Time Series Analysis on the Effect of Light Exposure on Sleep Quality

DSC 180 Capstone Report

Shubam Kaushal
UC San Diego
Data Science
shkausha@ucsd.edu

Yuxiang Hu
UC San Diego
Data Science
yuh365@ucsd.edu

Alex Liu
UC San Diego
Data Science
ajl128@ucsd.edu

I. ABSTRACT

The increase of artificial light exposure through the increased prevalence of technology has an effect on the sleep cycle and circadian rhythm of humans. The goal of this project is to determine how different colors and intensities of light exposure prior to sleep affect the quality of sleep through the classification of time series data.

II. INTRODUCTION

A. Background Information

As the world undergoes technological advancement on an unprecedented scale, artificial light from man-made sources is becoming ever more prevalent. The extent of this anthropogenic increase in artificial light has become a pollutant, with extensive research showing both ecological and medical consequences [1]. This is due to the importance of light from the sun on the survival and function of the majority of organisms and thus ecosystems on earth. These organisms have developed day/night cycles that cause physiological, behavioral, and metabolic changes which optimize function and are essential for survival. Artificial light interferes with these processes due to differences in wavelength, intensity, and timing from that of light with origins from the sun. A study of satellite images done in 2001, showed that artificial light at night (ALAN) affects 18.7% of global land area [1], through which roughly two-thirds of the human population and 99% of humans living in the United States and the European Union, “live in areas where the night sky is above the threshold set for polluted status.” [2] The rapid development of technology and thus a rapid increase in artificial light within the last two hundred years has undoubtedly had effects on the biological function of organisms around the world.

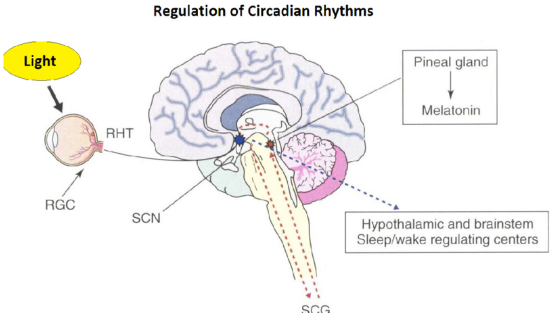
What is most concerning for the health of humans, however, is the ever-increasing use of devices with light-up displays such as phones, TVs, and computers for entertainment, work, and communication. Currently, there are an estimated 16 billion mobile devices worldwide [3] with many individuals spending over five hours a day looking at a screen. One biological mechanism that is affected by this increased exposure to artificial light in human beings is the circadian rhythm, through which the body undergoes changes during the night in preparation for sleep and changes during the day in preparation for activity. The circadian rhythm plays major roles in many “physiological processes, such as body temperature, blood pressure, hormone secretion, gene expression, and immune functions” [4], which all have some reliance on diurnal light patterns from the sun and thus the optimized function of these human body processes are impacted by stimulus from artificial sources of light. When light enters the eyes and is picked up by photosensitive ganglion cells, this information is then communicated to the suprachiasmatic nuclei of the hypothalamus, and then to other parts of the brain and body (*fig 1.*). One result is that the brain experiences an increase in

[1] The actual percentage of global land area affected by artificial light is actually smaller due to this data being taken from satellite images and the measure of skyglow.

wakefulness[2] and reduction in homeostatic sleep pressure in the presence of light [5] through the suppression of melatonin, a hormone released by the pineal gland which facilitates sleep and the circadian rhythm.

As a result light exposure during unnatural times can detrimentally affect sleep, which is necessary for human health and function. Sleep deprivation or impairment can lead to many health issues such as impairment to cognition [6], metabolism [7], and immune response [8]. This leads to the focus of this project, which is to determine the effects of light exposure on sleep quality.

fig 1.


Regulation of Circadian Rhythms
Light
RHT
RGC
SCN
Pineal gland
Melatonin
Hypothalamic and brainstem
Sleep/wake regulating centers
SCG

B. Data

The data used in this project comes from the Sueño Ancillary study done by The Hispanic Community Health Study / Study of Latinos (HCHS/SOL). The data is composed of wrist-worn actimetry sensor[3] data taken over the course of one week for each participant (*n*=2252). Measurements are taken from the sensor in thirty-second intervals and consist of blue, green, red, and white light intensities, locomotor activity, time, and sleep interval indicators [9]. One notable feature that we use is the interval indicator, which describes whether the patient is asleep, awake or resting for a given epoch. This uses the study's sleep/wake detection algorithm to determine.

III. METHODS

A. Feature Engineering

The outcome variable that we used for the data is sleep efficiency which is defined by the ratio between the duration of time the participant spent sleeping over the duration of time spent in bed for a given night [10].

The sleep efficiency equation is shown below:

$$SleepEfficiency = \frac{Total\ Sleep\ Time}{Time\ in\ Bed}$$

To calculate this quantity, we isolated the epochs when a subject switches from one activity to another. Thus, the amount of sleep can be calculated by finding the difference between the epochs when a subject sleeps and when they get up from the bed. Similarly, the amount of time spent in the bed can be calculated by finding the difference between the epochs when a subject comes to bed and when they get up from the bed. For a given sleep event x_i, sleep quality is defined as:

$$\left\{ \begin{array}{l} \text{Good, if } SleepEfficiency(x_i) > 0.95 \\ \text{Bad, if } SleepEfficiency(x_i) \leq 0.95 \end{array} \right\}$$

We created our classifier using the *sktime* library. The classifiers in this library take nested series as feature inputs for univariate classification, and nested series within DataFrames as feature input for multivariate classification. These nested series are indexed and represent the value of the observation that is changing with time, which in our case is the light intensity for white, blue, green, and red colors. The target is simply a series of labels of the corresponding feature inputs. These labels are the sleep quality of a sleep event. In order to get the light exposure time series corresponding to a certain sleep event, we isolated the light exposure until 2 hours before a subject went to sleep. The series for any color of light for different sleep events were then placed within another series, thus creating a nested series. This creates a feature input and target feature that can be used for multivariate time series classification using the *sktime* library. Below is an example of how a DataFrame in this format would be structured:

[2] Wakefulness here is defined as improved auditory reaction time, improved ECG readings indicating alertness, and reduced attentional lapses.

[3] Actiwatch Spectrum, Philips Respironics

$$\left[\begin{array}{cccc|cc} L_1^W & L_1^B & L_1^G & L_1^R & SE_1 & SQ_1 \\ L_2^W & L_2^B & L_2^G & L_2^R & SE_2 & SQ_2 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ L_n^W & L_n^B & L_n^G & L_n^R & SE_n & SQ_n \end{array}\right]$$

The DataFrame is a $n\,x\,6$ matrix where n is the number of complete active-rest intervals. Each element L is a series of length 240 (equivalent to 2 hours worth of 30-second epochs) where the superscript represents the associated color of light. SE and SQ represent the sleep efficiency ratio and sleep quality rating respectively. The value of SQ for any given SE is dependent on the threshold, which we chose to be 0.95 because this value represents a sleeping event where a person spent about 8 hours in the bed but did not sleep until about 25 minutes passed by. This choice is completely discretionary - other studies have used a value of 0.85 [https://mhealth.jmir.org/2016/4/e125] based on their discretion.

B. ROCKET [https://arxiv.org/abs/1910.13051]

Usually time series classification models are highly complex and require long training times even for small quantities of data. Some methods focus on properties such as frequency or shape, while others use learned convolutional kernels to perform classification. However, these methods are not scalable at all.

Rand**O**m **C**onvolutional **KE**rnel **T**ransform (ROCKET) involves creating features from time series using random convolutional kernels. These kernels have random length, weights, biases, dilation and padding. Due to the randomness in the kernels and the resulting features, it is virtually impossible to interpret these features. In fact, interpretability remains a huge challenge in the realm of time series classification. The number of kernels is usually 10,000 but the transformations still take place extremely fast. For *k* kernels and *n* time series, where each time series is of length *l,* the complexity of ROCKET is *O(k . n . l).*

C. Classifiers with ROCKET

Once the random convolutional features are created, they can be used to train a linear classifier. In theory, ROCKET can be used with any classifier such as Logistic Regression or Ridge Regression.

Logistic Regression: ROCKET can be used with logistic regression and stochastic gradient descent. This is particularly suitable for very large datasets because it provides for fast training with a memory cost fixed by the size of each minibatch. The complexity of stochastic gradient descent is proportional to the number of features and the number of classes (which is 2 in our case) but is linear in relation to the number of training examples.

Ridge Regression: For our practical use case, however, we use a ridge regression classifier. The ridge regression classifier is significantly faster than logistic regression on smaller datasets because it can make use of so-called generalized cross-validation to determine appropriate regularization. Regularization is critically important where the number of features is significantly greater than the number of training examples, allowing for the optimization of linear models, and preventing pathological behavior in iterative optimization, e.g., for logistic regression. [cite]. The ridge regression classifier can exploit generalized cross-validation to determine an appropriate regularization parameter quickly. In the case of our model, the number of features is in the order of ten thousand while the number of training examples is in the order of thousands, so ridge regression is computationally efficient while maintaining high accuracy. The non-scalability of ridge classification to larger datasets did not pose a problem for us.

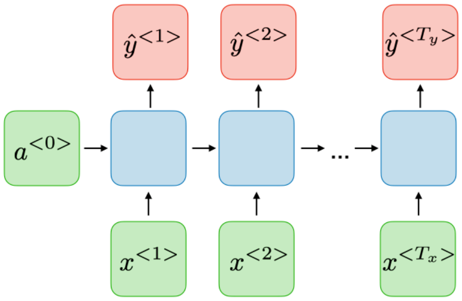
D. Evaluation metric for ROCKET-Ridge Classifier

After performing feature engineering, it turned out that only about 13% of sleep events were of 'Bad' sleep quality. In the case of such imbalanced data labels, the usual *accuracy* metric would not be appropriate because a naive model that always predicts 'Good' sleep quality would have 87% accuracy, but it would be practically useless. Therefore, *recall* with respect to 'Bad' sleep quality is a better evaluation metric. *Recall* with respect to 'Bad' sleep quality is the ratio between the number of correctly identified 'Bad' sleep quality events and a total number of actual 'Bad' sleep quality events. When this ratio is high, it means that the model is able to correctly predict more of the ‘Bad’ sleep quality events.

E. LSTM

Considering our data is time series with the frequency of 1 observation per 30 seconds and we are trying to predict the sleep-awake status of an individual based on the series of light he/she received. A recurrent neural network (RNN) might be ideal to accomplish this goal as RNN will learn the relationship from time step to time step. It will produce predictions not only from the current light intensity but also from the previous light the individual received.

Long Short-Term Memory(LSTM)[12] is an RNN model that fixed the problem of losing the long-term dependency of traditional RNN models. We are using the many-to-many architecture where $x^{<t>}$ is the light intensity and $y^{<t>}$ be the prediction of the sleep-awake status of t^{th} time step.



F. Evaluation metric for LSTM

In the model, we are directly predicting the sleep-awake status of every time step and compute the sleep quality from the predictions. As we are not directly predicting sleep quality, we are not too concerned about the imbalanced label of sleep quality. The main metric to evaluate the LSTM model is still the accuracy of the individual prediction in each time step. But we will also evaluate the accuracy of the calculated sleep quality and recall.

IV. RESULTS

A.Training the LSTM

Unlike the ROCKET model, we trained the LSTM model with raw unbalanced data, because the sequence of data passed into the model is important for the model to learn the correlation between each time step. As the dimensionality of the model is high and they are inter-correlated, we don’t want our oversampling on the bad sleep quality data to generate unnecessary noise to the model.

We used the first 1600 records of patients for the training set and the remaining part of the dataset as a test set. We used the 80-20 train-validation split on the training set and trained the model for 10 epochs. We saved the model with the lowest validation loss to avoid overfitting the model and generate the test result from it.

B. LSTM results

After training the LSTM model and getting the best model from the validation set, we tested it against the test set. It turns out that our model performs pretty well. The accuracy for the prediction of sleep-wake states in each time step is 92.94% and the accuracy is 98.26% for classifying the good or bad sleep quality calculated from the predicted sleep-wake states. The accuracy is high, but it doesn’t mean this is a perfect model. The True negative rate of this model is 0. This means that given light intensity received by a patient with bad sleep quality, the model never predicts the patient is having bad sleep quality. This may be caused by the imbalanced data input to the LSTM for training, as we are concerned with the high dimensionality of inter-correlation between time steps so we didn’t oversample the bad sleep quality data.

However, though this LSTM model is not suitable for sleep quality classification, it is capable of sleep state classification based on light intensity as it has a cross-entropy loss of 0.82, which is very low.

C. ROCKET-Ridge Models

All the attempted models in this section are similar in the sense that they apply ROCKET on the light exposure time series upto two hours before any sleep event to create features and then use Ridge Classifier on those features to predict the sleep quality of any sleep event. Given that the data is imbalanced, we chose to filter the data such that there was a 50/50 ratio between ‘Good’ and ‘Bad’ quality sleep events in the training dataset. The ROCKET-Ridge classifiers do not need any hyperparameter tuning because ROCKET creates tens of thousands of random features and these features are easily learned by the Ridge classifier without overfitting due to effective regularization. Therefore, we split the data into a 75/25 training-test split with no validation set.

I. All Lights Model: Baseline

The first baseline model included light exposure time series for all four colors: red, blue, green, and white.
'Bad' Sleep Quality Recall: 0.64
'Good' Sleep Quality Recall: 0.58

II. Red Light Model

This model only included light exposure time series for the red color.
'Bad' Sleep Quality Recall: 0.63
'Good' Sleep Quality Recall: 0.60

III. Green Light Model

This model only included light exposure time series for the green color.
'Bad' Sleep Quality Recall: 0.64
'Good' Sleep Quality Recall: 0.60

IV. Blue Light Model

This model only included light exposure time series for the blue color.
'Bad' Sleep Quality Recall: 0.59
'Good' Sleep Quality Recall: 0.58

V. White Light Model

This model only included light exposure time series for the white color. This model performs the best when it comes to models with only a single light color.
'Bad' Sleep Quality Recall: 0.62
'Good' Sleep Quality Recall: 0.61

VI. Vote Model: Best

This model uses Red, Blue, Green and White Light Models individually to get their predictions. Then, for any sleep event, the predictions from the four models are put to vote and the most common outcome is selected as the final prediction. In case of ties, the prediction of White Light Model determines the outcome because it is the best performing model with only a single color. This model turns out to be the best performing model.
'Bad' Sleep Quality Recall: 0.63
'Good' Sleep Quality Recall: 0.62

D. Observations from ROCKET-Ridge Models

I. Vote Model performance difference based on activity level

To assess the model performance on different subsets of dataset, we first split it into two groups: first group included participants whose mean activity over the week of recording was less than the overall median of mean activity levels of all the participants (less active), and the second group included the rest of the participants (more active).

'Bad' Sleep Quality Recall (less active): 0.61
'Good' Sleep Quality Recall (less active): 0.62

'Bad' Sleep Quality Recall (more active): 0.65
'Good' Sleep Quality Recall (more active): 0.62

So, the Vote Model performs similarly for 'Good' sleep quality events, both overall and across groups. However, the model performs better for 'Bad' sleep quality detection for more active participants in comparison to less active participants. This gives us an intuition that among less active participants, the light wave features are not good enough at predicting 'Bad' sleep quality, as compared to that in more active participants. This means that there must be other confounding features among less active participants that are causing 'Bad' sleep quality. These confounding features among less active individuals could be age, health conditions or events happening outside the two-hour window before sleeping that are dictating the 'Bad' sleep quality among less active participants.

II. Vote Model performance difference based on sleep level

Next, we split our dataset into two groups based on the amount of time that participants spent sleeping: first group included participants whose total sleep duration over the week of recording was less than the overall median of sleep duration of all the participants (less sleep), and the second group included the rest of the individuals (more sleep).

'Bad' Sleep Quality Recall (less sleep): 0.65
'Good' Sleep Quality Recall (less sleep): 0.66

‘Bad’ Sleep Quality Recall (more sleep): 0.61
‘Good’ Sleep Quality Recall (more sleep): 0.58

So, the Vote Model performs better for participants that sleep less and worse for participants that sleep more, irrespective of the kind of sleep quality. This gives us an intuition that among participants who sleep more, the light wave features are not good enough at predicting their sleep quality to be ‘Good’ or ‘Bad’, as compared to participants who sleep less. So there must be other confounding features that are dictating this classification among participants who sleep more. Again, these confounding features among individuals that sleep more could be age, health conditions or events happening outside the two-hour window before sleeping that are dictating the sleep quality classification among participants that sleep more.

III. Correlation between different colors of light exposure levels and its consequences

We know that for any average individual, the light exposure at any instant is a mixture of different colors. In fact, any color can be decomposed into its red, blue and green components. White light is a mixture of all three colors at their maximum intensity and typical sources include sunlight and usual artificial lighting indoors. So, as per our expectations, the correlation value between light exposure levels between any pair of colors in the two-hour window before a sleep event is very high. The least median correlation occurs between red and blue light (= 0.73) and the highest median correlation occurs between green and blue light (= 0.94).

We know that ROCKET creates random convolutional features from the time series and those features are then used by the Ridge Classifier for time series classification. Since the time series for different colors have high correlation, we would expect that ROCKET would create similar wave features across different light time series. Since similar features are being used, the recall for any sleep quality by Red, Blue and Green Light Model should be the same; and if that is not the case, it must be due to the difference in the color of the light.

To determine the difference in model performance between the Red, Green and Blue Light Model on ‘Bad’ sleep quality events, we perform bootstrapping on the test set and get a distribution of recall values with respect to ‘Bad’ sleep quality events for each of the three models. Then, we use the Kolmogrov-Smirnov statistic (KS-statistic) to determine whether the three distributions of recall values are the same or not. The mean recall values with respect to ‘Bad’ sleep quality are as follows:

Red Light Model: 0.63
Green Light Model: 0.64
Blue Light Model: 0.59

With a significance level of 1%, we find out that:

a) The recall values for Red and Green Light Models come from the same distribution (p-value = 0.91)

b) The recall values for Red and Blue Light Models come from different distributions (p-value = 0.002)

c) The recall values for Green and Blue Light Models come from different distributions (p-value = $0.6\text{x}10^{-4}$)

Thus, the recall with respect to ‘Bad’ sleep quality for the Blue Light Model is clearly less than that of the other two models. The three models are the same in every aspect except the color of the light, which is not a feature in our ROCKET-Ridge Models. This gives us an intuition that the same wave features that work well in Red and Green Light Models do not work well enough in Blue Light Model to determine ‘Bad’ sleep quality. So, it must be that the reason why random convolutional features in the Blue Light Model aren’t enough to determine ‘Bad’ sleep quality events is because the color blue itself causes ‘Bad’ sleep quality to some extent, and our model is unable to capture that because color is not a feature. This is in line with the popular theory that blue light negatively affects sleep quality because it suppresses the secretion of melatonin - the sleep-causing hormone.

V. CONCLUSION

VI. ACKNOWLEDGEMENTS

We would like to acknowledge Professor Benjamin Smarr for his guidance on this project.

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) was performed as a collaborative study supported by contracts from the NHLBI to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237) (AG05407, AR35582, AG05394, AR35584, AR35583, AG08415). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

VII. REFERENCES

[1] Gaston, K. J., Bennie, J., Davies, T. W., &amp; Hopkins, J. (2013). The ecological impacts of nighttime light pollution: A mechanistic appraisal. Biological Reviews, 88(4), 912–927. https://doi.org/10.1111/brv.12036

[2] P. Cinzano, F. Falchi, C.D. Elvidge, The first World Atlas of the artificial night sky brightness, Monthly Notices of the Royal Astronomical Society, Volume 328, Issue 3, December 2001, Pages 689–707, https://doi.org/10.1046/j.1365-8711.2001.04882.x

[3] Published by S. O'Dea, S. O. D. (2021, September 24). Number of mobile devices worldwide 2020-2025. Statista. Retrieved March 4, 2022, from https://www.statista.com/statistics/245501/multiple-mobile-device-ownership-worldwide/

[4] Cable, J., Schernhammer, E., Hanlon, E. C., Vetter, C., Cedernaes, J., Makarem, N., Dashti, H. S., Shechter, A., Depner, C., Ingiosi, A., Blume, C., Tan, X., Gottlieb, E., Benedict, C., Van Cauter, E., &amp; St‑Onge, M. P. (2021). Sleep and circadian rhythms: Pillars of Health—a keystone symposia report. Annals of the New York Academy of Sciences, 1506(1), 18–34. https://doi.org/10.1111/nyas.14661

[5] Rahman SA, Flynn-Evans EE, Aeschbach D, Brainard GC, Czeisler CA, Lockley SW. Diurnal spectral sensitivity of the acute alerting effects of light. Sleep. 2014 Feb 1;37(2):271-81. doi: 10.5665/sleep.3396. PMID: 24501435; PMCID: PMC3900613.

[6] Killgore, W. D. S. (2010). Effects of sleep deprivation on cognition. Progress in Brain Research, 105–129. https://doi.org/10.1016/b978-0-444-53702-7.00007-5

[7] Knutson, K. L., Spiegel, K., Penev, P., &amp; Van Cauter, E. (2007). The metabolic consequences of sleep deprivation. Sleep Medicine Reviews, 11(3), 163–178. https://doi.org/10.1016/j.smrv.2007.01.002

[8] Spiegel K, Sheridan JF, Van Cauter E. Effect of Sleep Deprivation on Response to Immunization. JAMA. 2002;288(12):1471–1472. doi:10.1001/jama.288.12.1469

[9] Patel SR, Weng J, Rueschman M, Dudley KA, Loredo JS, Mossavar-Rahmani Y, Ramirez M, Ramos AR, Reid K, Seiger AN, Sotres-Alvarez D, Zee PC, Wang R. Reproducibility of a Standardized Actigraphy Scoring Algorithm for Sleep in a US Hispanic/Latino Population. Sleep. 2015 Sep 1;38(9):1497-503. doi: 10.5665/sleep.4998. PMID: 25845697; PMCID: PMC4531418.

[10] Sathyanarayana A, Joty S, Fernandez-Luque L, Ofli F, Srivastava J, Elmagarmid A, Arora T, Taheri S; Sleep Quality Prediction From Wearable Data Using Deep Learning; JMIR Mhealth Uhealth 2016;4(4):e125; doi: 10.2196/mhealth.6562; PMID: 27815231; PMCID: 5116102

[11] Dempster, A., Petitjean, F. & Webb, G.I. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Min Knowl Disc 34, 1454–1495 (2020). https://doi.org/10.1007/s10618-020-00701-z

[12]Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation. 9 (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276. S2CID 1915014.