Sleep Stage Classification For Medical Purposes: Machine Learning Evaluation For Imbalanced Data

Bens Pardamean (bdsrc@binus.edu)
Bina Nusantara University https://orcid.org/0000-0002-7404-9005
Arif Budiarto
Bina Nusantara University
Bharuno Mahesworo
Bina Nusantara University
Alam Ahmad Hidayat
Bina Nusantara University
Digdo Sudigyo
Bina Nusantara University

Research Article

Keywords: Sleep Stage, Classification, Machine Learning, Imbalanced Data, Neural Network

Posted Date: January 5th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1208553/v1

Abstract

Background: Sleep is commonly associated with physical and mental health status. Sleep quality can be determined from the dynamic of sleep stages during the night. Data from the wearable device can potentially be used as predictors to classify the sleep stage. Robust Machine Learning (ML) model is needed to learn the pattern within wearable data to be associated with the sleep-wake classification, especially to handle the imbalanced proportion between wake and sleep stages. In this study, we incorporated a publicy available dataset consists of three features captured from a consumer wearable device and the labelled sleep stages from a polysomnogram. We implemented Random Forest, Support Vector Machine , Extreme Gradiet Boosting Tree, Densed Neural Network (DNN), and Long Short-Term Memory (LSTM), complemented by three strategies to handle the imbalanced data problem.

Results: In total, we included more than 24,815 rows of preprocessed data from 31 samples. The proportion of minority-majority data is 1:10. In classifying this extreme imbalanced data, the DNN model was found to have the best performance compared to the previous best model, which is based on basic Multi-Layer Perceptron. Our best model successfully achieved a 12% higher specificity score (prediction score for minority class) and 1% improvement on the sensitivity score (prediction score for majority class) by including all features in the model. This achievement was affected by the implementation of custom class weight and oversampling strategy. In contrast, when we only used two features, XGB achieved a specificity improvement only by 1%, while keeping the sensitivity at the same level.

Conclusions: The non-linear operation within the DNN model could successfully learn the hidden pattern from the combination of three features. Additionally, the class weight parameter avoided the model ignoring the minority class by giving more weight for this class in the loss function. The feature engineering process seemed to obscure the time-series characteristics within the data. This is why LSTM, as one of the best methods for time-series data, failed to perform well in this classification task.

Background

Sleep is an important part of the daily routine for all ages to nourish the mind, restore fatigue, and strengthen almost every system in the body. Adequate and good quality sleep at the right time is essentially as important for survival as food and water. Our understanding of sleep patterns is required to maintain the quality of the body in carrying out daily activities and avoid chronic health problems. Lack of sleep may lead to bad behavior and lower performance. Furthermore, long-term effects of sleep deprivation may lead to the development of obesity, diabetes, hypertension, heart disease, and stroke that can contribute, in the long term, to premature death (1). People who have poor sleep quality are often recognized to experience irregular sleep-wake patterns.

The gold standard for a sleep quality analysis is to measure and observe sleep patterns using a Polysomnogram (PSG) that also requires various psychological parameters (2). However, polysomnogram only measures longitudinal ambulatory sleep for one to two nights assessment. In

addition to using a polysomnogram, another FDA-approved method such as actigraphy also measures longitudinal ambulatory sleep. Actigraphy utilizes a wearable accelerometry device to estimate sleep quality based on users' movement activities. The usage of wearable devices with actigraphy is convenient for users to evaluate sleeping habits without the need for complicated sleep laboratory equipment (3, 4). However, this method is still expensive compared to other sleep tracking technologies, which is the main drawback of actigraphy for use in personalized sleep monitoring. In addition, relying on the movement's observation during sleep, actigraphy is considered difficult to accurately determine the time to wake up during the patient's or user's sleep period (5). Therefore, the limitations of the system integration between health data recording platforms and actigraphy need to be addressed using more sophisticated yet affordable methods to evaluate sleep quality accurately. A validation of sleep tracking data to monitor users' sleep quality in some studies indicates a clinical utility of commercial wearable devices for such purposes (6).

Commercial wearables that have their algorithmic method for tracking ambulatory sleep have advantages such as affordable prices, high availability in the market, and high capabilities for their system integration with various health-oriented platforms. However, evaluating commercial sleep tracking data problems compared with polysomnograms as the gold standard for measuring ambulatory sleep cannot be utilized for medical approval in certain clinical research cases (7, 8). The algorithm has implemented in this commercial wearable is a company secret and is rarely published for research from the production side, which causes clinical evaluation problems. Interestingly, this problem becomes a challenge for researchers to investigate and develop an effective and accurate method for measuring longitudinal outpatient sleep, which is implemented in commercial wearable devices (9, 10).

The development of technology and sensors is manageable for researchers to analyze massively integrated data sets. Many studies that use commercial wearable devices to measure sleep quality use Microelectromechanical Systems (MEMS) for data acquisition. This accelerometer may gather acceleration signal data before the program processes data (11, 12) Researchers conveniently access data for use in analyzing the algorithms they've developed. Researchers employ Photoplethysmography (PPG) in wearable devices to quantify sleep quality by monitoring levels of blood volume, in addition to MEMS. PPG uses an optical technique that can also accurately measure heart rhythm through changes in blood volume. The FDA has approved the use of PPG in clinical trials for evaluating abnormal heart rhythms in wearable commercial devices (13–15). The use of these two sensor technologies makes it easier to predict sleep metrics from signal data.

The capabilities of commercial wearable devices for sleep quality prediction have been suggested in sleep studies. For example, a recent study by Miller et al. demonstrated comparable performances of a commercial wearable device with research-grade actigraphy and polysomnography to estimate sleep-wake classification and sleep stages (16). Furthermore, machine learning-based methods have shown promising results for more accurate and flexible data modelling, including sleep quantification by utilizing multi-modal data from wearable sensors (17). A study by Walch et al. employed acceleration data and heart rate obtained from commercial wearable devices to perform Machine Learning (ML)-based sleep

stage prediction (18). They established a multi-classification task for categorizing sleep into several sleep stages. Logistic regression, K-nearest neighbors, random forest, and Multi Layer Perceptron (MLP) were proposed in the study and different features inputs from the data were also compared. The MLP model performed better with the highest accuracy for both classification tasks by incorporating all features inputs. This study also provides a publicly available dataset to be used by other researchers to build a more robust sleep stage classification model.

Furthermore, the robustness of deep learning models for wearable time-series data to accurately predict sleep quality has been explored extensively. An earlier study in 2015 employed an Long Short-Term Memory (LSTM) model to recognize sleep-wake state and offset-onset classification using multimodal data (actigraphy and skin-related data from wrist sensors and daily smartphone activities) (19). The proposed method had the highest classification accuracy and F1 scores when compared with non-temporal models (MLP, logistic regression, and support vector machine). Moreover, Sathyanarayana et al. used actigraphy devices to measure physical activity data during the awaken time and the sleep time for binary sleep efficiency classification (20). The study investigated predictive performances of deep learning models (MLP, CNN), and variants of RNN models). They found that the time batched version of LSTM achieved the highest evaluation AUC score but fares slightly poorer than the CNN model that had the higher F1 and accuracy among all models.

More advanced variants of deep learning architectures and feature engineering have also been proposed in many studies about supervised learning tasks of quantification of sleep using wearable data. For example, a bidirectional LSTM architecture was proposed for sleep stage categorization by learning multi-level features heart rate and actigraphy data (21). Chen et al. showed that crafting features from Heart Rate Variability (HRV) and acceleration features learned using Local Feature-based Long Short-Term Memory (LF-LSTM) to build an ensemble learning model can boost the performance of sleep-wake classification (22).

Inspired by these previous works, which suggests the opportunity to build an advanced ML-based sleep monitoring tool, we proposed various ML methods to be implemented specifically to the dataset from Walch et al (18). They highlighted the nature of imbalanced data within their dataset that significantly affects the classifiers' performance, especially for the binary classification task (i.e., Wake vs Sleep). Therefore, we compared multiple ML methods with three different approaches for handling imbalanced data problems that were aimed to increase the predictive capability for the minority class (wake class) while keeping the majority class prediction score to be the same or even higher.

Results

In total there were 17 different models, from five main methods, implemented in this study. Each method consists of four variations: (1) standard model, (2) model with custom class weight; (3) model with under-sampling approach; (4) model with the oversampling approach. The customized variants were not implemented in the LSTM model because of the different input data formats. Each variant model was

then applied to two different feature sets. The first set only included heart rate and motion count features, while the other set included all features.

The performances for all models are shown in Table 1 and Table 2. Among all proposed models and feature set scenarios, the Densed Neural Network (DNN) model complemented by custom class weight and oversampling strategy was the best classifier. It achieved an 8.5% balanced accuracy improvement from the best model in the previous study, from 73.5–82%. More specifically, this model performed well in predicting the minority class, where it was the main problem in the previous study. The specificity score of our best model is 16% higher than the previous best model, while also slightly improve the sensitivity by 1%.

Table 1
Model performance comparison. Feature Set: Heart Rate (HR) and Motion Count

Method	Variants	Accuracy	Specificity	Sensitivity	Balanced accuracy
MLP	Previous study	90%	41%	95%	68.0%
RF	Previous study	90%	39%	95%	67.0%
RF	Standard	94%	23%	99%	61.0%
	cw: {0:4.6, 1:1.1}	91%	41%	95%	68.0%
	cw: {0:1.09, 1:1.1} under sampling	91%	40%	95%	67.5%
	cw: {0:1.09, 1:1.65} over sampling	91%	41%	95%	68.0%
SVM	Standard	93%	28%	98%	63.0%
	cw: {0:3.8, 1:1}	90%	40%	94%	67.0%
	cw: {0:1.8, 1:1} under sampling	90%	41%	94%	67.5%
	cw: {0:.78, 1:1.1} over sampling	90%	41%	94%	67.5%
XGB	Standard	94%	23%	99%	61.0%
	spw: [0.31]*	91%	42%	95%	68.5%
	spw: [0.725] under sampling	91%	41%	95%	68.0%
	spw: 1.65 over sampling	91%	41%	95%	68.0%
DNN	standard	94%	24%	99%	61.5%
	cw: {0: 5, 1: 1}	91%	39%	95%	67.0%
	cw: {0: 1.16, 1: 1} under sampling	91%	38%	95%	66.5%
	cw: {0: 3.3, 1: 1.1} over sampling	91%	41%	95%	68.0%
LSTM	cw: {0:4, 1:1}	91%	33%	95%	64.0%

cw = class weight parameter

spw = scale pos weight parameter

* Best proposed model in 2-features set

Table 2
Model performance comparison. Feature Set: HR, Motion Count, and Circadian Clock

Method	Variants	Accuracy	Specificity	Sensitivity	Balanced accuracy
MLP	Previous study	91%	52%	95%	73.5%
RF	Previous study	91%	51%	95%	73.0%
RF	Standard	95%	42%	99%	70.5%
	cw: {0: 4.5, 1: 1}	95%	57%	98%	77.5%
	cw: {0: 3.2, 1: 1} under sampling	93%	60%	96%	78.0%
	cw: {0:1.09, 1:1.65} over sampling	93%	62%	95%	78.5%
SVM	Standard	95%	44%	99%	71.5%
	cw: {0:2, 1:1}	94%	56%	97%	76.5%
	cw: {0: 2.7, 1: 1} under sampling	93%	60%	95%	77.5%
	cw: {0:1, 1:1.7} over sampling	93%	48%	96%	72.0%
XGB	Standard	95%	47%	99%	73.0%
	spw: [0.07]	94%	63%	96%	79.5%
	spw: [0.35] under sampling	93%	64%	95%	79.5%
	spw: [1.8] over sampling	93%	64%	95%	79.5%
DNN	Standard	95%	51%	98%	74.5%
	cw: {0: 2.5, 1: 1}	94%	67%	96%	81.5%
	cw: {0: 3.2, 1: 1.2} under sampling	93%	66%	95%	80.5%
	cw: {0: 1., 1: 1.8} over sampling *	94%	68%	96%	82.0%
LSTM	cw: {0:3.3, 1:1}	93%	48%	96%	72.0%

cw = class weight parameter

spw = scale pos weight parameter

* Best proposed model for the whole experiment

Our best model consists of 6 dense layers with 128 neurons, except for the output layer. In total, 66,818 parameters were trained in this model for 40 epochs. The full architecture of this model is depicted in Fig. 1. The model was optimized using Adadelta optimizer,(23) with a static 0.01 learning rate. Binary cross-entropy was used as the loss function, and as the metric evaluation during training. The training process only lasted for 80 seconds.

In the two features scenario, our XGB model outperformed the previous best model in this scenario. The class weight parameter was the only hyperparameter that was applied to the model without under-sampling and over-sampling methods. However, the improvement was not quite impressive with only a 1% increase in the specificity score. It is seen that in each method, the performance was boosted by the application of class weight to handle imbalanced data. However, the implementation of under sampling and over sampling strategies did not consistently yield better performance.

Discussion

In this binary classification task, all the models with three inputs successfully outperformed the models with the same methods that only used two features. This finding is in accordance with the previous study outcome. It indicates that the circadian clock feature, which was modelled from the ambulatory data from each sample, gave a significant booster to help the models in learning the hidden pattern within the data. This particular feature represents the routine biological cycle of each sample and indirectly provides unique information in regard to the sample's sleep habits. By only collecting the seven-days ambulatory data before the laboratory observation, this feature could complement the other two superficial features to categorize the sleep stages. The method to generate this feature by looking at the samples' daily activity data (especially step count) shows a promising strategy to infer the routine cycle of this person. Furthermore, since step count data was commonly captured in the majority of consumer wearable devices, this strategy can be implemented easily so that the captured data can give more benefits to the users, not only related to the sleep stages but also other medical-related information, such as disease or disorders early screening. By knowing this information, we will have enough confidence to infer that a certain condition that is deviates from our routine circadian clock is something worth paying attention to.

We successfully implemented various advanced ML to recognize hidden patterns from the preprocessed features from wearable-based data in relation to sleep habits. We included a complex ML method to reduce the gap from the previous study. This gap is related to the model capability in learning from extreme imbalanced data. In our main referred study, MLP was the best model that can classify 91% of sleep epochs into the correct class. However, this high accuracy was slightly disappointing in this case since the predictive capability for the minority class was quite low, just slightly higher than 50%.

XGB models consistently outperformed other conventional ML models in both feature set combinations. RF model could not perform well even though based on a similar basic method to the XGB model. This difference outcome was the result of a distinct ensemble learning strategy from both methods. Random Forest performs voting mechanisms from several decision trees to get the final predicted class. In contrast, XGB uses a slightly more advanced strategy by stacking multiple weak decision trees to improve the prediction performance. Yet, this model could not outperform the DNN model as our best model, except for the 2-feature set where the XGB model scored a slight improvement on the specificity.

In our best model, assigning class weights and applying SMOTE oversampling approach was found to be effective in addressing the imbalanced data problem when using all features as predictors. The same strategy could not achieve similar success when using only two features. This result was strong evidence to say that the circadian clock could offer the powerful predictive capability to complement heart rate and motion count. Additionally, heart rate and motion count were found to have a stronger correlation than between heart rate and circadian clock or motion count and circadian clock as illustrated in Fig. 2. This correlation may lead to the low performance of all models when only use these two features.

This study was not without any limitations. Firstly, we only included the preprocessed features to be fed to the models. It potentially obscures or removes the essential information from the raw data. Our proposed models were scoped to the ML method with relatively low computational cost, yet still, have enough capability to learn the data. The consideration was because the model was intended to be implemented in a mobile or web-based application, where the wearable users can directly use this model to analyses their sleep routine.

LSTM model, as the most common model for time series data also failed to learn the training data. The possibility is that the time series information within the data was disguised as the result of the feature engineering process. Despite this low performance, LSTM still can be a promising option, if we can use the raw data from the device and formulate it into a neat multivariate time series data, then this LSTM model can potentially yield a higher score for both specificity and sensitivity, as reflected in some previous works from other domains (24). An additional variant of the Nural Network (NN)-based method, called CNN, can also be implemented on top of LSTM layers to perform convolution operations among the features over the time steps. The combination of CNN and LSTM has been proven to have good predictive power in time-series prediction problems (25).

Feature importance analysis is also one angle that can be explored in our next study. The black-box characteristic of the deep NN-based model hinders the mechanism of the model in creating a prediction. In a clinical setting, this analysis is very important to know which feature is the most influential, so that a precision treatment or action can be carried out to the user.

Conclusions

In the present study, we proposed alternative classification models to the previous best model for classifying sleep stages. We limited this classification task to a binary problem. Additionally, we also

focused on addressing the imbalance data problem in this task. Among all proposed models, the DNN model was our best classifier with significant improvement from the previous best model. This finding indicates the potential of the NN-based method in handling imbalanced classification, even with limited features. It also opens the opportunities to implement other types of NN-based methods, specifically the ones with the capability to learn time-series information. Our proposed model is relatively computational efficient so that it is possible to be embedded the model into mobile or web-based applications. Ultimately, it can help the wearable device users to take maximum benefits of their data in regards to monitor their health status.

Methods

Dataset

In this research, we used a publicly available dataset consisting of consumer wrist-worn wearable and medical-grade PSG measurements (18) Each subject was asked to wear an Apple Watch to capture the daily activities data for a week. This 1-week session was then followed by a one-night sleep observation in the laboratory. Wrist band data collection was also still conducted during this observation which includes acceleration and heartbeat. In total, 31 subjects were confirmed to have good quality data based on several inclusion and exclusion criteria, such as issues in data transmission, and several sleep disorders.

In this study, we used the processed features that were provided in the previous study. These features are motion count that was derived from acceleration data, HR measurement from Apple Watch, and circadian clock that was calculated from the 1-week ambulatory data. Motion count data was gathered by looking at the fluctuation in the acceleration raw data which can be interpreted as a motion. HR was processed by calculating the standard deviation from the average of each sample's heart rate. This approach was taken to remove the individual heart rate bias because each person has a unique pattern of heart rate depending on age, gender, and other physical characteristics. All these features were aggregated to meet the sleep epoch (30 s) from the PSG data. Each sleep epoch was categorized into 5 different classes, 0 for a wake stage, 1-4 for non- Rapid Eye Movement (REM), and 5 for REM sleep. In total, 24,815 sleep epochs were included in this study.

Sleep stage classification can be considered as outlier detection, or anomaly detection, due to the imbalance data proportion, if we formulate the problem into binary classification. It means that around 90% of the sleep epochs were categorized as sleep class (nonREM and REM). This extreme discrepancy between the minority (wake) and majority (sleep) classes can be seen in Fig. 3. This Figure depicts a huge difference between the majority class and the other. Ignoring this problem may limit the model's performance.

Classification Model

We employed two different types of ML methods. The first one is a group of machine learning methods that are commonly used for tabular data, while the other group is a series of NN-based methods that offer relatively complex algorithms. The first group is also commonly called conventional machine learning. In total, five different supervised classification methods were compared with the best model from the previous study (18). This best model, MLP, is also considered as conventional machine learning, even though based on the neural network basic technique. Support Vecotr Machine (SVM) (26), Random Forest (RF) (18), and Extreme Gradient Boosting Tree (XGB) (27) were among the existing methods that were selected to be implemented in this study. These three methods offer a non-linear approach in mapping the input data to its desired output data.

On the other hand, NN-based models can be differentiated based on their hidden layer types. The first model is developed by stacking multiple DNN layers to perform a non-linear operation on the data. Although, a single neuron in each layer merely just performs a simple linear regression. Each layer was also complemented by an activation function to select which information can be passed from one neuron to another neuron. We used Rectified Linear Unit (ReLU) as the activation function in all dense layers, except the output layer (28). This last layer, which consists of 2 neurons that represent the number of classes (sleep and wake), was complemented by a Softmax function to generate a probability of a sample belonging to a certain class. While ReLU always generates a number between zero and infinite number as can be seen in Fig. 4, Softmax will only generate a number within a range of 0 and 1 using a formula in Equation 1. In addition to this DNN model, we also developed a model using LSTM layer to take into account the time series characteristic within the data.

$$\sigma\left(\vec{z}\right)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

1

Where:

σ = softmax

$\vec{z}$ = input vector

e^{z_i} = standard exponential function for input vector

K = number of classes in the multi-class classifier, 2 for binary classification

e^{z_j} = standard exponential function for output vector

Prior to train the models, the entire dataset was split into training, validation, and testing subset with the proportion of 60%, 20%, and 20%, respectively. Each model was trained using the training set and validated using the validation set. Eventually, the trained model was evaluated on the test subset to measure the performance in the prediction of the new data that has not to be shown during the training.

To keep the data order in each sample, the data split was done manually by splitting the entire data based on the sample id. It prevents the data from being shuffled which can consequently break the temporal information within the data.

Hyperparameters tuning was done for each model to boost its performance. This tuning was applied specifically only on the training subset. Each model has specific parameters that need to be tuned. RF and XGB have similar tunable parameters since these two methods are based on a decision tree as the main technique for ensemble learning.

Handling Imbalance Data

The main challenge in this classification task was the extreme imbalance of data between wake and sleep. The proportion between these two classes is more than 10% for the whole dataset. The summary visualization of stage proportion in each sample can be seen in Fig. 3. This data proportion is a normal condition in certain topics such as anomaly detection. The imbalance between these two groups causes the typical model to ignore the minority group and consider it as noise.

Consequently, the model accuracy shows a spectacular result, with a clear disparity between specificity and sensitivity. The specificity, in this case, is the count of the correct wake predictions, while the sensitivity is the count of the correct sleep predictions, as illustrated in Table 3. Based on this problem formulation, the main objective in this study was to increase the specificity while keeping a sensitivity score. We applied two strategies of handling this imbalance data by adding weights for each class and performing under a sampling approach to the training data.

Table 3
Confusion matrix

	Predicted Wake	**Predicted Sleep**
Wake	True Negative (TN)	False Negative (FN)
Sleep	False Positive (FP)	True Positive (TP)

In the first approach, the basic intuition was to limit the loss function when calculating the error for the majority class. In contrast, it will give a booster to the minority score, so that the model will predict more on the minority group. We applied different class weights for each model as can be seen in Table 1 and Table 2. In complement to the class weight approach, we also applied a sample-based approach which aimed to balance the amount of data between two classes. To achieve this, we applied two strategies by reducing the amount of sleep class and adding synthetic wake data based on the existing data distribution. These strategies were aimed to balance the proportion of the two classes that can avoid the model only focusing on the majority class. This sample-based approach was not applied to the RNN model since it contradicts the objective of the model that emphasizes the temporal characteristics of the data. In the under-sampling approach, we reduced 50% of the majority class, while in the other strategy

we added augmented data into minority class as much as 50% of the total data in the majority class using the implementation of the Synthetic Minority Over-Sampling Technique (SMOTE) method (29).

Data Evaluation

To measure the performance of each proposed model we calculate five scores, namely accuracy, specificity, sensitivity, and balanced accuracy. These scores are based on the number of correct and incorrect predictions for each class from the confusion matrix as can be seen in Table 3, where the formulas for obtaining these scores are:

- Accuracy = (TP+TN)/(TP+FP+FN+TN)
- Specificity = TN/(TN+FP)
- Sensitivity = TP/(TP+FN)
- Balanced Accuracy = (Specificity+Sensitivity)/2

In a typical binary classification with balance data, accuracy will be the main performance score to be evaluated. However, in imbalance data, this score alone cannot determine the overall performance of the model for both classes. As an illustration, using the dataset in this study, the number of data is 2,5481; 2,152; 23,329, for whole data, wake data and sleep data, respectively. If the model predicts all data as sleep class, then it achieves the accuracy of 91.55% (similar accuracy score for the best model in the previous study). On the other hand, the specificity is zero, which indicates that the model ignores the minority class. Therefore, we focused on the improvement of the specificity score compared to the previous model. At the same time, we also tried to maintain a sensitivity score at least has the same score as the previous best model. The combination of these two scores can be summarized into one score called the balanced accuracy score. This metric has been proven to be an effective scoring in evaluating the ML model for imbalanced data.(30, 31)

All the model training was done in Python using SKLearn, XGBoost, and Keras library for RF and SVM, XGB, and NN-based models, respectively. The hyperparameter tuning was helped by using the grid search function from the SKLearn library. All the plots were generated using the Matplotlib and Seaborn libraries. The computational operations were performed in a LINUX-based Personal Computer with i5-8 cores Central Processing Unit and GeForce RTX 2060 Graphical Processing Unit.

Abbreviations

DNN: Densed Neural Network

FN: False Negative

FP: False Positive

HR: Heart Rate

HRV: Heart Rate Variability

LF-LSTM: Local Feature-based Long Short-Term Memory

LSTM: Long Short-Term Memory

MEMS: Microelectromechanical Systems

ML: Machine Learning

MLP: Multi-Layer Perceptron

NN: Neural Network

PC: Personal Computer

PPG: Photoplethysmography

PSG: Polysomnogram

ReLU: Rectified Linear Unit

REM: Rapid Eye Movement

RF: Random Forest

SVM: Support Vector Machine

SMOTE: Synthetic Minority Over-Sampling Technique

TN: True Negative

TP: True Positive

XGB: Extreme Gradient Boosting Tree

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the PhysioNet repository: https://doi.org/10.13026/hmhs-py35.

Competing interests

The authors declare that they have no competing interests.

Funding

None.

Authors' contributions

BP and AB formulated the research idea with input from BM, AAH, and DS. AB and BM administered and verified the data, and all authors had full access to the data. AB and BM conducted the data analysis. All authors contributed to the interpretation of the analysis. All authors contributed to writing the first draft of the manuscript. All authors approved the final version and were involved in the decision to submit the manuscript for publication.

Acknowledgements

None.

References

1. Medic G, Wille M, Hemels MEH. Short-and long-term health consequences of sleep disruption. Nat Sci Sleep. 2017;9:151.
2. Berry RB, Brooks R, Gamaldo CE, Harding SM, Marcus C, Vaughn BV, et al. The AASM manual for the scoring of sleep and associated events. Rules, Terminol Tech Specif Darien, Illinois, Am Acad Sleep Med. 2012;176:2012.
3. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep. 2003;26(3):342–92.
4. Smith MT, McCrae CS, Cheung J, Martin JL, Harrod CG, Heald JL, et al. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an American Academy of Sleep Medicine clinical practice guideline. J Clin Sleep Med. 2018;14(7):1231–7.
5. Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen JM, Solet JM, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep. 2013;36(11):1747–55.
6. de Souza L, Benedito-Silva AA, Pires MLN, Poyares D, Tufik S, Calil HM. Further validation of actigraphy for sleep studies. Sleep. 2003;26(1):81–5.
7. Budiarto A, Febriana T, Suparyanto T, Caraka RE, Pardamean B. Health Assistant Wearable-Based Data Science System Model: A Pilot Study. In: 2018 International Conference on Information

Management and Technology (ICIMTech). IEEE; 2018. p. 438–42.

8. Budiarto A, Pardamean B, Caraka RE. Computer vision-based visitor study as a decision support system for museum. In: Proceedings - 2017 International Conference on Innovative and Creative Information Technology: Computational Intelligence and IoT, ICITech 2017. 2018. p. 1–6.
9. Pardamean B, Soeparno H, Mahesworo B, Budiarto A, Baurley J. Comparing the Accuracy of Multiple Commercial Wearable Devices: A Method. Procedia Comput Sci. 2019 Jan 1;157:567–72.
10. Baron KG, Duffecy J, Berendsen MA, Mason IC, Lattie EG, Manalo NC. Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep. Sleep Med Rev. 2018;40:151–9.
11. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. Br J Sports Med. 2014;48(13):1019–23.
12. Goldstone A, Baker FC, de Zambotti M. Actigraphy in the digital health revolution: still asleep? Sleep. 2018;41(9):zsy120.
13. Spierer DK, Rosen Z, Litman LL, Fujii K. Validation of photoplethysmography as a method to detect heart rate during rest and exercise. J Med Eng \& Technol. 2015;39(5):264–71.
14. Fonseca P, Weysen T, Goelema MS, Møst EIS, Radha M, Lunsingh Scheurleer C, et al. Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults. Sleep. 2017;40(7).
15. Gottlieb S. Statement from FDA Commissioner Scott Gottlieb, MD, and Center for Devices and Radiological Health Director Jeff Shuren, MD, JD, on Agency Efforts to Work with Tech Industry to Spur Innovation in Digital Health. FDA Statement; Sept. 2018. p. 12.
16. Miller DJ, Roach GD, Lastella M, Scanlan AT, Bellenger CR, Halson SL, et al. A Validation Study of a Commercial Wearable Device to Automatically Detect and Estimate Sleep. Biosensors. 2021;11(6):185.
17. Perez-Pozuelo I, Zhai B, Palotti J, Mall R, Aupetit M, Garcia-Gomez JM, et al. The future of sleep health: a data-driven revolution in sleep science and medicine. npj Digit Med. 2020;3(1):42.
18. Walch O, Huang Y, Forger D, Goldstein C. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. Sleep. 2019 Dec 1;42(12).
19. Sano A, Chen W, Lopez-Martinez D, Taylor S, Picard RW. Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks. IEEE J Biomed Heal Informatics. 2019;23(4):1607–17.
20. Sathyanarayana A, Joty S, Fernandez-Luque L, Ofli F, Srivastava J, Elmagarmid A, et al. Sleep quality prediction from wearable data using deep learning. JMIR mHealth uHealth. 2016;4(4):e125.
21. Zhang X, Kou W, Chang EI-C, Gao H, Fan Y, Xu Y. Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device. Comput Biol Med. 2018;103:71–81.

22. Chen Z, Wu M, Gao K, Wu J, Ding J, Zeng Z, et al. A Novel Ensemble Deep Learning Approach for Sleep-Wake Detection Using Heart Rate Variability and Acceleration. IEEE Trans Emerg Top Comput Intell. 2020;1–10.
23. Zeiler MD. ADADELTA: An Adaptive Learning Rate Method. 2012 Dec 22.
24. Gao J, Zhang H, Lu P, Wang Z. An Effective LSTM Recurrent Network to Detect Arrhythmia on Imbalanced ECG Dataset. J Healthc Eng. 2019;2019.
25. Xie H, Zhang L, Lim CP. Evolving CNN-LSTM Models for Time Series Prediction Using Enhanced Grey Wolf Optimizer. IEEE Access. 2020;8:161519–41.
26. Noble WS. What is a support vector machine? Nat Biotechnol 2006 2412. 2006 Dec;24(12):1565–7.
27. Chen T, Guestrin C. XGBoost:. A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min.
28. Agarap AF. Deep Learning using Rectified Linear Units (ReLU). 2018 Mar 22.
29. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002 Jun 1;16:321–57.
30. Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, et al. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet Epidemiol. 2007 May 1;31(4):306–15.
31. Korkmaz S. Deep learning-based imbalanced data classification for drug discovery. J Chem Inf Model. 2020 Sep;28(9):4180–90. 60(.

Figures

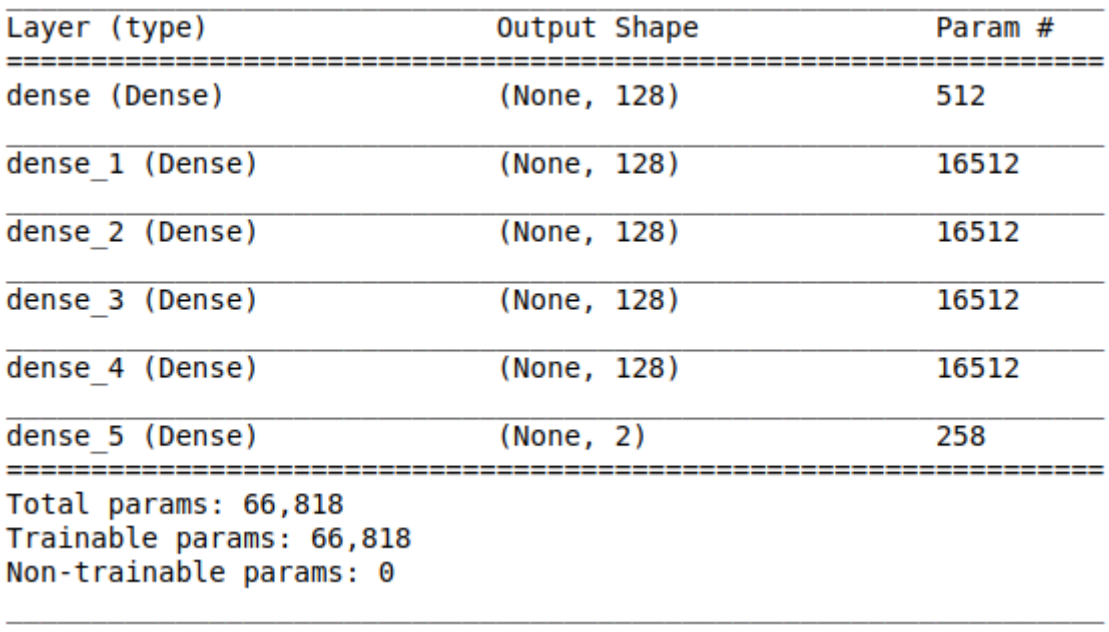

Figure 1

Best model architecture

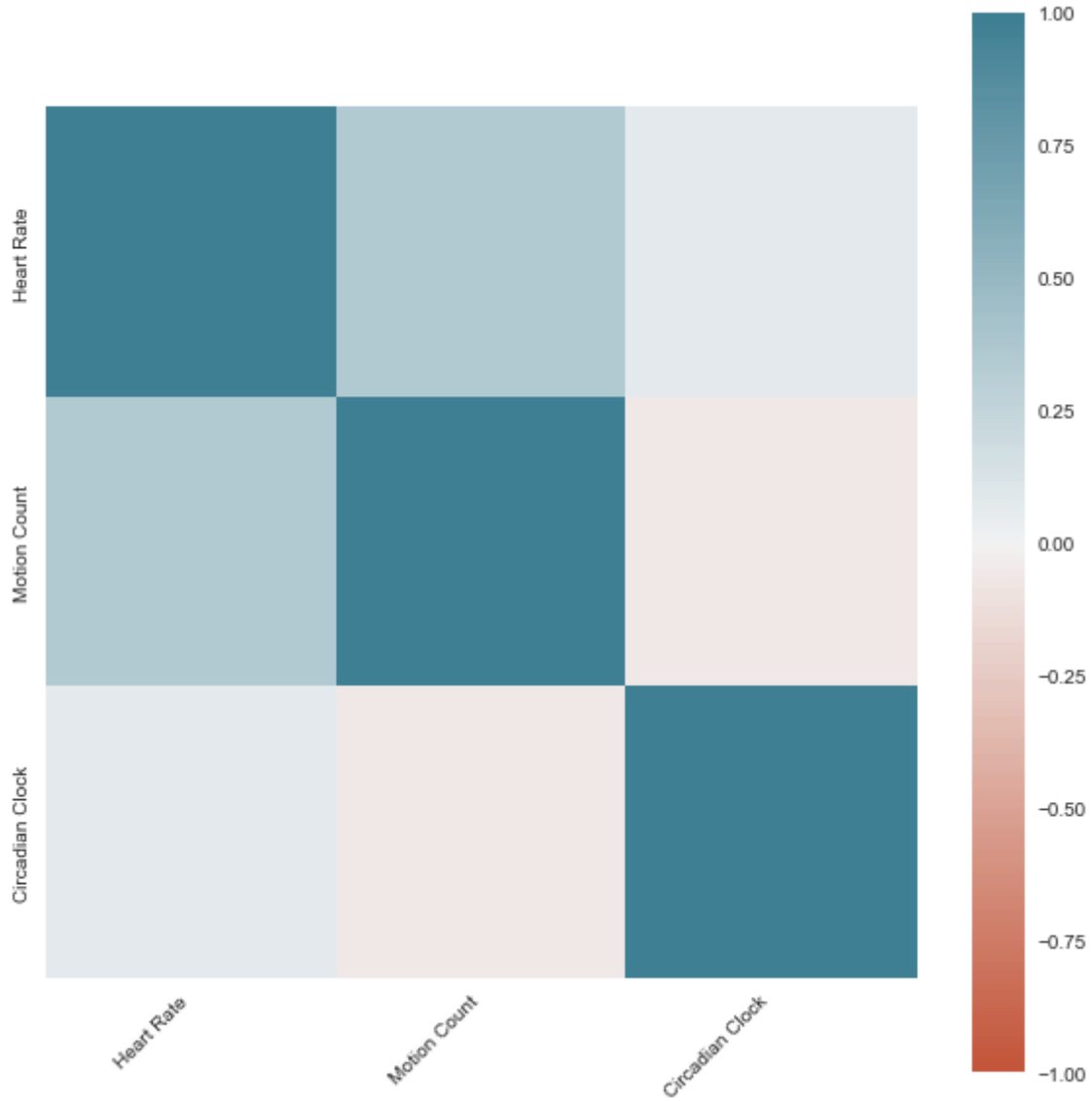

Figure 2

Pairwise correlation of all features

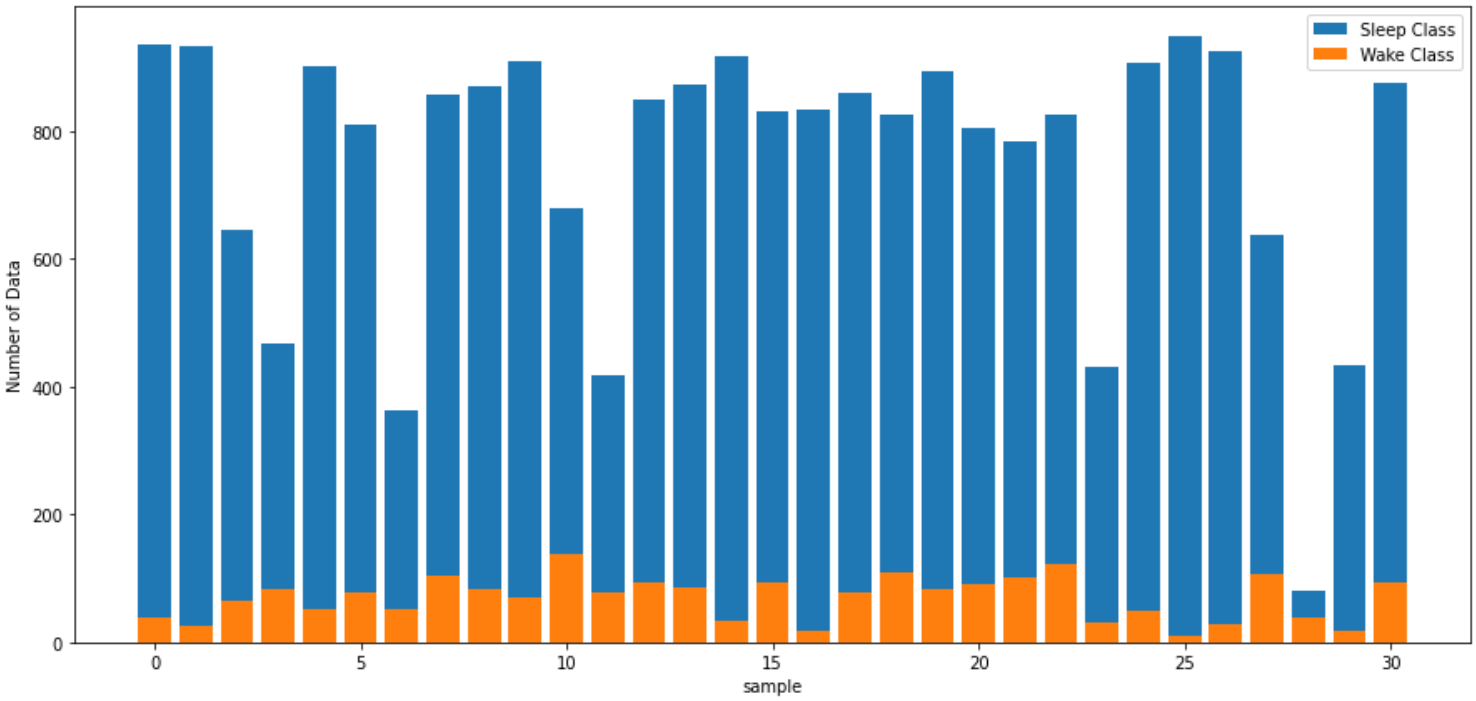

Figure 3

The proportion of sleep and wake classes in the dataset

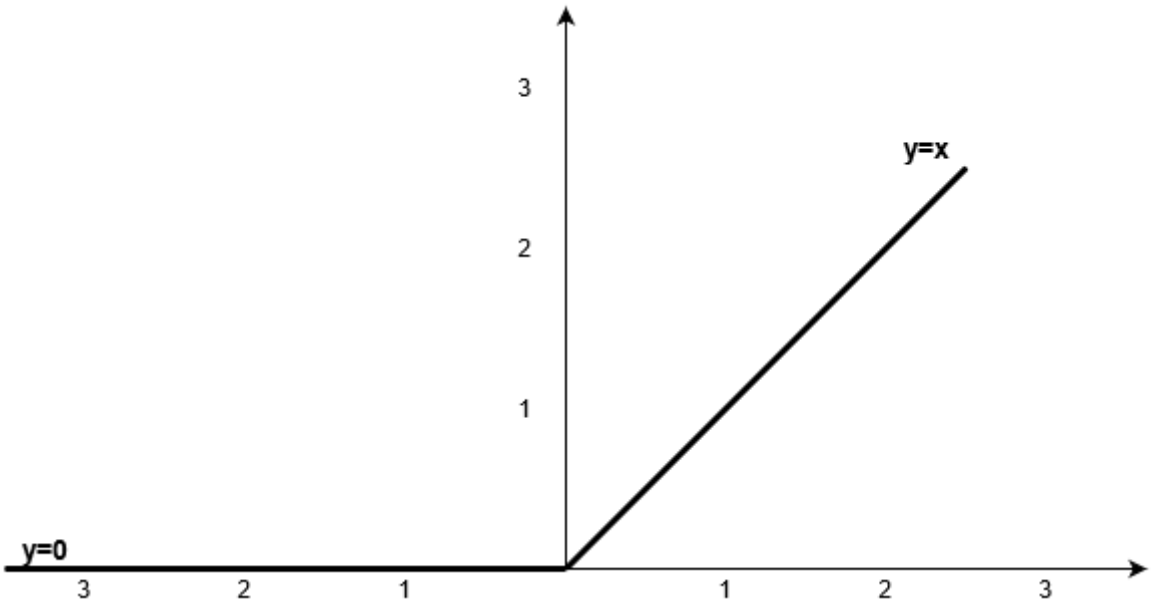

Figure 4

ReLU activation function