting to 448 configurations in total. The various model configurations were trained on an 80-20 training-test split using a categorical cross-entropy loss function, which optimizes the average number of bits needed to identify an event drawn from a set of probability distributions. Note that the training is implemented on one night's sleep data per patient, similar to that which would be undertaken as in a clinical setting.

After training, the models were evaluated using the 20% test set data. For comparison, a confusion matrix of sleep/wake stage in both the true condition (PSG) and the predicted condition (LSTM on actigraphy data) was created, and from this confusion matrix a number of contingency table parameters were generated, including accuracy, sensitivity, specificity, precision, and the F1 score. Because of the imbalanced distribution in sleep/wake in a typical night's sleep, accuracy is not the correct outcome to focus on. Instead, we evaluated model performance using specificity, an area where traditional actigraphy scoring algorithms have suffered. Finally, to compare the output of the LSTM against the control state (traditional actigraphy algorithm scoring), the control scoring was interpolated using last-observation-carried-forward.

III. Results

A confusion matrix was generated for the test set (n=3512 epochs from 4 subjects) comparing traditional PSG for W, N1, N2, N3, and REM versus LSTM model for W and Sleep (S = N1, N2, N3, or REM) (Table 2). Concordance between PSG and the LSTM model was highest for N2, N3, and REM sleep, and lowest for N1 sleep. These data are consistent with the fact that N1 sleep is widely regarded as the least distinguishable stage from W.

The best LSTM model (Fig. 4; with a dropout rate of 10% and a sequence length of 15 epochs) increased the specificity (detecting wake) to 37.7%; more than double of the specificity of the traditional actigraphy scoring methods (15%). Sensitivity (detecting sleep) was moderately lower in the LSTM model compared to traditional actigraphy methods (60.2% vs. 76%). Classification accuracy and precision (67.7% and 90.7%) were similar to the threshold-based approach (75% and 94%).

IV. Discussion and Future Work

Our results show that machine learning approaches, specifically neural network LSTM modeling, can be successfully applied to parse sleep/wake states from actigraphy data when

compared to a common gold standard among older and non-normative patients. Because gold standard polysomnography is so challenging to widely obtain in older, cognitively impaired populations, and novel methods for sleep analysis monitoring and classifying have previously focused on healthy normative patients, our approach highlights the potential for actigraphic analyses in the non-normative population.

While our model specificity is relatively low, it holds a high clinical relevance by nearly doubling the traditional scoring specificity of detecting wakefulness: 37.7% compared to 15% seen in traditional actigraphy scoring. From a clinical perspective, it could be argued that knowing when patients are awake when they should be asleep is the most critical piece of nighttime sleep analysis. Further, because nighttime studies are biased towards "sleep", accuracy is not a particularly useful metric. For the same reason, the F_1 score, the harmonic mean of recall and precision, is problematic in this application because it equally weights recall and precision.

Further, the best LSTM model used quadratic interpolation for activity, nearest-neighbor interpolation for light, and has a sequence length of 15 epochs, meaning it considered the current epoch and the following 14 when classifying the sleep/wake for the current period. Additionally, the chosen sequence length means that the final 14, 30-second epochs are not assigned a sleep stage. For a clinical setting, missing five minutes of data at the tail-end of a night of sleep is insignificant, thus we have no concerns about our model's ability to translate to healthcare applications.

Other groups [23] have achieved considerably higher categorical accuracy with LSTM networks by using multimodal data, including more complicated wrist-worn devices, such as the Affectiva Q-Sensor, which records both skin conductance and temperature, and mobile phone data. However, in addition to access to higher-dimensional data, most groups that achieve higher accuracy work with younger, healthy populations. Furthermore, skin conductance and temperature is more variable in older populations and more difficult to record reliably. While we did not apply multimodal techniques in this study, our proof-of-concept that machine learning can be applied to wrist actigraphy in our older, non-normative sample is promising nonetheless.

Future directions include expansion of these analyses to additional subjects. A larger cohort could allow us to compare the effectiveness of the network on different subsets of the population, investigating the effects of age, sex, and disease burden on wakefulness detection. With more subjects, we would also like to see how the network performs at the level of an individual subject, which is not currently feasible given our small test data sample size. An additional future step is to explore feature engineering, rather than feeding raw data into the neural network.

V.   Conclusion

In conclusion, we have developed a neural network LSTM model which we trained on a sample of older and non-normative patients and have shown that machine learning approaches to wrist actigraphy can be useful for sleep/wake estimation in older and non-

normative patients. The 22 nights of data were trained on gold-standard PSG and were compared against the clinical standard threshold-based method for sleep detection. The LSTM model we have implemented more than doubled traditional method's ability to detect clinically-relevant wakefulness during sleep (37.7% vs. 15%) without sacrificing accuracy (67.7% vs. 75%) or precision (90.7% vs. 94%) of sleep classification. This LSTM model displayed considerably higher specificity at the cost of slightly reduced sensitivity. Neural network models should be considered for the estimation of sleep and wake states in non-normative subjects.

Acknowledgments

References

[1]. Shi L, Chen S-J, Ma M-Y, Bao Y-P, Han Y, Wang Y-M, Shi J, V Vitiello M, and Lu L, "Sleep disturbances increase the risk of dementia: A systematic review and meta-analysis.," Sleep Med. Rev, 7. 2017.

[2]. Boespflug EL and Iliff JJ, "The Emerging Relationship Between Interstitial Fluid–Cerebrospinal Fluid Exchange, Amyloid-β, and Sleep," Biol. Psychiatry, vol. 83, no. 4, pp. 328–336, 2. 2018. [PubMed: 29279202]

[3]. Sadeh A , Hauri PJ, Kripke DF, and Lavie P. "The Role of Actigraphy in the Evaluation of Sleep Disorders," Sleep. vol. 18, no. 4, pp. 288–302, 5 1995. [PubMed: 7618029]

[4]. Sadeh A, "The role and validity of actigraphy in sleep medicine: An update," Sleep Medicine Reviews,; vol. 15, no. 4, pp. 259–267, 8. 2011. [PubMed: 21237680]

[5]. Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen JM, Solet JM, Dulin H, Berkman LF, and Buxton OM, "Measuring Sleep: Accuracy, Sensitivity, and Specificity of Wrist Actigraphy Compared to Polysomnography," Sleep, vol. 36, no. 11, pp. 1747–1755, 11. 2013. [PubMed: 24179309]

[6]. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, and Poliak CP, "The role of actigraphy in the study of sleep and circadian rhythms.," Sleep, vol. 26, no. 3, pp. 342–92, 5 2003. [PubMed: 12749557]

[7]. Kushida CA, Chang A, Gadkary C C, Guilleminault C, Carrillo O, Dement WC, "Comparison of actigraphic, polysomnographic, and subjective asessment of sleep parameters in sleep-disordered patients", Sleep Med. vol. 2, no. 5, pp. 398–396, 9. 2001.

[8]. Trenkwalder C, Stiasny K, Pollmacher T, Wetter T, Schwarz J, Kohnen R, Kazenwadel J, Kruger HP, Ramm S, Kunzel M and Oertel WH, "L-Dopa therapy of uremic and idiopathic restless legs syndrome: A double-blind, crossover trial,' Sleep. vol. 18, no. 8, pp. 681–688, 10. 1995. [PubMed: 8560135]

[9]. Collado-Seidel V, Kazenwadel J, Wetter TC, Kohnen R, Winkelmann J, Selzer R, Oertel WH, Trenkwalder C, "A controlled study of additional sr-L-dopa in L-dopa-responsive restless legs syndrome with late-night symptoms," Neurology. vol. 52, no. 2, pp. 285–290, 1. 1999. [PubMed: 9932945]

[10]. Gruwez A, Bruyneel AV, and Bruyneel M. "The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients," PLoS ONE vol 14, no. 1, 1. 2019.

[11]. Blackwell T, Ancoli-Israel S, Redline S, Stone KL, and Osteoporotic Fractures in Men (MrOS) Study Group, "Factors that may influence the classification of sleep-wake by wrist actigraphy: the MrOS Sleep Study.,"J. Clin. Sleep Med, vol. 7, no. 4, pp. 357–67, 8. 2011. [PubMed: 21897772]

[12]. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, and Gillin JC, "Automatic sleep/wake identification from wrist activity.," Sleep, vol. 15, no. 5, pp. 461–9, 10. 1992. [PubMed: 1455130]

[13]. Sadeh A, Sharkey KM, and Carskadon MA, "Activity-based sleep-wake identification: an empirical test of methodological issues.," Sleep, vol. 17, no. 3, pp. 201–7, 4. 1994. [PubMed: 7939118]

[14]. Lötjönen J, Korhonen I, Hirvonen K, Eskelinen S, Myllymäki M, and Partinen M, "Automatic sleep-wake and nap analysis with a new wrist worn online activity monitoring device vivago WristCare.," Sleep, vol. 26, no. 1, pp. 86–90, 2. 2003. [PubMed: 12627738]

[15]. Paquet J, Kawinska A, and Carrier J, "Wake detection capacity of actigraphy during sleep.," Sleep, vol. 30, no. 10, pp. 1362–9, 10. 2007. [PubMed: 17969470]

[16]. Fang H, He L, Si H, Liu P, and Xie X, "Human activity recognition based on feature selection in smart home using back-propagation algorithm," ISA Trans., vol. 53, no. 5, pp. 1629–1638, 9. 2014. [PubMed: 25016308]

[17]. Schmidhuber J. "Deep learning in neural networks: An overview.", Neural Networks, vol. 61, pp. 85–117, 1. 2015. [PubMed: 25462637]

[18]. Tu JV, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," J. Clin. Epidemiol, vol. 49, no. 11, pp. 1225–1231, 11. 1996. [PubMed: 8892489]

[19]. N. C. for the P. of H. S. of B. and B. Research., "The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.," Bethesda, MD, 1978.

[20]. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse D, Choi S, Cook K, Devellis R, DeWalt D, Fries JF, Gershon R, Hahn EA, Lai J-S, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R, and PROMIS Cooperative Group, "The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008.," J. Clin. Epidemiol, vol. 63, no. 11, pp. 1179–94, 11. 2010. [PubMed: 20685078]

[21]. Soble JR, Silva MA, Vanderploeg RD, Curtiss G, Belanger HG, Donnell AJ, and Scott SG, "Normative Data for the Neurobehavioral Symptom Inventory (NSI) and Post-Concussion Symptom Profiles Among TBI, PTSD, and Nonclinical Samples," Clin. Neuropsychol, vol. 28, no. 4, pp. 614–632, 5 2014. [PubMed: 24625213]

[22]. OpenXDF Consortium "Open eXchange Data Format Specification." 2009.

[23]. Sano A, Chen W, Martinez DL, Taylor S, and Picard RW, "Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks," IEEE J. Biomed. Heal. Informatics, pp. 1–1, 2018.
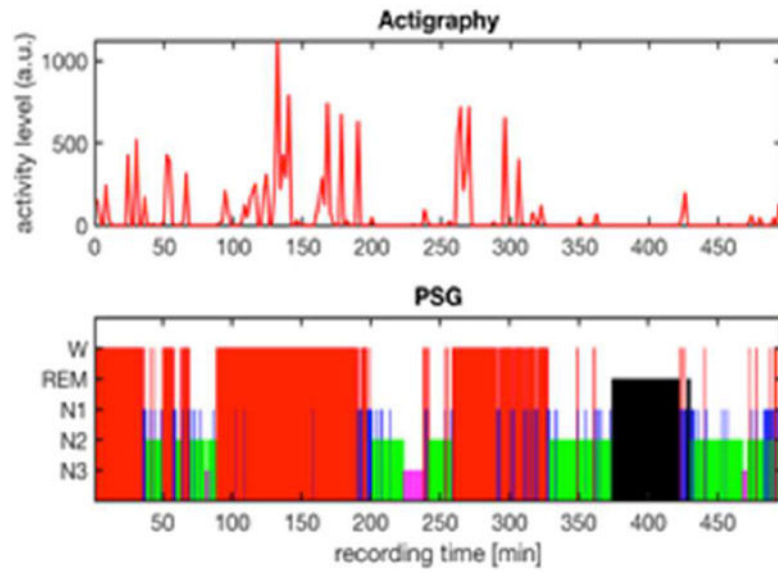
Figure 1.
Raw actigraphy data indicating activity levels (red) and PSG data indicating sleep stages:
W=wake, REM=rapid eye movement, N1=non-REM stage 1, N2=non-REM stage 2,
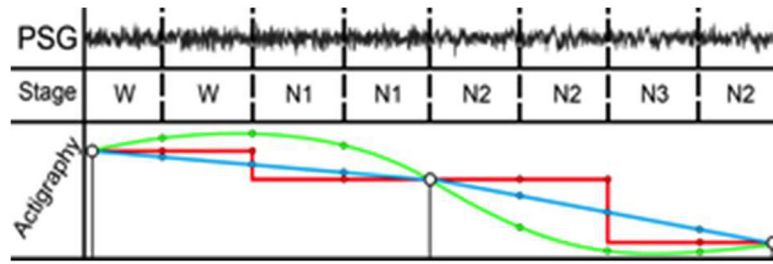N3=non-REM stage 3.

Figure 2.
Study consort diagram.

Figure 3.
Smoothed (green), nearest-neighbor interpolation (red), and linear interpolation (blue) compared with interpolated actigraphy and PSG data. Note that due to the sampling frequency disparity with PSG (2 Hz), actigraphy data (0.5 Hz) was interpolated to match the physician-validated sleep scoring.
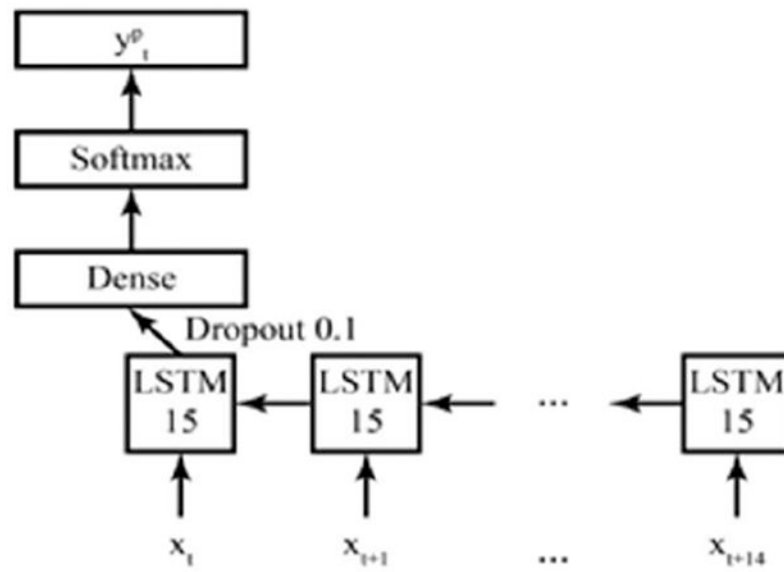
Figure 4.
LSTM model with a dropout rate of 10% and a sequence length of 15 epochs.

Table I.

Patient Demographics

Demographics	*n* = 22	Range
Age (years)	49.3 ± 17.6	26 – 72
Sex (% male)	77%	
NSI Cognitive Subscale	5.0 ± 4.6	0 – 16
PROMIS Cognitive Function	14.5 ± 5.3	4 – 20

Table 2.

Confusion matrix

LSTM Model			
		W	S
PSG	W	1005	443
	N1	72	297
	N2	41	1007
	N3	1	263
	REM	9	374