Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2023, 2023:131
https://doi.org/10.28919/cmbn/8297
ISSN: 2052-2541

SUPERVISED LEARNING FOR IMBALANCE SLEEP STAGE CLASSIFICATION PROBLEM

BENS PARDAMEAN[1,2,*], ARIF BUDIARTO[1,3], BHARUNO MAHESWORO[1,4], ALAM AHMAD HIDAYAT[1,5], DIGDO SUDIGYO[1]

[1]Bioinformatics & Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia

[2]Computer Science Department, BINUS Graduate Program - Master of Computer Science Program, Bina Nusantara University, Jakarta 11480, Indonesia

[3]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

[4]Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

[5]Mathematics Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Abstract: Sleep is commonly associated with physical and mental health status. Monitoring sleep quality from the dynamic of sleep stages during the night can be valuable. Data from the wearable device has the potential to be used as predictors to predict the sleep stage. Machine learning methods have been proposed to learn patterns within the data for the sleep-wake classification. The main challenge is the nature of imbalanced sleep, which means more sleep stages will be found in the data than in the wake stages. In this study, we utilized five different supervised methods complemented by three strategies to handle the imbalanced data problem. We implemented Random Forest, Support Vector Machine, XGBoost, Dense Neural Network (DNN), and Long-Short Term Memory (LSTM), to a publicly available dataset that consists of three features captured from a consumer wearable device and the labelled sleep stages. Among all the models, the DNN method was found to have the best performance, achieving a 12% higher specificity score (predictive capability for minority class) while using all features in the model. This achievement was affected

[*]Corresponding author
E-mail address: bdsrc@binus.edu
Received October 22, 2023

by the implementation of custom class weight and SMOTE oversampling strategy. 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. Our proposed method therefore can provide an insight into constructing more robust ML-based sleep quality prediction pipelines.

Keywords: classification; data imbalance; machine learning; sleep quality; wearable.

2020 AMS Subject Classification: 92B25.

1. INTRODUCTION

Our understanding of sleep patterns is required to maintain the quality of the body in carrying out daily activities and avoid chronic health problems. 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 [1]. However, a polysomnogram only measures longitudinal ambulatory sleep for one to two nights of assessment. Another FDA-approved method such as actigraphy also measures longitudinal ambulatory sleep. Actigraphy utilizes a wearable accelerometer device to estimate sleep quality based on users' movement activities. The use of actigraphy is convenient for evaluating sleeping habits without the need for complicated sleep laboratory equipment [2,3]. 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 [4–7]. 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 the clinical utility of commercial wearable devices for such purposes [8].

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 [9,10]. The algorithm 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 [11–13].

Many studies that use commercial wearable devices to measure sleep quality rely on microelectromechanical systems (MEMS) for data acquisition. This accelerometer may gather acceleration signal data before the program processes the data [14,15]. On the other hand, photoplethysmography (PPG) in wearable devices was also used in some studies to quantify sleep quality. 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 [16,17]. 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 [18]. 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 [19]. A study by Walch et al. employed acceleration data and heart rate obtained from commercial wearable devices to perform machine learning-based sleep stage prediction [20]. They established a multi-classification task for categorizing sleep into several sleep stages using various ML models and found that Multi Layer Perceptron (MLP) model performed better with the highest accuracy for classification tasks by incorporating all feature inputs.

Further, 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 a Long Short Term Memories (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) [21]. The proposed method had the highest classification accuracy and F1 scores when compared with non-temporal models. Moreover, Sathyanarayana et al. used actigraphy devices to measure physical activity data during the awakening time and the sleep time for binary sleep efficiency classification [22]. They found that the time-batched version of LSTM achieved the highest evaluation AUC score but fares slightly poorer than the CNN model and had the higher F1 and accuracy among all models.

Further, more advanced variants of deep learning architectures and feature engineering have been also proposed for sleep prediction tasks. For example, a bidirectional LSTM architecture was proposed for sleep stage categorization by learning multi-level features heart rate, and actigraphy data [23]. Chen et al. showed that crafting features from HRV and acceleration features learned using local feature-based LSTM (LF-LSTM) to build an ensemble learning model can boost the performance of sleep-wake classification [24]. Another variant of RNNs such as the CNN-LSTM model along with heart rate variability (HRV) and actigraphy data demonstrated accurate scoring of sleep quality prediction [25]. Additionally, using a transfer learning strategy, Phan et al. proposed a sequence-to-sequence neural network called SeqSleepNet trained on a public dataset to predict subject-specific sleep scoring [26], which is a suitable method to deploy in personalized wearable devices.

Inspired by these previous works, we proposed various ML methods to be implemented specifically to the sleep dataset from Walch et al. [20]. 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 at increasing the predictive capability for the minority class (wake class) in the binary classification task.

2. METHODS

2.1. Dataset. In this research, we used a publicly available dataset consisting of consumer wrist-worn wearable and medical-grade polysomnography (PSG) measurements [20]. Each subject was asked to wear an Apple Watch to capture the daily activity data for a week. This one-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, which was derived from acceleration data, heart rate (HR) measurement from Apple Watch, and circadian clock calculated from the 1-week ambulatory data. Motion count data was gathered by employing 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 PPG 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.

Sleep stage classification can be considered as outlier 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 (non-REM and REM). This extreme discrepancy between the minority (wake) and majority (sleep) classes can be seen in Figure 1. The figure depicts a huge difference between the majority class and the other. Ignoring this problem may limit the model's performance.

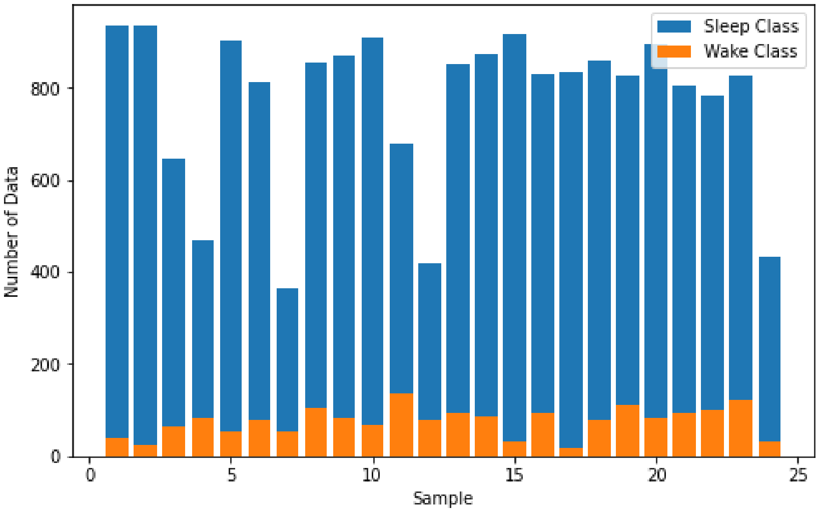


Fig 1. The Proportion of Sleep and Wake Classes in the Dataset.

2.2. Classification Models. We employed two different types of machine learning (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 neural network (NN)-based methods that offer relatively complex algorithms. In total, five different supervised classification methods were compared with the best model from the previous study [20]. This best model, Multi-Layer Perceptron (MLP), is also considered conventional machine learning. Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB) were among the existing methods that were selected to be implemented in this study because of their proven performance in previous classification tasks, especially for tabular data. These three methods offer a non-linear approach to 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 dense neural network layers to perform a non-linear operation on the data. However, a single neuron in each layer merely 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. 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.

Before training the models, the entire dataset was split into training, validation, and testing subsets with the proportions of 60%, 20%, and 20%, respectively. The model trained and validated on the training and validation sets was evaluated on the test subset to measure the performance in the prediction of the testing set. To keep the data order in each sample, the entire data was manually split based on the sample ID. It avoids the data being shuffled which can consequently break the temporal information within the data.

Hyperparameter tuning was done for each model to boost its performance. This tuning was applied specifically only to the training subset. RF and XGB have similar tuneable parameters since these two methods are based on a decision tree as the main technique for ensemble learning.

2.3. 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 Figure 1. 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. Based on this problem formulation, the main objective of this study was to increase the specificity while keeping a sensitivity score. We applied two strategies for handling this imbalanced data by adding weights for each class and performing under a sampling approach to the training data.

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. 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 reducing the amount of sleep class and adding synthetic wake data based on the existing data distribution. These strategies were aimed at balancing the proportion of the two classes which 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 and 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 the minority class as much as 50% of the total data in the majority class.

2.4. 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, where the formulas for obtaining these scores are given in Table 1.

Table 1. Evaluation Metrics

Definition	**Formula**
Accuracy	(TP+TN)/(TP+FP+FN+TN)
Specificity	TN/(TN+FP)
Sensitivity	TP/(TP+FN)
Balanced Accuracy	(Specificity+Sensitivity)/2

In our imbalanced data scenario, accuracy cannot be the only metric to determine the overall performance of the model for both classes. As an illustration, using the dataset in this study, the number of data is 25481, 2152, and 23329, for whole data, wake data, and sleep data, respectively. If the model predicts all data as sleep class, then it achieves an accuracy of 91.55% (a 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 of at least 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.

Table 2. Model Performance Comparison (Heart Rate and Motion Count)

Method	Variants	Accuracy	Specificity	Sensitivity	Binary 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

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 SKlearn. All the plots were generated using the Matplotlib and Seaborn libraries. The computational operations were performed in a LINUX-based portable Personal Computer (PC) with an i5 8 cores CPU and GeForce RTX 2060 GPU.

3. Results

Table 3. Model Performance Comparison (Heart Rate, Motion Count, and Circadian Clock).

Method	Variants	Accuracy	Specificity	Sensitivity	Binary 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

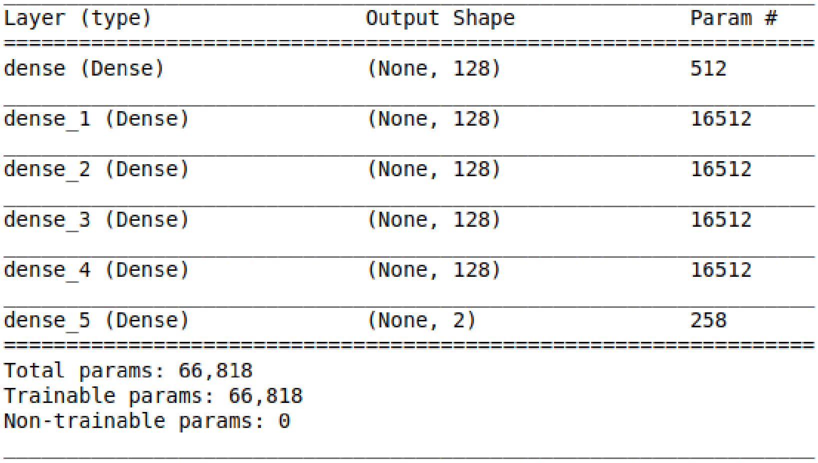
spw = scale pos weight parameter

** best proposed model*

In total, there were 17 different models, from five methods in this study. Each method consists of four variations: (1) standard model, (2) model with custom class weight; (3) model with under-sampling approach; and (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 2 and Table 3. Among all proposed models and feature set scenarios, the 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% to 82%. More specifically, this model performed well in predicting the minority class, which 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 improving the sensitivity by 1%.

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 Figure 2. The model was optimized using an Adadelta optimizer [27], 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.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 128)               512
_________________________________________________________________
dense_1 (Dense)              (None, 128)               16512
_________________________________________________________________
dense_2 (Dense)              (None, 128)               16512
_________________________________________________________________
dense_3 (Dense)              (None, 128)               16512
_________________________________________________________________
dense_4 (Dense)              (None, 128)               16512
_________________________________________________________________
dense_5 (Dense)              (None, 2)                 258
=================================================================
Total params: 66,818
Trainable params: 66,818
Non-trainable params: 0
_________________________________________________________________
```

Fig 2. The Best Model Architecture.

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.

4. 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 [28,29]. 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 feature represents the routine biological cycle of each sample and indirectly provides unique information concerning the sample's sleep habits. By only collecting the seven-day 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 most 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 [30]. By knowing this information as the basis, we will have enough confidence to infer that a certain condition that is different from our circadian clock is something worth paying attention to. In general, sleep stage classification can be very useful to be implemented in the medical realm to help clinicians understand the health condition of the patient. While consumer wearable devices or fitness trackers have been becoming very ubiquitous, an Artificial Intelligence (AI) program is needed to automatically learn the data and inform the users and their clinicians regarding any health issues. However, it cannot replace clinicians to decide the action to address those issues.

Our study was aimed towards this goal by building an effective sleep stage classification as the starting point. We successfully implemented various advanced ML to recognize hidden patterns from the engineered 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's capability to learn from extremely 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 (wake stage) was quite low, just slightly higher than 50%.

XGBoost 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 different 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.

In our best model, assigning class weights and applying the 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 Figure 3. This correlation may lead to the low performance of all models in this feature set scenario.

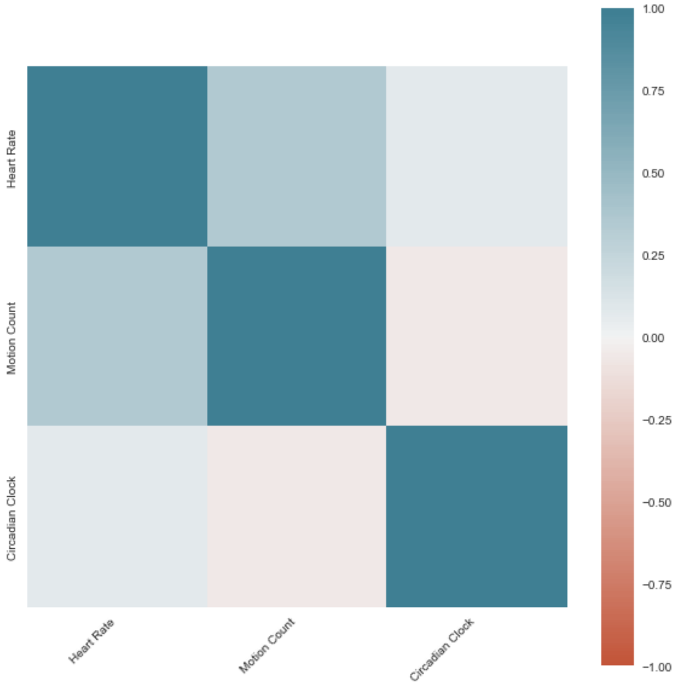


Fig 3. Pairwise Correlation of All Features.

The 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 [31]. An additional variant of the NN-based method, called Convolutional Neural Network (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 [32].

5. Conclusion

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. In total, there were 17 variant models from 4 different ML methods that were successfully implemented in two scenarios based on the number of features included. Three different strategies for handling imbalanced data were also applied to boost the performance of the models. Model performance was measured by looking at the specificity and sensitivity scores, which means the capability to correctly classify wake and sleep classes, respectively. Among all the models, the XGB model with additional class weight assignment was the best in both scenarios. In the 3-features scenario, this model achieved a tremendous improvement, by achieving a specificity score of 21% more than the previous best model plus a 1% improvement in the sensitivity score. NN-based models, as the most advanced ML method, could not achieve a good performance. It was mainly caused by the data used in this study that has been engineered from the original raw data from the wearable device. This feature engineering process seemed to obscure the temporal information within the data. This study can be extended by implementing an NN-based model into raw data to get more benefits from its time series characteristics [33–35]. Moreover, it is suggested that the prediction models need to be deployed in a robust information health system to realize its implementation for a wide use of sleep and mental health research [36].

Conflict of Interests

The author(s) declare that there is no conflict of interest.

References

[1] R.B. Berry, R. Brooks, C. Gamaldo, et al. The AASM manual for the scoring of sleep and associated events, Rules, Terminology and Technical Specification Version 2.4. 2017.

[2] S. Ancoli-Israel, R. Cole, C. Alessi, et al. The role of actigraphy in the study of sleep and circadian rhythms,

Sleep. 26 (2003), 342-392. https://doi.org/10.1093/sleep/26.3.342.

[3] M.T. Smith, C.S. McCrae, J. Cheung, et al. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an american academy of sleep medicine systematic review, meta-analysis, and grade assessment, J. Clinic. Sleep Med. 14 (2018), 1209-1230. https://doi.org/10.5664/jcsm.7228.

[4] M.L. Blood, R.L. Sack, D.C. Percy, et al. A comparison of sleep detection by wrist actigraphy, behavioral response, and polysomnography, Sleep. 20 (1997), 388-95. https://doi.org/10.1093/sleep/20.6.388.

[5] J. Paquet, A. Kawinska, J. Carrier, Wake detection capacity of actigraphy during sleep, Sleep. 30 (2007), 1362–1369. https://doi.org/10.1093/sleep/30.10.1362.

[6] J.M. Lee, W. Byun, A. Keill, et al. Comparison of wearable trackers' ability to estimate sleep, Int. J. Environ. Res. Public Health. 15 (2018), 1265. https://doi.org/10.3390/ijerph15061265.

[7] M. Marino, Y. Li, M.N. Rueschman, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography, Sleep. 36 (2013), 1747-1755. https://doi.org/10.5665/sleep.3142.

[8] L. de Souza, A.A. Benedito-Silva, M.L.N. Pires, et al. Further validation of actigraphy for sleep studies, Sleep. 26 (2003), 81-85. https://doi.org/10.1093/sleep/26.1.81.

[9] A. Budiarto, T. Febriana, T. Suparyanto, et al. Health assistant wearable-based data science system model: a pilot study, in: 2018 International Conference on Information Management and Technology (ICIMTech), IEEE, Jakarta, 2018: pp. 438-442. https://doi.org/10.1109/ICIMTech.2018.8528102.

[10] R.E. Caraka, N.T. Nugroho, S.K. Tai, et al. Feature importance of the aortic anatomy on endovascular aneurysm repair (EVAR) using Boruta and Bayesian MCMC, Commun. Math. Biol. Neurosci. 2020 (2020), 22. https://doi.org/10.28919/cmbn/4584.

[11] B. Pardamean, H. Soeparno, B. Mahesworo, et al. Comparing the accuracy of multiple commercial wearable devices: a method, Procedia Computer Sci. 157 (2019), 567-572. https://doi.org/10.1016/j.procs.2019.09.015.

[12] K.G. Baron, J. Duffecy, M.A. Berendsen, et al. Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep, Sleep Med. Rev. 40 (2018), 151-159. https://doi.org/10.1016/j.smrv.2017.12.002.

[13] M.A. Hamza, A.H. Abdalla Hashim, H. Alsolai, et al. Wearables-assisted smart health monitoring for sleep quality prediction using optimal deep learning, Sustainability. 15 (2023), 1084. https://doi.org/10.3390/su15021084.

[14] R.P. Troiano, J.J. McClain, R.J. Brychta, et al. Evolution of accelerometer methods for physical activity research, Br. J. Sports Med. 48 (2014), 1019-1023. https://doi.org/10.1136/bjsports-2014-093546.

[15] A. Goldstone, F.C. Baker, M. de Zambotti, Actigraphy in the digital health revolution: still asleep?, Sleep. 41 (2018), zsy120. https://doi.org/10.1093/sleep/zsy120.

[16] P. Fonseca, T. Weysen, M.S. Goelema, et al. Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults, Sleep. 40 (2017), zsx097. https://doi.org/10.1093/sleep/zsx097.

[17] D.K. Spierer, Z. Rosen, L.L. Litman, et al. Validation of photoplethysmography as a method to detect heart rate during rest and exercise, J. Med. Eng. Technol. 39 (2015), 264-271. https://doi.org/10.3109/03091902.2015.1047536.

[18] D.J. Miller, G.D. Roach, M. Lastella, et al. A validation study of a commercial wearable device to automatically detect and estimate sleep, Biosensors. 11 (2021), 185. https://doi.org/10.3390/bios11060185.

[19] I. Perez-Pozuelo, B. Zhai, J. Palotti, et al. The future of sleep health: a data-driven revolution in sleep science and medicine, Npj Digit. Med. 3 (2020), 42. https://doi.org/10.1038/s41746-020-0244-4.

[20] O. Walch, Y. Huang, D. Forger, et al. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device, Sleep. 42 (2019), zsz180. https://doi.org/10.1093/sleep/zsz180.

[21] A. Sano, W. Chen, D. Lopez-Martinez, et al. Multimodal ambulatory sleep detection using LSTM recurrent neural networks, IEEE J. Biomed. Health Inform. 23 (2019), 1607-1617. https://doi.org/10.1109/jbhi.2018.2867619.

[22] A. Sathyanarayana, S. Joty, L. Fernandez-Luque, et al. Sleep quality prediction from wearable data using deep learning, JMIR mHealth uHealth. 4 (2016), e125. https://doi.org/10.2196/mhealth.6562.

[23] X. Zhang, W. Kou, E.I.C. Chang, et al. Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device, Computers Biol. Med. 103 (2018), 71-81. https://doi.org/10.1016/j.compbiomed.2018.10.010.

[24] Z. Chen, M. Wu, K. Gao, et al. A novel ensemble deep learning approach for sleep-wake detection using heart rate variability and acceleration, IEEE Trans. Emerg. Top. Comput. Intell. 5 (2021), 803-812. https://doi.org/10.1109/tetci.2020.2996943.

[25] S. Haghayegh, S. Khoshnevis, M.H. Smolensky, et al. Deep neural network sleep scoring using combined motion and heart rate variability data, Sensors. 21 (2020), 25. https://doi.org/10.3390/s21010025.

[26] H. Phan, K. Mikkelsen, O.Y. Chén, et al. Personalized automatic sleep staging with single-night data: a pilot study with Kullback-Leibler divergence regularization, Physiol. Measure. 41 (2020), 064004. https://doi.org/10.1088/1361-6579/ab921e.

[27] M.D. Zeiler, ADADELTA: An adaptive learning rate method, preprint, 2012. https://arxiv.org/abs/1212.5701.

[28] A.A. Hidayat, A. Budiarto, B. Pardamean, Long short-term memory-based models for sleep quality prediction from wearable device time series data, Procedia Computer Sci. 227 (2023), 1062–1069.

https://doi.org/10.1016/j.procs.2023.10.616.

[29] B. Mahesworo, A. Budiarto, A.A. Hidayat, et al. Sleep quality and daily activity association assessment from wearable device data, in: 2020 International Conference on Information Management and Technology (ICIMTech), IEEE, Bandung, Indonesia, 2020: pp. 197-202. https://doi.org/10.1109/ICIMTech50083.2020.9211281.

[30] B. Pardamean, H. Soeparno, A. Budiarto, et al. Quantified self-using consumer wearable device: predicting physical and mental health, Healthc. Inform. Res. 26 (2020), 83-92. https://doi.org/10.4258/hir.2020.26.2.83.

[31] J. Gao, H. Zhang, P. Lu, et al. An effective LSTM recurrent network to detect arrhythmia on imbalanced ECG dataset, J. Healthc. Eng. 2019 (2019), 6320651. https://doi.org/10.1155/2019/6320651.

[32] H. Xie, L. Zhang, C.P. Lim, Evolving CNN-LSTM models for time series prediction using enhanced grey wolf optimizer, IEEE Access. 8 (2020), 161519-161541. https://doi.org/10.1109/access.2020.3021527.

[33] K. Purwandari, J.W.C. Sigalingging, T.W. Cenggoro, et al. Multi-class weather forecasting from twitter using machine learning aprroaches, Procedia Computer Sci. 179 (2021), 47-54. https://doi.org/10.1016/j.procs.2020.12.006.

[34] R.E. Caraka, S.A. Bakar, B. Pardamean, A. Budiarto, Hybrid support vector regression in electric load during national holiday season, in: 2017 International Conference on Innovative and Creative Information Technology (ICITech), IEEE, Salatiga, 2017: pp. 1–6. https://doi.org/10.1109/INNOCIT.2017.8319127.

[35] A.A. Hidayat, T.W. Cenggoro, B. Pardamean, Convolutional neural networks for scops owl sound classification, Procedia Computer Sci. 179 (2021), 81-87. https://doi.org/10.1016/j.procs.2020.12.010.

[36] B. Pardamean, W. Gazali, H.H. Muljo, et al. The fundamental variable of stress detection in health information system to measure health worker's current mental health, Int. J. Med. Eng. Inform. 13 (2021), 397. https://doi.org/10.1504/IJMEI.2021.117727.