Sleep Stage Classification from ECG: A Comparative Study of Traditional and Deep Learning Models

Team Members: Ian Lee, Aarushi Bhardwaj

Abstract

This study aims to compare machine learning and deep learning methods for classifying sleep stages using ECG data. By applying both approaches to the same dataset under consistent conditions, we assess their relative performance across key classification metrics. Statistical and wavelet-based features extracted from 30-second ECG epochs were used to train machine learning models, including Random Forest, XGBoost, and Multilayer Perceptron. Concurrently, raw ECG signals were used to train deep learning architectures such as a 1D Convolutional Neural Network (CNN), a CNN with bidirectional LSTM, and a multi-scaled hybrid model integrating CNN, Transformer, and BiLSTM layers. Among all models, Random Forest achieved the highest F1-score (76.5%), highlighting the effectiveness of feature engineering. The CNN+BiLSTM model was the most effective among deep learning approaches, capturing temporal dynamics but exhibiting reduced performance on underrepresented classes. T The results demonstrate the potential of ECG-based sleep staging while emphasizing trade-offs between model complexity, interpretability, and generalization. Future work should explore deeper architectures, multimodal inputs, and improved strategies for handling class imbalance.

Table of Contents

Background

Sleep progresses through several stages: light sleep (N1, N2), deep sleep (N3), and REM (rapid eye movement). These stages cycle during the night, contributing to brain and body restoration. Monitoring them reveals sleep quality and abnormalities, such as disrupted REM or insufficient deep sleep.

Polysomnography (PSG), the clinical gold standard for sleep staging, uses EEG, EOG, and EMG to measure brain, eye, and muscle activity. Although highly accurate, PSG is expensive, time-consuming, and impractical for continuous or large-scale use. Electrocardiography (ECG) offers a more accessible alternative by reflecting autonomic nervous system activity, which changes across sleep stages—heart rate slows during deep sleep and becomes more variable in REM—making ECG a useful indirect signal for sleep analysis.

ECG can be collected non-invasively via clinical monitors or wearables, enabling long-term, at-home monitoring and large-scale sleep studies. Traditional machine learning approaches classify sleep stages by extracting features like heart rate variability and applying models such as random forests or support vector machines. These methods are effective with limited data and easily interpretable.

Deep learning has gained popularity for its end-to-end approach, processing raw ECG data and predicting sleep stages without manual feature extraction. Models like CNNs and LSTMs excel at learning complex patterns and often outperform traditional methods. However, few studies compare these approaches under standardized conditions, highlighting the need for comprehensive evaluations to inform practical deployment.

Existing Work

Machine Learning (ML) Approaches

Study	Subjects	Input	Model	Classes	Performance
Adnane et al. (2010)	—	HRV	SVM-RFE	2	Accuracy: 79.9%
Xiao et al.	45	HRV	Random Forest	3	Accuracy: 72.5%
Ebrahimi et al. (2015)	30	HRV, Resp.	SVM	4	Accuracy: 89.3%
Singh et al. (2016)	20	RR interval	SVM	2	Accuracy: 72.8%
Yücelbaş et al.	10	ECG	Random Forest	3	Accuracy: 78.0%

Fonseca et al. (2015)	48	ECG + RIP	Linear Discriminant Classifier	4	Accuracy: 69%, Kappa: 0.49
Tataraidze et al. (2016)	625	ECG + RIP	—	4	Accuracy: 71.4%, Kappa: 0.57
Fonseca et al. (2017)	342	ECG + RIP	Conditional Random Fields	4	Avg. accuracy: 87.38%, Kappa: 0.41 (varies by class and subject group)
Yoon et al. (2017)	26	ECG	—	2	Accuracy: 87.03%, Kappa: 0.61

Deep Learning (DL) Approaches

Study	**Subjects**	**Input**	**Model**	**Classes**	**Performance**
ECG-SleepNet (2024)	18	ECG	Feature Imitating Networks + KAN	5	Accuracy: 80.79%, Kappa: 0.73
Urtnasan et al. (2022)	112	ECG	DCR (CNN + RNN)	3, 5	74.2% (5-class), 86.4% (3-class)
Pini et al. (2022)	994 (CinC), 52 (Z3Pulse)	ECG (HR)	DL (unspecified)	2, 3, 4	Accuracy: 88% (2-class), Kappa: 0.61 (3-class)
Li et al. (2018)	Multiple datasets	ECG-derived Resp, HRV	CNN + SVM	2, 3, 4	Accuracy: 75.4
Sridhar et al. (2020)	10,000+ nights (SHHS, MESA)	Instantaneous HR	CNN	4	Accuracy: 77%, Kappa: 0.66

Radha et al. (2019)	292	HRV	LSTM	4	Accuracy: 77%, Kappa: 0.61
Sun et al. (2020)	8682	ECG + Respiration	CNN + LSTM	3, 5	Kappa: 0.585 (5-class), 0.76 (3-class)
Wei et al. (2017)	18	ECG	Stacked Autoencoder (DNN)	3	Accuracy: 77%, Kappa: 0.56

Common Challenges

A key limitation of ECG-based sleep staging is that ECG is an indirect measure of sleep. Unlike EEG, which reflects brain activity that defines sleep stages, ECG captures autonomic nervous system responses that only correlate with those stages. This makes it difficult to accurately distinguish stages with subtle differences in autonomic activity, such as N1 and quiet wake. Inter-subject variability in heart rate dynamics, due to age, fitness, medications, or health conditions, further reduces generalizability across populations. ECG signals are also sensitive to noise, motion artifacts, and poor signal quality, particularly in wearable or home-based recordings. These artifacts can distort heart rate variability patterns and lead to incorrect predictions. Additionally, sleep stage distributions are naturally imbalanced, with stages like N1 and REM occurring less frequently, making them harder to detect. Sleep disorders, such as obstructive sleep apnea, introduce abnormal autonomic patterns that can confuse models unless explicitly addressed during training. Overcoming these issues is critical for improving robustness and reliability in real-world applications.

Research Gap

Despite the growing number of studies on ECG-based sleep staging, few offer a direct comparison between traditional machine learning and deep learning approaches on the same dataset. This limits our ability to objectively evaluate performance differences and understand the trade-offs between model complexity, interpretability, and accuracy.

Objectives

The primary objective of this analysis is to train and evaluate multiple machine learning (ML) and deep learning (DL) models on the same dataset to compare their performance across key classification metrics. Specifically, the goal is to:

- Assess precision, recall, F1-score, and accuracy of each model.
- Identify strengths and weaknesses of each algorithm in handling class imbalance and prediction quality.
- Determine which model offers the best overall and class-wise performance.
- Use these findings to guide model selection for further optimization or deployment

Methodology

Datasets

Our study utilized two publicly available datasets to develop and evaluate ECG-based sleep stage classification models:

1. **St. Vincent's University Hospital / University College Dublin Sleep Apnea Database (UCDDB)**

   The UCDDB comprises 25 full overnight polysomnography (PSG) recordings from adult subjects suspected of having sleep-disordered breathing. Each recording includes simultaneous three-channel Holter ECG data. The dataset encompasses a range of subjects (21 males, 4 females) aged between 28 and 68 years, with varying body mass indices (BMI) and apnea-hypopnea indices (AHI). Sleep stages were annotated by experienced technologists following standard Rechtschaffen and Kales rules. Despite its comprehensive data collection, the limited sample size posed challenges for training robust machine learning models.

2. **PhysioNet/Computing in Cardiology Challenge 2018 Dataset**

   This extensive dataset includes 1,985 PSG recordings from subjects monitored at the Massachusetts General Hospital (MGH) sleep laboratory for sleep disorder diagnoses. The recordings feature multiple physiological signals, including EEG, EOG, EMG, ECG, and oxygen saturation (SaO_2), with sleep stages annotated by clinical experts. The dataset was divided into balanced training and test sets, each containing approximately 994 and 989 recordings, respectively. The large sample size and diverse subject pool make this dataset well-suited for developing and validating machine learning models for sleep stage classification.

Initially, we conducted experiments using the UCDDB dataset. However, due to its limited sample size and class imbalance, model performance was suboptimal. Consequently, we transitioned to the PhysioNet Challenge 2018 dataset, for its larger and more diverse sample to enhance model training and evaluation.

From this dataset, we selected a subset of 100 subjects for training, validation, and testing evaluation, balancing the need for representative data with available computational resources. We focused on both deep learning and classical machine learning approaches to compare performance and generalizability.

Feature Extraction

Statistical Feature Extraction

To represent the signal characteristics over time, we extracted a range of statistical features from each 30-second epoch. These included the mean, standard deviation, minimum, maximum, and median values, as well as the 25th and 75th percentiles. Additional measures such as variance and peak-to-peak range were computed to capture the overall variability and amplitude range of the signal. To assess temporal dynamics, the mean and standard deviation of signal differences within each epoch were also computed.

Wavelet-Based Feature Enrichment

We applied the Continuous Wavelet Transform (CWT) to analyze the signal in both time and frequency domains, as it effectively handles non-stationary characteristics by offering multi-resolution insights.

CWT was chosen for its ability to capture localized frequency-specific features, and we specifically selected the Morlet wavelet for its optimal balance between time and frequency localization, achieved through its Gaussian-shaped waveform. Using scales 1 to 31, we extracted the mean and standard deviation of wavelet coefficients to represent scale-specific signal traits. Additionally, signal energy and entropy were computed to quantify the distribution and complexity of energy across frequency bands, allowing for a comprehensive characterization of the signal.

Combined Feature Set for Model Input

The final feature set was constructed by combining the statistical features and the wavelet-based features. This enriched feature representation served as the input to classical machine learning models, aiming to leverage both time-domain and time-frequency characteristics of the signal for improved predictive performance.

Machine Learning Models

To evaluate classification performance, we employed three machine learning models: Random Forest (RF), XGBoost, and Multilayer Perceptron (MLP). The Random Forest model was chosen for its ability to handle high-dimensional feature sets and its robustness against overfitting due to ensemble learning. XGBoost, a gradient boosting algorithm, was selected for its computational efficiency—particularly with GPU acceleration—and its effectiveness in handling class imbalance. Finally, the Multilayer Perceptron provided a simple neural network architecture capable of learning nonlinear relationships within the data.

Data Processing Pipeline

The input to all models consisted of features derived from Heart Rate Variability (HRV) statistics and Continuous Wavelet Transform (CWT) analysis. Prior to training, all features were standardized using StandardScaler to ensure consistent scaling across variables. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training set. The dataset was split into 80% for training and 20% for testing to evaluate model generalizability.

Deep Learning Models

We developed three deep learning architectures to classify sleep stages directly from raw ECG signals. The first was a 1D Convolutional Neural Network (CNN) with three convolutional blocks, each followed by batch normalization and max pooling, designed to extract local temporal features. The second model extended this by incorporating a bidirectional Long Short-Term Memory (BiLSTM) layer after the CNN, allowing the network to capture both short- and long-range dependencies in the signal. The third architecture was a multi-scaled hybrid model that stacked a CNN feature extractor with a Transformer encoder and a BiLSTM layer, followed by a classification head. This design aimed to capture temporal features across multiple scales and apply attention mechanisms to enhance representation learning.

Data Processing Pipeline

The deep learning models were trained directly on raw ECG signals, segmented into 30-second epochs (6000 samples per epoch at 200 Hz). Preprocessing steps included filtering out undefined stages (label 4) and remapping the remaining labels to ensure consistency across stages. To address class imbalance, we applied a weighted loss strategy using compute_class_weight, assigning higher importance to underrepresented classes. The dataset was divided using PyTorch's random_split into 80% training, 10% validation, and 10% testing subsets.

Training Procedure

All models were trained using the AdamW optimizer with weight decay for improved generalization. The loss function used was CrossEntropyLoss, enhanced with class weights and label smoothing (smoothing factor = 0.05) to reduce model overconfidence and improve calibration. This regularization technique helps the model make more robust predictions in imbalanced or noisy settings.

Metrics

The performance of each model was evaluated using a confusion matrix and class-wise accuracy for each sleep stage. Overall accuracy was calculated as the proportion of correctly classified epochs across all stages. To better understand model behavior on imbalanced data, accuracy was also reported separately for each class (e.g., Wake, N1, N2, N3, REM). The confusion matrix provided insight into common misclassifications between similar stages, such as N1 and Wake or REM and N2.

Results

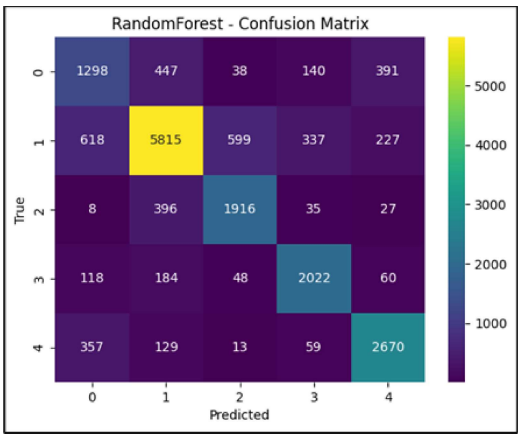
The performance of both classical machine learning and deep learning models was evaluated using precision, recall, and F1-score as primary metrics. These results are summarized in the table below:

Model	Precision	Recall	F1-score
Random Forest (RF)	76.8%	76.4%	76.5%
XGBoost	69.7%	68.6%	68.7%
Multilayer Perceptron	66.9%	62.4%	62.7%
CNN	63.9%	49.8%	50.2%
CNN + BiLSTM	72.1%	68.5%	68.5%
Multi-Scaled Network	69.7%	66.7%	66.8%

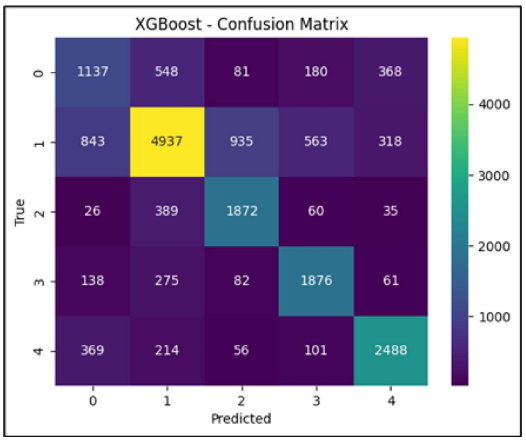
Table 1: Performance metrics of deep learning and machine learning algorithms

Machine Learning Models Confusion Matrices

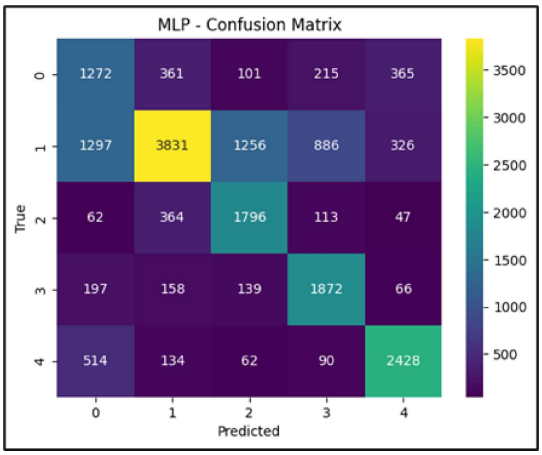
Random Forest



XGBoost

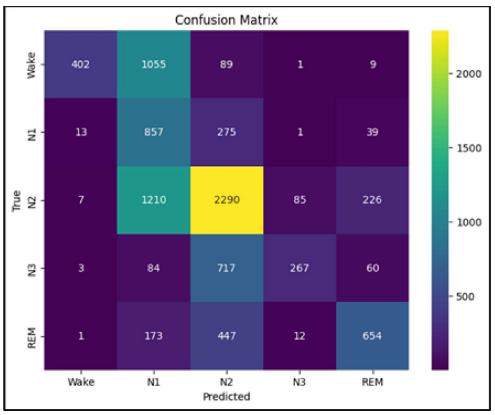


MLP

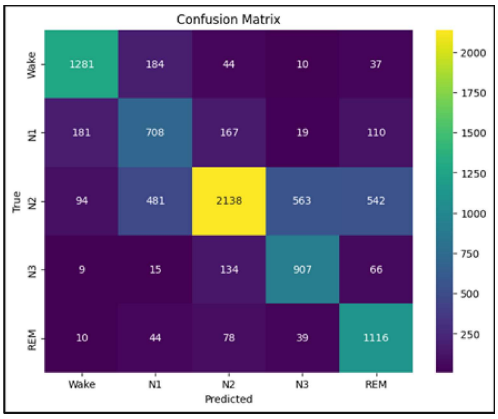


Deep Learning Models Confusion Matrices

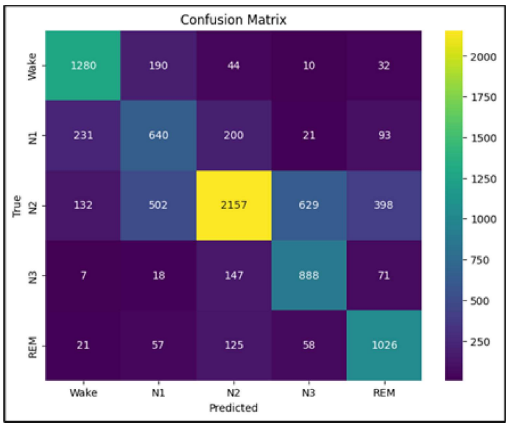
1D CNN



CNN BiLSTM



Stacked Model



Discussion

Machine Learning Models

Among the classical machine learning models, Random Forest (RF) achieved the highest performance with an F1-score of 76.5%, indicating balanced precision and recall across all sleep stages. The confusion matrix shows that RF performed particularly well in identifying stages N2 and N3, with relatively fewer misclassifications across classes. This result highlights the effectiveness of ensemble methods in leveraging the enriched statistical and wavelet-based features.

XGBoost, while slightly lower than RF, also showed solid performance with an F1-score of 68.7%. It misclassified a larger proportion of N1 and REM samples, as evident from the confusion matrix, but maintained competitive performance overall. The Multilayer Perceptron (MLP) had the lowest scores among the ML models, with an F1-score of 62.7%. Its confusion matrix shows difficulty in differentiating between adjacent sleep stages (e.g., N1 vs N2), likely due to the model's limited capacity to capture complex feature interactions.

Deep Learning Models

In the deep learning category, the CNN + BiLSTM model performed best, achieving an F1-score of 68.5%. The confusion matrix shows significant improvement in classifying deeper sleep stages (N3) and REM compared to the basic CNN. The addition of bidirectional LSTMs helped the model learn temporal dependencies, leading to more accurate stage transitions.

The Multi-Scaled Network, which incorporated CNNs, Transformers, and BiLSTMs, followed closely with an F1-score of 66.8%. This architecture successfully balanced temporal modeling and attention mechanisms but only marginally improved over the simpler CNN + BiLSTM model.

The standalone CNN model yielded the lowest F1-score (50.2%) among all models. Its confusion matrix reveals frequent misclassification of N1 and REM as N2, hinting at limited temporal context awareness. This supports the need for sequence-aware layers in sleep staging tasks, particularly where transitions and context are crucial.

General Observations

ML models using handcrafted features (especially RF) outperformed DL models trained on raw ECG, suggesting that statistical and wavelet features are rich and discriminative for this task. Deep learning models struggled with class imbalance and temporal overlap between sleep stages. Despite this, models incorporating sequence information (BiLSTM, Transformer) showed clear improvements. Confusion matrices highlight persistent difficulty distinguishing between lighter sleep stages (Wake, N1, N2), a common challenge in sleep staging tasks.

Conclusions

This study compared classical machine learning and deep learning approaches for sleep stage classification using ECG-derived features and raw signals. Among all models, the Random Forest classifier achieved the highest overall performance, demonstrating the effectiveness of enriched statistical

and wavelet-based features. Deep learning models, particularly the CNN + BiLSTM and Multi-Scaled Network, showed promising results by capturing temporal dependencies in raw ECG data. However, their performance was still slightly lower than feature-based ML models, suggesting that handcrafted features remain highly valuable for this task. Overall, these findings highlight the potential of combining physiological signal processing with both traditional and modern AI techniques for automated sleep staging, and pave the way for future enhancements such as hybrid models, improved handling of class imbalance, and real-time deployment in wearable health technologies.

Limitations and Future Work

Despite promising results, this study faced several limitations, primarily related to computational resources. Training deep learning models on large ECG datasets was highly resource-intensive, with runtimes ranging from 10 to 50 hours depending on the architecture and preprocessing pipeline. Although Google Colab provided access to GPUs, limitations in runtime duration and memory capacity restricted the number of training epochs and subjects. Locally, only one machine was equipped with a capable GPU, further constraining the scope of experimentation. As a result, the analysis was limited to 100 subjects, and input sequences were restricted to 30-second segments. While this aligns with standard sleep scoring protocols, longer input windows (e.g., 60 seconds) could better capture transitions and improve classification performance.

Another challenge lies in the nature of ECG as an indirect signal for sleep staging. Unlike EEG, ECG does not directly reflect neural activity, making it difficult to distinguish between stages with similar autonomic signatures, such as Wake and N1. Classification performance was also hindered by class imbalance and signal noise, especially for underrepresented stages like N1 and REM.

For traditional machine learning models, features were limited to statistical and wavelet-based metrics. Although effective, future work could explore frequency-domain features, nonlinear dynamics (e.g., entropy or fractal measures), and time-frequency coherence to enhance stage discrimination.

To improve deep learning performance, future work should prioritize developing deeper and more expressive model architectures. The CNN+BiLSTM model yielded the best results among the networks tested, but the limited benefit of the Transformer encoder in the multi-scaled model suggests that architectural depth and complexity were insufficient. Designing deeper models with more convolutional layers, stacked recurrent or attention modules, and richer hierarchical representations may help better capture subtle patterns in ECG data. This effort will require enhanced computational resources to support longer training, larger batch sizes, and more extensive hyperparameter tuning for more comprehensive evaluations.

References

Adnane, M., Jiang, Z., & Yan, Z. (2012). Sleep–wake stages classification and sleep efficiency estimation using single-lead electrocardiogram. *Expert Systems with Applications, 39*(1), 1401–1413. https://doi.org/10.1016/j.eswa.2011.08.022

Aghaomidi, P., & Wang, G. (2024). ECG-SleepNet: Deep learning-based comprehensive sleep stage classification using ECG signals. *arXiv*. https://doi.org/10.48550/arXiv.2412.01929

Ebrahimi, F., Setarehdan, S.-K., & Nazeran, H. (2015). Automatic sleep staging by simultaneous analysis of ECG and respiratory signals in long epochs. *Biomedical Signal Processing and Control, 18*, 69–79. https://doi.org/10.1016/j.bspc.2014.12.003

Fonseca, P., Long, X., Radha, M., Haakma, R., & Aarts, R. M. (2015). Sleep stage classification with ECG and respiratory effort. *Physiological Measurement, 36*(10), 2027–2040. https://doi.org/10.1088/0967-3334/36/10/2027

Fonseca, P., et al. (2017). Cardiorespiratory sleep stage detection using conditional random fields. *IEEE Journal of Biomedical and Health Informatics, 21*(4), 956–966. https://doi.org/10.1109/JBHI.2016.2550104

Li, Q., et al. (2018). Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram. *Physiological Measurement, 39*(12), 124005. https://doi.org/10.1088/1361-6579/aaf339

Meng, X., Yan, H., Song, J., Yang, Y., & Yang, X. (2013). Sleep stages classification based on heart rate variability and random forest. *Biomedical Signal Processing and Control, 8*(6), 624–633. https://doi.org/10.1016/j.bspc.2013.06.001

Pini, N., et al. (2022). An automated heart rate-based algorithm for sleep stage classification: Validation using conventional polysomnography and an innovative wearable electrocardiogram device. *Frontiers in Neuroscience, 16*, 974192. https://doi.org/10.3389/fnins.2022.974192

Radha, M., Fonseca, P., Moreau, A., et al. (2019). Sleep stage classification from heart-rate variability using long short-term memory neural networks. *Scientific Reports, 9*, 14149. https://doi.org/10.1038/s41598-019-49703-y

Singh, J., Sharma, R. K., & Gupta, A. K. (2016). A method of REM-NREM sleep distinction using ECG signal for unobtrusive personal monitoring. *Computers in Biology and Medicine, 78*, 138–143. https://doi.org/10.1016/j.compbiomed.2016.09.018

Sridhar, N., Shoeb, A., & Stephens, P., et al. (2020). Deep learning for automated sleep staging using instantaneous heart rate. *npj Digital Medicine, 3*, 106. https://doi.org/10.1038/s41746-020-0291-x

Sun, H., et al. (2020). Sleep staging from electrocardiography and respiration with deep learning. *Sleep, 43*(7), zsz306. https://doi.org/10.1093/sleep/zsz306

Tataraidze, A., et al. (2016). Sleep architecture measurement based on cardiorespiratory parameters. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3478–3481. https://doi.org/10.1109/EMBC.2016.7591477

Urtnasan, E., et al. (2022). Deep convolutional recurrent model for automatic scoring sleep stages based on single-lead ECG signal. *Diagnostics (Basel), 12*(5), 1235. https://doi.org/10.3390/diagnostics12051235

Wei, R., et al. (2017). The research of sleep staging based on single-lead electrocardiogram and deep neural network. *Biomedical Engineering Letters, 8*(1), 87–93. https://doi.org/10.1007/s13534-017-0044-1

Yoon, H., et al. (2017). REM sleep estimation based on autonomic dynamics using R-R intervals. *Physiological Measurement, 38*(4), 631–651. https://doi.org/10.1088/1361-6579/aa63c9

ücelbaş, Ş., Yücelbaş, C., Tezel, G., Özşen, S., & Yosunkaya, S. (2018). Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal. *Expert Systems with Applications, 102*, 193–206. https://doi.org/10.1016/j.eswa.2018.02.034

St. Vincent's University Hospital & University College Dublin. (2000). *St. Vincent's University Hospital / University College Dublin Sleep Apnea Database (UCDDB)* PhysioNet. https://physionet.org/content/ucddb/1.0.0/

Goldberger, A. L., Goldberger, Z. D., & Mark, R. G. (2018). *PhysioNet/Computing in Cardiology Challenge 2018 Dataset* [Dataset]. PhysioNet. https://physionet.org/content/challenge-2018/1.0.0/