Sleep Quality

Haipeng LI, Lucas MONTI

March 11, 2019

Encadrant: Sorin MOGA

1 Objective

1.1 Sleep cycles

2 Sleep staging with polysomnography

2.1 Sleep dataset

CCSHS The Cleveland Childrens Sleep and Health Study dataset: with full-night polysomnographic recordings from 515 participants (16-19 years) with minority population representation from the National Sleep Research Resource website (Hibbs et al. (2013)).

Participants	F/M	Age	Population	Epochs	Acquisition period
50	25/25	16-19	Healthy	67217	2006-2010

Sleep-EDFx The public Sleep-EDF dataset: from the PhysioNet database featuring 39 full-night recordings from 20 participants (Kemp et al. (2000)).

Participants	F/M	Age	Population	Epochs	Acquisition period
39	19/20	21-101	Healthy	50559	1987-1991

UCD The public St. Vincent dataset: with full-night recordings from 25 participants from the PhysioNet database (Goldberger et al. (2000)).

Participants	F/M	Age	Population	Epochs	Acquisition period
25	4/21	28-68	Obese and apnoea	20789	2002-2003

EMSA The private Episodic Memory and Sleep Assessment dataset: with 51 participant full-night polysomnographic recordings including 18 school children (8-12 years) and 33 adults (18-30 years) divided into the subsets EMSAch (children) and EMSAad (adults) (Goldberger et al. (2000)).

Participants	F/M	Age	Population	Epochs	Acquisition period
33	16/17	18-30	Healthy	32105	2012-2015

SHHS todo.

Participants	F/M	Age	Population	Epochs	Acquisition period
todo	todo	todo	todo	todo	todo

All datasets were staged in 30-second epochs, according to AASM(American Academy of Sleep Medicine), the following electrodes will be used in the sleep-staging process (Wang et al. (2018)):

- EEG: an electrophysiological monitoring method to record electrical activity of the brain.
- EMG: an electrodiagnostic medicine technique for evaluating and recording the electrical activity produced by skeletal muscles

- EOG: a technique for measuring the corneo-retinal standing potential that exists between the front and the back of the human eye.

The hypnogram annotations contain the following labels:

- W: subject is awake
- 1: sleep stage 1
- 2: sleep stage 2
- 3: sleep stage 3
- 4: sleep stage 4
- R: REM sleep

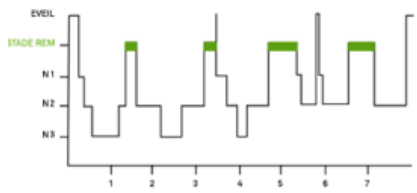


Figure 1: Hypnogram.

2.2 Sleep staging algorithm

Today, there are many algorithms to carry out the sleep scoring procedure, but there exist some problems and shortcomings:

- Lack of large training databases: the reason is that the acquisition of sleep data is time-costly, and nowadays the quality of dataset is not enough.
- Inter-subject variability: depending on age, current sleep disorders and other neurological disorders, sleep EEG in the same sleep stage may look very different.
- Inter-subject variability: depending on age, current sleep disorders and other neurological disorders, sleep EEG in the same sleep stage may look very different.

so we've chosen two algorithms to automatize the sleep scoring procedure:

- AutoSleepScorer: Automatic Sleep Stage Classification using Convolutional Neural Networks with Long Short-Term Memory (Kern (2017)).
- DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG (Supratak et al. (2017)).

as a result, we gain some advantages from the automatic sleep scoring procedures:

- The result of staging would be consistent.
- Save a lot of time for patients and doctors.

2.3 AutoSleepScorer

AutoSleepScorer is a pilot open-source project to create a robust sleep stage using Convolutional Neural Networks with Long Short-Term Memory.

During the process, a Convolutional Neural Network with Long Short-Term Memory is used for the detection of sleep stages. This approach has the advantage that:

- it can automatically detect and extract features from the raw EEG, EMG and EOG signal.

The network architecture was trained, validated, and tested on several different public and private datasets to ensure a good generalizability (CCSHS, EDFx, EMSA).

Currently the classifier reaches the state-of-the-art of automatic sleep stage classification while obtaining a similar performance to a human scorer(80%). And below is a hypnogram of stage predictions of the CNN+LSTM on one participant of the CCSHS(The Cleveland Childrens Sleep and Health Study) dataset:

To evaluate the the performance of models, normally we carry out the metric of evaluation consists of **Accuracy**, **Recall**, **Precision**, **F1-score**. Here are some evaluations obtained within one dataset concluding training set and testing set:

Data Set	Accuracy	F1-score
CCSHS	89%	81%
EDFx	87%	80%
EMSA	87%	77%

Below is the flowchart to explain the architecture of Autosleepscorer:

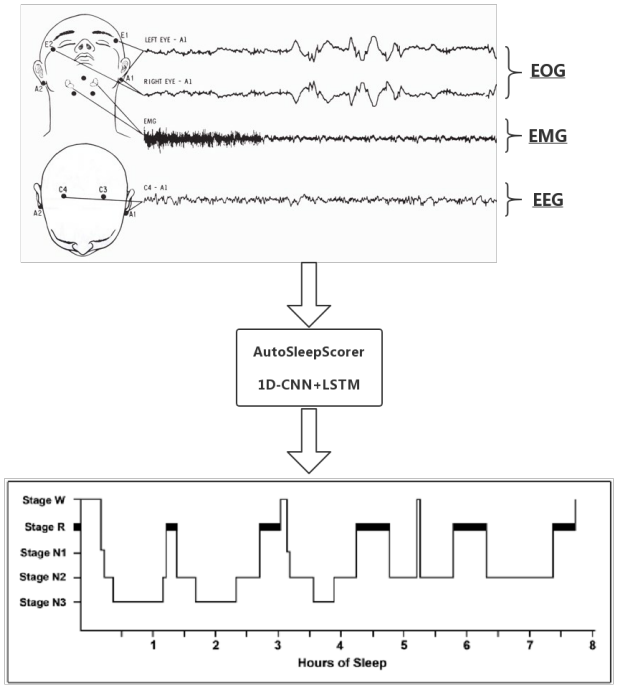


Figure 2: Convert polysomnography to sleep stages

2.4 DeepSleepNet

DeepSleepNet is a deep learning model for automatic sleep stage scoring based on raw, single-channel EEG.

The model is based on DeepSleepNet: a Model for Automatic Sleep Stage Scoring based on Raw Single-Channel EEG by Akara Supratak, Hao Dong, Chao Wu, Yike Guo from Data Science Institute, Imperial College London(Supratak et al. (2017)).
Compared with the first model(AutoSleepScorer), DeepSleepNet carries out the automatic sleep staging process based on raw single-channel EEG. DeepSleepNet gets the similar architecture with AutoSleepScorer. Precisly, it utilizes:

- Convolutional Neural Networks to extract time-invariant features.
- Bidirectional-Long Short-Term Memory to learn transition rules among sleep stages automatically from EEG epochs.

Within the model, they implement a two-step training algorithm to train the model efficiently which are:

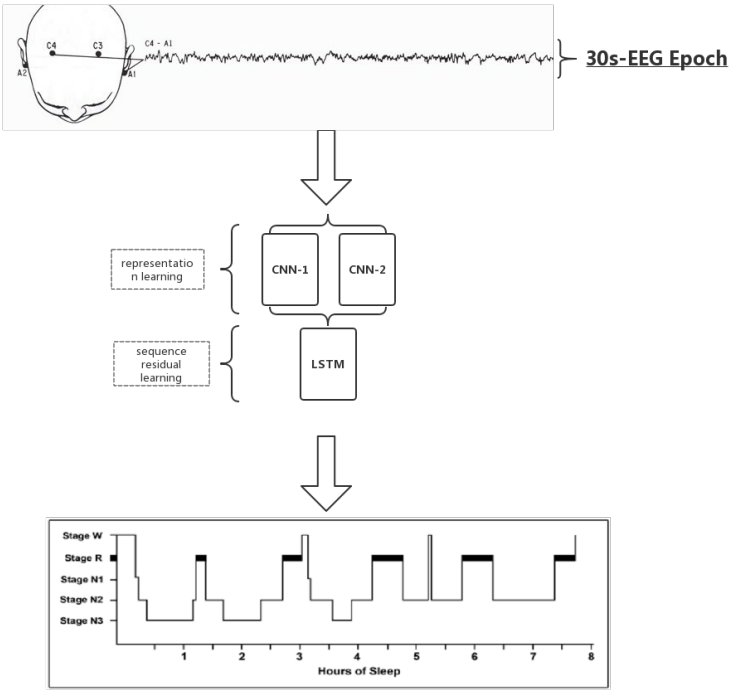


Figure 3: Convert polysomnography to sleep stages

- Representation learning.
- Sequence residual learning.

The results showed that the model achieved similar overall accuracy and macro F1-score(Supratak et al. (2017)):

Data Set	Accuracy	F1-score
MASS	86.2%	81.7%
Sleep-EDF	85.9%	80.5%

compared to the state-of-the-art methods(Tsinalis et al. (2016)):

Data Set	Accuracy	F1-score
MASS	85.9%	80.5%
Sleep-EDF	78.9%	73.7%

3 Sleep Quality Dataset

3.1 Sleep dataset: SHHS

We collecte a processed dataset of polysomnograpy from National Sleep Research Resource(NSRR) named The Sleep Heart Health Study(SHHS).

The Sleep Heart Health Study (SHHS) is a multi-center cohort study implemented by the National Heart Lung Blood Institute to determine the cardiovascular and other consequences of sleep-disordered breathing.

Participants	F/M	Age	Population	Epochs	Acquisition period
5,804		40-100	Healthy		1995-1998&2001-2003

In all, 6,441 men and women aged 40 years and older were enrolled between November 1, 1995 and January 31, 1998 to take part in SHHS Visit 1. During exam cycle 3 (January 2001- June 2003), a second polysomnogram (SHHS Visit 2) was obtained in 3,295 of the participants.

3.2 Data processing

	nsrrid	pptid	ecgdate	lvh3_1	lvh3_3	st4_1_3	st5_1_3	lvhst	mob1	part2deg	...	Abdoqual	EEG1qual	EEG2qual	EOGRqual	EOGLqual	Chinqual
0	200001	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	4.0	3	3	4	4	4
1	200002	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2.0	3	2	2	2	2
2	200003	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	3.0	4	4	4	4	4
3	200004	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	3.0	3	3	3	3	3
4	200005	5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	4.0	4	4	4	4	4
5	200006	6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	4.0	4	4	4	4	4

Figure 4: SHHS(5804 rows,1279 columns)

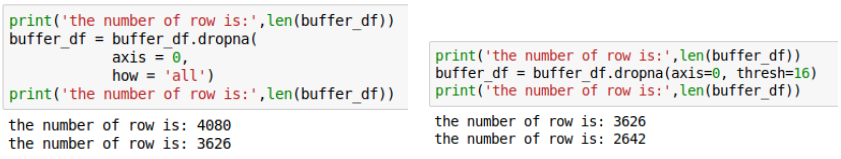
After doing some basic analysis, we find that there exists a lot of empty values and too much features, so firstly we're going to do feature selecting. We choose those features related to **Sleep Architecture**, **Heart Rate**, **Blood Pressure**, **Breathe**:

- avg23bpd_s2: Average Diastolic blood pressure (BP).
- avg23bps_s2: Average Systolic blood pressure (BP).
- ai_all: Arousal Index.
- havbrbh: Average Heart Rate.
- rdi0p: Overall Respiratory Disturbance Index (RDI) all oxygen desaturations – The prevalence of obstructive sleep apnea (OSA) depends on the definition of the respiratory disturbance index (RDI) or apneahypopnea

index (AHI) criteriaObstructive Sleep Apnea in Adults: Epidemiology and Variants.

- nsupinep: Percent Time non-supine
- pctlt75: Percent of sleep time with less than 75% oxygen saturation (SaO2)
- pctlt80: Percent of sleep time with less than 80% oxygen saturation (SaO2)
- pctlt85: Percent of sleep time with less than 85% oxygen saturation (SaO2)
- pctlt90: Percent of sleep time with less than 90% oxygen saturation (SaO2), Ratio of the number of minutes with oxygen saturation (SaO2) under 90% to the total sleep time expressed in hours.
- slp_eff: Percentage of time in bed that was spent sleeping, or the ratio of total sleep time to total time in bed, expressed as a percentage.
- slp_lat: Time from lights out time to beginning of sleep, rounded to nearest minute.
- slpprdp: Sleep Time
- slptime: Total Sleep Time
- supinep: Percent Time supine
- times34p: Percent Time in Stage 3/4
- timest1p: Percent Time in Stage 1
- timest2p: Percent Time in Stage 2
- waso: Total amount of time spent awake after going to sleep

Then we do the data cleaning, we're going to keep those rows with at least nine features and delete those columns with all empty values:



(a) step one (b) step two

Figure 5: Data Cleaning

For the rest empty values, we're going to calculate and fill in the average values. After finishing the data cleaning process, we get our dataset with **2642 rows** and **20 features** for the next step of sleep-quality modeling.

avg23bpd_s2	13
avg23bps_s2	8
ai_all	32
rdi0p	0
nsupinep	11
pctlt75	2
pctlt80	2
pctlt85	2
pctlt90	2
slp_eff	0
slp_lat	575
slpprdp	0
supinep	11
times34p	25
timest1p	25
timest2p	25
waso	0
ms204a	42
ms204b	80
ms204c	58
dtype: int64	

(a) before filling

avg23bpd_s2	0
avg23bps_s2	0
ai_all	0
rdi0p	0
nsupinep	0
pctlt75	0
pctlt80	0
pctlt85	0
pctlt90	0
slp_eff	0
slp_lat	0
slpprdp	0
supinep	0
times34p	0
timest1p	0
timest2p	0
waso	0
ms204a	0
ms204b	0
ms204c	0
dtype: int64	

(b) after filling

Figure 6: Calculate and Fill In the Average

3.3 Indices of sleep quality

During a sleep, there are two different moments when a person is awake(REM,W) and when he is asleep(N1,N2,N3). And the process from awake time to the sleep period is called **the sleep onset time**, and on the other hand, the process from sleep period to awake time is referred to **the sleep awakening time**. Whats more, the period of time between the self-reported time to bed and the sleep onset time is called the **latency**.

To measure the sleep quality, there are a lot of factors, but the most important factor is called **sleep efficiency**, which is the ratio of total minutes asleep to total minutes in bed. Below is the formula:

$$\begin{aligned} \textit{Sleep Efficiency} &= \frac{TotalSleepTime}{TotalMinutesinBed} \\ &= \frac{||SleepPeriod|| - WASO}{||SleepPeriod|| + Latency} \end{aligned}$$

Figure 7: Sleep Efficiency(Sathyanarayana et al. (2016))

- **TotalMinutesInBed** represents the amount of time that an individual spends asleep as well as the amount of time the individual takes to fall asleep, that is, latency.
- **TotalSleepTime** represents the amount of time that an individual spends asleep, minus the amount of time the person awakens. This is calculated by subtracting the wake after sleep onset (**WASO**) from the duration of the sleep period.
- **WASO** is the sum of all moments of wakefulness lasting longer than 5 minutes.

In addition to **sleep efficiency**, there are some factors that affect sleep quality, such as(Krystal and Edinger (2008)):

- total sleep time (TST)
- sleep onset latency (SOL)
- total wake time
- spontaneous arousals or apnea
- the percentage or temporal amounts of stage 1 sleep, stage 2 sleep
- slow wave sleepor rapid eye movement (REM)

3.4 The definition of sleep quality

In order to build the sleep quality model, we select the indicators in the **chapter 3.3** as training features. For training labels, we carry out the simple Likert-style rating of (the previous nights) sleep quality, commonly included as an item on sleep diaries, as the core sleep quality indicator(Krystal and Edinger (2008)), and we focus on three indicators(Silva et al. (2007)):

- **ms204a**: Morning Survey (Sleep Heart Health Study Visit Two (SHHS2)), Quality of sleep light/deep.
- **ms204b**: Morning Survey (Sleep Heart Health Study Visit Two (SHHS2)): Quality of sleep: short/long.
- **ms204c**: Morning Survey (Sleep Heart Health Study Visit Two (SHHS2)): Quality of sleep: restless/restful.

Based on the above three indicators, we can approximately evaluate the quality of a person's sleep for one night. Until now, we finish the preparation of data set(SHHS) including data processing, feature selecting, definition of sleep quality, etc.. In the next part, we're going to build our sleep-quality evaluation system.

4 Sleep Quality Algorithms

As explained in more detail below, in this project we're going to explore the use of machine learning and deep learning methods to predict sleep quality based on polysomnography data. The models used in our study were as follows:

- **DNN**: an artificial neural network (ANN) with multiple layers between the input and output layers(Deng et al. (2014)Bengio et al. (2009)), a deep learning model
- **Random Forest**: an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees(Ho (1995)), a machine learning model.
- **LightGBM**: a gradient boosting framework that uses tree based learning algorithms.

Data Partitioning: To train our models and test their performance, we create a random partitioning of the dataset. The data were split to:

Training	Validation
80%	20%

Input of the models: The input of the models are selected-feature vectors, X=(x1,x2,x3,...,x17), representing the core datas of a person's sleep time. Each vector correspond to a specific indicator during sleep time.

Output of the models: The output of the model was a three-criteria(0,1,2) classification decision between poor and good sleep quality based on the core indicators. These classifications corresponded to the sleep indicators as described earlier. In addition to the three-criteria decision, the model also gave its confidence (a score between 0.0 and 1.0) in that decision.

4.1 Deep Neural Networks

To be able to predict, firstly we're going to train the DNN model on the training dataset. We designed a six-layer neural network including four hidden layers and one output layer, and the first hidden layer contains 75 neurons, the second hidden layer contains 50 neurons, the third hidden layer contains 50 neurons, the fourth hidden layer contains 35 neurons, finally, the output layer calculates the probability by the **SOFTMAX** activation function to get three results.

During the training process, in order to speed up training and prevent overfitting, to avoid gradient explosion or gradient disappearance, and to optimize model, we will do some optimization measures.

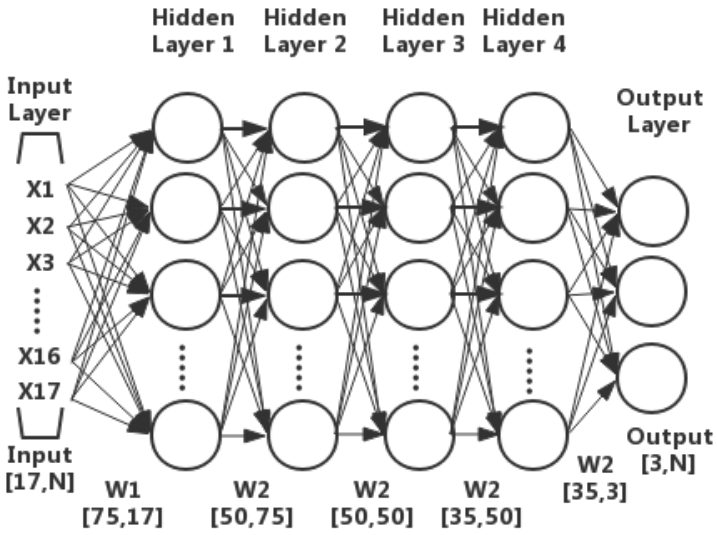


Figure 8: DNN for sleep quality

In order to convert numbers to vectors, for example, for one good sleep quality where **204a = 3**, we need to convert it to one vector so that our DNN model can be trained on, such as [0 0 1]. We're going to use **One Hot Encoding**

Optimization measures:

- Feature Normalization
- Xavier Initialization for weights and Zero Initialization for biases.
- Batch Normalization
- L1 and L2 regularization
- AdamOptimizer
- Mini-batch Gradient Descent

4.2 Random Forest

First we focus on the parameters of the RF bagging framework, compared with GBDT(Gradient Boosting Decision Tree), GBDT has more frame parameters to tune, there are less parameters to tune the hyperparameters of RF. During the training process, in order to speed up training and prevent over-fitting, we will do some optimization measures:

Optimization measures:

- **n_estimators**: maximum number of iterations of the weak learner, or the number of the largest weak learners. Generally, if **n_estimators** is too small, the model is easy to underfit, on the other hand, if **n_estimators** is too large, the amount of calculation will be too large, moreover, after **n_estimators** to a certain number, the model lift will be small, so generally choose a moderate value. In this model we choose 100.
- **oob_score**: Whether to use out-of-bag samples to estimate the generalization accuracy. We selected true, because the extra-bag score reflects the generalization ability of a model.
- **criterion**: The function to measure the quality of a split. Supported criteria are gini for the Gini impurity and entropy for the information gain.
- **max_features**: The number of features to consider when looking for the best split:
  - If int, then consider max_features features at each split.
  - If auto, then max_features=sqrt(n_features).
  - If log2, then max_features=log2(n_features).
- **max_depth**: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

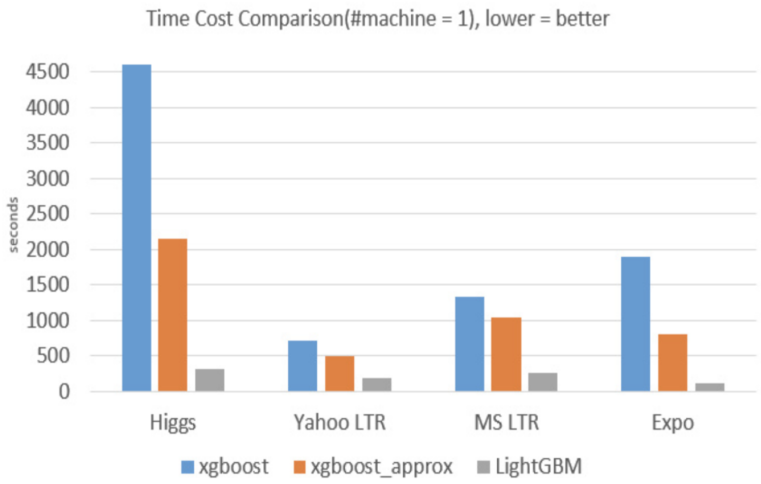
4.3 lightGBM

The Gradient Boosting Decision Tree (GBDT) model is a sustainable model of automatic learning, the main idea of which is to use the weak classifier, the decision tree to practice iteratively in order to obtain the model optimal(Ke et al. (2017)).

The model has a good ripple effect and is not easy to be "overfitting". GBDT is often used for tasks such as click rate prediction and search sorting. The GBDT is also a lethal weapon in various data mining competitions: according to statistics, more than half of the Kaggle champions are based on the GBDT. In this project, we use LightGBM (Light Gradient Boosting Machine), which

is a framework for the implementation of the GBDT algorithm. It supports efficient parallel training and has the following advantages:

- Faster training speed
- Reduced memory consumption
- Better accuracy
- Distributed support for fast processing of large data



(a) Time Cost Comparison

Data	Xgboost	xgboost_approx	LightGBM
Higgs	4.853GB	4.875GB	**0.822GB**
Yahoo LTR	1.907GB	2.221GB	**0.831GB**
MS LTR	5.469GB	5.600GB	**0.745GB**

(b) Space Cost Comparison

Figure 9: lightGBM(Ke et al. (2017))

5 Performance Evaluation

For the evaluation of the performance of the different models, we carry out several metrics such as accuracy, precision, recall, F1-score. These metrics are commonly used in data mining and clinical decision support systems.

- **Accuracy**: It is the ratio of number of correct predictions to the total number of input samples, both positive and negative (sum of true positives and true negatives divided by the number of all instances in the dataset).
- **Precision**: It is the fraction of the number of true positive predictions to the number of all positive predictions (true positives divided by the sum of true positives and false positives). In our project, precision described what percentage of the time the model predicted **self-reported sleep quality** correctly. Note that precision is also known as positive predictive value.
- **Recall or Sensitivity**: It is the fraction of the number of true positive predictions to the actual number of positive instances in the dataset (true positives divided by the sum of true positives and false negatives). In our project, recall referred to the percentage of the correctly predicted sleep quality to the total number of **self-reported sleep quality** instances in the dataset. Note that recall is also known as true positive rate or sensitivity.
- **F1-Score**: There is usually an inverse relationship between precision and recall. That is, it is possible to increase the precision at the cost of decreasing the recall, or vice versa. Therefore, it is more useful to combine them into a single measure such as F1 score.

5.1 Result

As shown in the table below, the performance of models:

Data models	Accuracy	Precision	Recall	F1Score
Logistic Regression	45.9%	47.1%	39.5%	36.9%
Deep Neural Network	40%	37.8%	35%	33%
Random Forest	46.3%	43.4%	39.3%	37.9%
lightGBM	47.5%	44.1%	41.2%	39.5%

Quality of sleep light/deep

Data models	Accuracy	Precision	Recall	F1Score
Logistic Regression	38.3%	34.0%	34.4%	33.4%
Deep Neural Network	39.3%	37%	35.2%	33.1%
Random Forest	48.7%	48.6%	44.6%	44.8%
lightGBM	43.0%	42.2%	39.8%	40.0%

Quality of sleep short/long

Data models	Accuracy	Precision	Recall	F1Score
Logistic Regression	42.4%	41.8%	40.2%	39.3%
Deep Neural Network	36.2%	33.2%	31.2%	31.8%
Random Forest	40.3%	40.7%	40.3%	40.0%
lightGBM	40.1%	39.6%	39.7%	39.4%

Quality of sleep restless/restful

We found that DNN performed the worst, by analyzing the reasons, we think the main reason is:

- At the level of the model:
  - Neural network structure is too simple(4 hidden layers)

Then on the overall level, we find that other machine learning models did not perform well, for a 3-criteria problem, by random guessing we have a correct rate of 33.3%, however our best model gets a 47.5% accuracy, it's not too far from the random result, and we analyze the reason again, we find the main reason is:

- At the level of the data:
  - The number of samples is too small(2041 examples for training set)
  - Sample distribution is uneven, we have much less zero(bad sleep quality) label comparing to one or two (good sleep quality) label.
  - The sleep quality label is too subjective, it lacks accuracy, for example:
    - For the sleep last night, I felt like I slept very well, so I scored a good sleep quality, but in fact, my real sleep may be poor, vice versa.

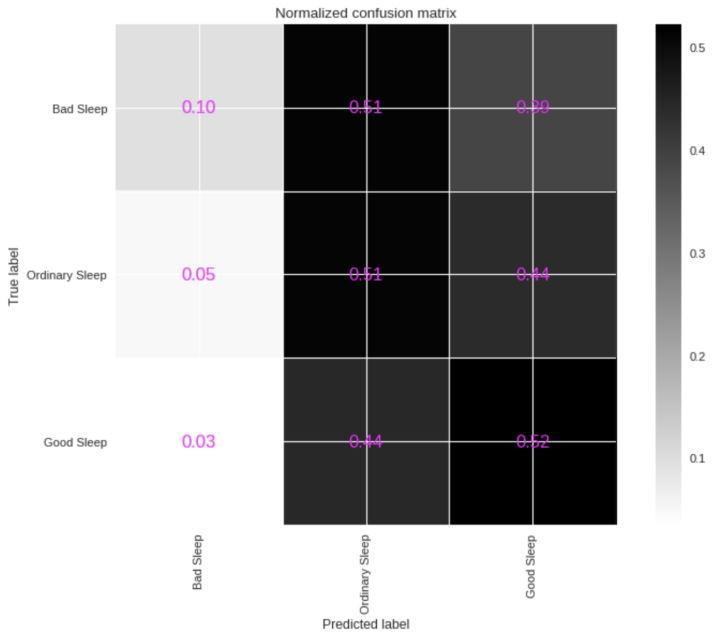


Figure 10: Confusion Matrix for quality of sleep light/deep

In order to improve models and solve data problems, we propose another solution, according to the Likert-style rating, we define sleep quality as five levels, from 0 to 5. On this basis, we do some feature aggregation:

- If the Likert-style rating is less than or equal to 3, we define it as **bad sleep**
- If the Likert-style rating is greater than 3, we define it as **good sleep**

As a result, we convert the problem to a binary classification problem, and now as shown in the table below, we have better result:

Data models	Accuracy	Precision	Recall	F1Score
Logistic Regression	62.4%	56.2%	52.5%	48.6%
Deep Neural Network	57.6%	54.2%	53.2%	53.6%
Random Forest	60.8%	56.6%	54.2%	52.4%
lightGBM	61.8%	58.7%	56.4%	55.5%

Quality of sleep light/deep

Data models	Accuracy	Precision	Recall	F1Score
Logistic Regression	67.3%	64.4%	56.1%	53.8%
Deep Neural Network	61.2%	57.3%	56.6%	56.8%
Random Forest	68.8%	64.6%	57.5%	56.5%
lightGBM	65.3%	57.3%	54.7%	53.9%

Quality of sleep short/long

Data models	Accuracy	Precision	Recall	F1Score
Logistic Regression	63.0%	65.5%	54.1%	48.0%
Deep Neural Network	54.5%	53.1%	51.3%	52.3%
Random Forest	63.8%	58.4%	55.0%	53.3%
lightGBM	62.4%	57.9%	55.3%	54.1%

Quality of sleep restless/restful

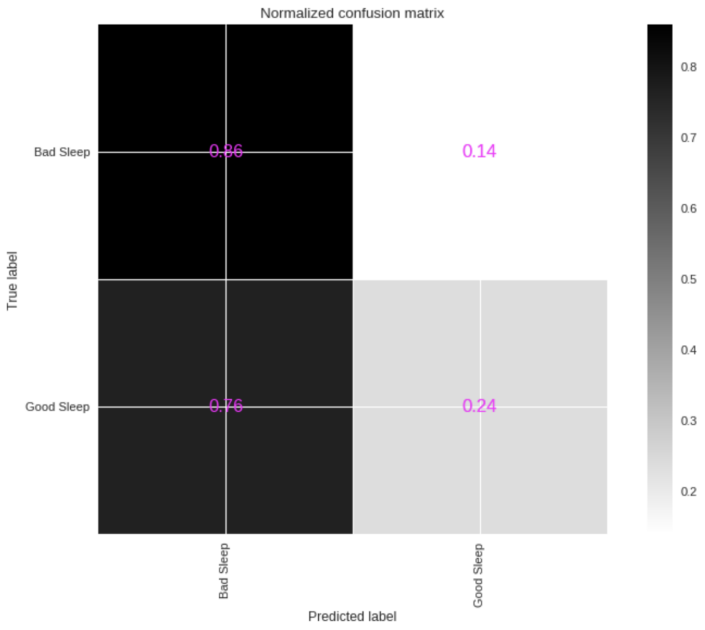


Figure 11: Confusion Matrix for quality of sleep light/deep

6 Conclusions

References

Bengio, Y. et al. (2009). Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127.

Deng, L., Yu, D., et al. (2014). Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Hibbs, A. M., Storfer-Isser, A., Rosen, C., Ievers-Landis, C., M Taveras, E., and Redline, S. (2013). Advanced sleep phase in adolescents born preterm. *Behavioral sleep medicine*, 12.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154.

Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A. C., and Oberye, J. J. L. (2000). Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194.

Kern, S. J. (2017). Automatic sleep stage classification using convolutional neural networks with long short-term memory.

Krystal, A. D. and Edinger, J. D. (2008). Measuring sleep quality. *Sleep medicine*, 9:S10–S17.

Sathyanarayana, A., Joty, S., Fernandez-Luque, L., Ofli, F., Srivastava, J., Elmagarmid, A., Arora, T., and Taheri, S. (2016). Sleep quality prediction from wearable data using deep learning. *JMIR mHealth and uHealth*, 4(4).

Silva, G. E., Goodwin, J. L., Sherrill, D. L., Arnold, J. L., Bootzin, R. R., Smith, T., Walsleben, J. A., Baldwin, C. M., and Quan, S. F. (2007). Relationship between reported and measured sleep times: the sleep heart health study (shhs). *Journal of Clinical Sleep Medicine*, 3(06):622–630.

Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008.

Tsinalis, O., Matthews, P. M., and Guo, Y. (2016). Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Annals of biomedical engineering*, 44(5):1587–1597.

Wang, J.-Y., Weber, F. D., Zinke, K., Inostroza, M., and Born, J. (2018). More effective consolidation of episodic long-term memory in children than adultsunrelated to sleep. *Child development*, 89(5):1720–1734.