Exploring Label Noise Reduction Techniques for Sleep Stage Classification Using Wearable Devices

Maria Yohanne Moreira [3], Vinicius Monteiro de Lira[1,3], Lívia Almada Cruz[2,3], José Antônio Fernandes de Macêdo[1,3]

[1]Department of Computer Science, Federal University of Ceará – Fortaleza – CE – Brazil

[2]Campus Quixadá, Federal University of Ceará – Quixadá – CE – Brazil

[3]Insight Data Science Lab, Federal University of Ceará – CE – Brazil

{yohannemoreira,vinicius.monteiro,livia,jose.macedo}@insightlab.ufc.br

Abstract. *Wearable devices offer a portable alternative to polysomnography for sleep stage classification using accelerometer and photoplethysmography (PPG) data. However, the performance of machine learning models heavily depends on the quality of the reference data. Consequently, the presence of incorrectly labeled data undermines the performance of these models. In this work, we conduct an exploratory analysis of different label noise reduction methods, including duration window and Isolation Forest. We evaluate the impact of these techniques on sleep stage prediction using several machine learning classifiers. Our results provide insights into the effectiveness and characteristics of label noise reduction methods for improving sleep stage classification.*

1. Introduction

Automatic sleep stage classification has become a foundational tool in both clinical research and consumer health applications, driving algorithms that convert physiological recordings into the recognized sleep and wakefulness stages (N1, N2, N3, REM). Supervised machine learning approaches to this problem depend critically on accurately labeled training data [Sekkal et al. 2022], which are most reliably obtained through in-laboratory polysomnography (PSG): PSG is widely recognized as the gold standard for sleep staging, as it provides a comprehensive, multidimensional assessment of sleep architecture. However, even PSG recordings can be affected by noise and artifacts, occasionally resulting in errors or ambiguity in the assigned sleep stage labels. Moreover, signals collected from wrist-worn devices (e.g., smartwatches) can also be impacted by noise and artifacts during data acquisition, which may further compromise the accuracy and reliability of models built using this data in conjunction with PSG-derived labels [Chen et al. 2019, Chriskos et al. 2018, Gunter et al. 2023, Chaparro-Vargas and Cvetkovic 2013, Correa and Leber 2011].

Polysomnography captures a rich array of signals—including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), respiratory effort, airflow, and pulse oximetry—and thus affords the most comprehensive basis for human sleep staging. In practice, however, PSG recordings are vulnerable to multiple noise sources [Metsis et al. 2015]. Biological artifacts may originate from muscle contractions or twitches (EMG), cardiac activity superimposed on the EEG (ECG), irregular breathing patterns or blood-oxygen fluctuations, and spontaneous eye movements (EOG) that mimic or obscure true sleep-related phenomena

[Chaparro-Vargas and Cvetkovic 2013]. Environmental interference—room noise, stray electromagnetic fields, loose or shifting electrodes, and cables—further degrades signal quality. Such contaminations are particularly problematic in transitional sleep stages (e.g., N1 vs. N2, or N2 vs. N3), where subtle changes in EEG spindle density or delta power must be detected against a background of artifact, often leading to inter-scorer variability and ambiguous "borderline" labels.

Wearable systems, by contrast, have become increasingly important for automatic sleep stage classification. These devices collect physiological signals such as raw acceleration and heart rate (via photoplethysmography, PPG), enabling the inference of sleep–wake cycles and the classification of different sleep stages. The widespread adoption of wearables allows for large-scale, continuous monitoring of sleep patterns in naturalistic settings, providing valuable data for both research and health applications [Charlton et al. 2025]. The integration of wearable device data with PSG-based ground truth provides a valuable approach for developing machine learning models for automatic sleep stage classification. Leveraging these rich, longitudinal datasets enables the development of systems that can accurately classify sleep phases outside the laboratory [Birrer et al. 2024].

In this paper, we present a framework designed to handle noise in PSG sleep stage signals, supported by an exploratory analysis of different noise detection and reduction techniques. Our key contributions are: (1) formalizing sleep stage label noise as an outlier detection problem; (2) developing a modular, classifier-agnostic pipeline for iterative noise identification and data cleaning; and (3) evaluating the effectiveness of various techniques on real-world PSG and wearable sensor data.

2. Related Work

As described by [de Zambotti et al. 2024], the gold standard for labeling sleep stages is through polysomnography. However, new proposals for approaches to classify time series of polysomnography signals aim to facilitate access to data that allow the assessment, for example, of sleep quality by detecting patterns in different sleep stages. In this section, we bring together works that present the current state of this research field in relation to the methods already applied and what are the main consolidated contributions and persistent limitations.

Different studies have focused specifically on optimizing signal preprocessing and developing advanced noise reduction techniques. [Kim et al. 2017] conducted a study that examined sleep stage-related features through Heart Rate Variability signals, applying the Empirical Mode Decomposition method to reduce the noise of the DFA sequence (DFAseq). The prediction model achieved 73.6% accuracy, but the limitation to only 13 subjects may restrict its applicability to a wider population.

Similarly focused on optimizing preprocessing, [Wang et al. 2019] used single-channel electroencephalogram signals, applying discrete wavelet adaptive thresholding and Butterworth filtering to improve the signal-to-noise ratio, establishing a stacking-based ensemble learning algorithm that achieved 96.6% accuracy.

[Permana et al. 2023] developed a sleep stage classification system using algorithms such as k-Nearest Neighbors (KNN), Decision Tree, and Random Forest, applying the Synthetic Minority Oversampling (SMOTE) technique to balance the training dataset and overcome the limitations of manual annotations, although they acknowledge that future work could explore more complex Deep Learning architectures.

Some studies have explored innovative approaches for sleep analysis using machine learning and deep learning techniques. [Fedorin and Slyusarenko 2021] present a proposal for a wearable system based on smartwatches for detailed analysis of the physiological state during sleep, seeking to overcome the limitations of traditional PSG through a Bidirectional Neural Network LSTM (Bi-LSTM) model to estimate sleep stages, respiratory events, snoring, and blood oxygen saturation level (SpO_2).

[Phan et al. 2019] proposed a joint classification and prediction framework based on Convolutional Neural Networks (CNNs) by transforming raw signals into preprocessed logarithmic-scale power spectra for frequency smoothing and dimensionality reduction, demonstrating the efficiency of the 1-max CNN model in automatically classifying sleep stages, achieving 81.9$\%$ accuracy and 73.8$\%$ F1-score.

Complementarily, [Huang et al. 2023] specifically addressed classification in children, implementing zero-mean and unit-variance normalization, duration window segmentation, and data augmentation via Gaussian noise injection to balance minority classes. Their DeConvolution- and Self-Attention-based Model (DCSAM) achieved 90.26$\%$ accuracy, although labeling overlapping segments can lead to inter-class classification errors.

2.1. Limitations and opportunities

Therefore, the literature analysis reveals a significant diversity of approaches in sleep classification research, from preprocessing techniques to more complex algorithms, each with specific advantages depending on the signal modality and application context. In our work, exploratory experiments with different label noise reduction methods allow us to better understand the complexity of sleep classification problems with time series from wearable devices. Furthermore, we can understand how these preprocessing techniques influence the performance of machine learning models, although the results indicate that isolated improvements in noise treatment do not guarantee substantial gains in final accuracy.

3. Problem Definition

Automatic sleep stage classification using wearable devices depends heavily on the quality and reliability of labeled training data, which is typically collected from two main sources: wearable sensor devices—such as those capturing raw acceleration and photoplethysmography (PPG) signals—and polysomnography (PSG), the clinical gold standard for sleep monitoring. However, labels derived from both wearables and PSG can be affected by a variety of noise sources. These noise sources can introduce errors and inconsistencies in sleep stage labels, potentially degrading the performance and robustness of automated sleep stage classifiers.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ represent the training dataset, where $x_i \in \mathcal{X}$ is a time series consisting of raw acceleration and photoplethysmography (PPG) heart rate signals, and $y_i \in \mathcal{Y} = \{S_1, S_2, \ldots, S_k\}$ is the corresponding sleep stage label. The objective is to learn a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by θ, that maps the input data x_i to a predicted label $\hat{y}_i$, minimizing the classification error where $\mathcal{L}$ is the loss function (e.g. cross-entropy loss).

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i)$$

Due to the aforementioned sources of noise in both PSG labels and wearable-generated signals, a significant challenge in this task is the presence of noisy or incorrectly labeled data within $\mathcal{D}$. Specifically, let $\tilde{\mathcal{D}} = \{(x_j, y_j)\}_{j=1}^{M} \subset \mathcal{D}$ represent the subset of mislabeled instances, where $y_j \neq y_j^*$ and y_j^* is the unknown ground truth label x_j. Training on such noisy data typically results in degraded model performance, as the learned classifier parameters, leading to:

$$\hat{\theta}_{\text{noisy}} = \arg\min_{\theta} \frac{1}{M} \sum_{j=1}^{M} \mathcal{L}(f_\theta(x_j), y_j)$$

where $\hat{\theta}_{\text{noisy}}$ represents the parameters learned with noisy data, biased by the corrupted labels, which results in higher error and decreased performance. Thus, addressing label noise in $\mathcal{D}$ is essential for robust and accurate model development.

4. Methodology

To mitigate the impact of noisy labels introduced in the previous section, we propose an iterative outlier detection framework grounded in our problem formulation. The process is illustrated in Figure 1, which outlines the key steps: starting with noisy labeled data, applying noise reduction techniques to identify and remove mislabeled or corrupted instances, and finally using the cleaned dataset for model training.

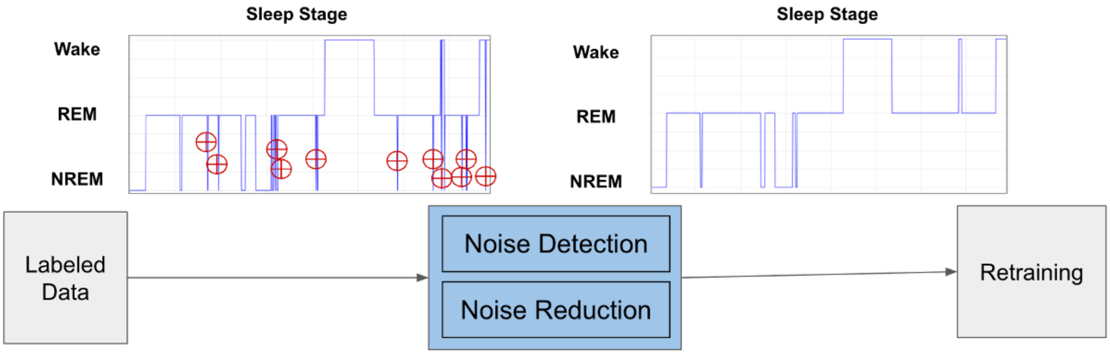

Figure 1. Overview of the proposed noise reduction pipeline. Raw labeled data (which may contain mislabeled instances due to noise from both wearables and PSG) is processed to filter out noisy labels, resulting in a cleaner dataset for classifier training.

Our approach capitalizes on the observation that noisy or mislabeled data ($\tilde{\mathcal{D}}$) can be detrimental to learning, and aims to systematically identify and eliminate such instances to enhance the integrity of the training process, producing a cleaner training set $\mathcal{D}_{\text{clean}}$.

1. **Noise Detection ($\mathcal{I}$):** Identify the noisy subset $\tilde{\mathcal{D}} \subset \mathcal{D}$:

$$\tilde{\mathcal{D}} = \mathcal{I}(\mathcal{D}, f_\theta)$$

2. **Noise Reduction** (R)**:** Refine $\tilde{\mathcal{D}}$ from $\mathcal{D}$ to obtain a cleaner dataset:

$$\mathcal{D}_{\text{clean}} = \mathrm{R}(\mathcal{D}, \tilde{\mathcal{D}})$$

3. **Model Retraining:** Retrain the classifier f_θ on $\mathcal{D}_{\text{clean}}$ to obtain an updated parameter set $\hat{\theta}_{\text{clean}}$ that minimizes the classification error:

$$\hat{\theta}_{\text{clean}} = \arg\min_\theta \frac{1}{|\mathcal{D}_{\text{clean}}|} \sum_{i=1}^{|\mathcal{D}_{\text{clean}}|} \mathcal{L}(f_\theta(x_i), y_i)$$

The objective is to obtain a $\hat{\theta}_{\text{clean}}$ that results in a more accurate and robust model compared to $\hat{\theta}_{\text{noisy}}$, thus improving sleep stage prediction performance.

This procedure can be repeated to further refine the dataset and improve model reliability. By directly confronting the challenge of label noise formulated in Section 3, our method aims to produce a classifier $\hat{\theta}_{\text{clean}}$ that surpasses $\hat{\theta}_{\text{noisy}}$ in both accuracy and robustness, thereby enabling superior sleep stage prediction from wearable sensor data.

5. Experiments

In this section, we present the experimental framework designed to evaluate the effectiveness of different preprocessing strategies for noise reduction in sleep stage data. Multiple supervised learning algorithms were used to assess the impact of these strategies, and their performance results are systematically reported.

5.1. Dataset

The dataset used in this study was originally made available at [Walch et al. 2019]. Participant data were collected in a controlled laboratory environment, where each individual wore an Apple Watch to capture physiological and movement data while simultaneously undergoing polysomnography (PSG) to provide ground truth sleep stage labels. The wearable device collected three types of data: motion (triaxial acceleration in g), heart rate (in beats per minute) and steps (count), while polysomnography provided the labels of the sleep stages used as the target variable. For our experiments, we consider 3 stages: Wake, NREM (i.e. N1, N2, and N3), and REM.

For our experiments, we used data from 29 of the 31 participants, due to issues related to the length of the time series in the data from two participants.

5.2. Data processing and Classification

All analyzed models received the same pre-processed data, following a methodologically consistent approach for handling noisy data, aiming to preserve relevant information during the cleaning process.

Noise Detection ($\mathcal{I}$): Two distinct scenarios were implemented for outlier identification: (1) Isolation Forest: Application of isolation forest algorithm with automatic outlier detection based on accelerometer characteristics (x, y, and z axes), heart rate, and step count. (2) Duration window: A minimum duration threshold was applied to define valid sleep stages. The duration window approach flags as outliers (noise) any stage transitions that last less than this threshold—set to one minute in our experiments.

Noise Reduction ($\mathcal{R}$): Following detection, targeted strategies were applied to address identified noise. (1) For Isolation Forest, all detected outliers were completely removed from the dataset. (2) For the duration window approach, noise labels were handled in two ways: (a) by replacing the outlier label value with the majority stage class present in the considered time window, or (b) by removing the entire row sample identified as an outlier.

Supervised algorithms: Four supervised models were implemented: (1) Random Forest (ensemble of trees), (2) Decision Tree (hierarchical rules), (3) XGBoost (gradient boosting), and (4) LSTM (long short-term memory). Evaluation metrics included accuracy (correct predictions), precision (true positives/positive predictions), recall (true positives identified), and F1-score (harmonic mean of precision and recall), with class-specific analyses (Wake, NREM, REM). For classification, we apply the Leave-One-Out Cross-Validation technique, in which a cross-validation is performed where, for a dataset with n individuals, the model is trained n times, leaving one single individual out at each iteration to be used as a test set.

5.3. Results

The following are the results obtained to evaluate the performance of different classification models after the application of noise reduction techniques. Table 1 presents the performance of different classification models under various noise detection and reduction strategies. The **baseline** for each model is represented by the rows where both the Detection and Reduction columns are set to "None," indicating that no noise handling technique was applied. These baseline results provide a reference point for evaluating the impact of the applied preprocessing strategies. The table also provides class-wise precision values, allowing for a more granular assessment of model performance across different sleep stages.

The results show that, in several cases, the introduction of noise detection and reduction methods led to notable improvements over the baseline. For example, for the Decision Tree model, applying the Isolation Forest for detection and removing outliers increased the overall accuracy from 69.47% (baseline) to 71.26%, and improved the F1-Score from 65.31% to 68.03%. Similarly, the Random Forest model saw an increase in accuracy from 71.36% (baseline) to 72.73% when using Isolation Forest for outlier removal, along with a corresponding improvement in F1-Score. These results demonstrate that appropriate noise handling strategies can enhance the performance of classification models, particularly for Decision Tree and Random Forest classifiers.

6. Conclusion and Future Work

This study demonstrates that addressing label noise through targeted detection and reduction strategies can meaningfully improve the performance of automatic sleep stage classification models. By applying methods such as Isolation Forest and duration window analysis to PSG data, we were able to identify and mitigate the impact of mislabeled or

Table 1. Experiment results obtained for each model

Model	Detection ($\mathcal{I}$)	Reduction ($\mathcal{R}$)	Accuracy	Precision	Recall	F1-Score	Precision by Class		
							Wake	NREM	REM
Decision Tree	None	None	69.47%	65.33%	69.47%	65.31%	67.34%	73.31%	33.44%
	Duration window	Replaced outlier	68.77%	66.00%	68.77%	64.81%	71.51%	73.15%	34.37%
	Isolation Forest	**Removed outlier**	**71.26%**	**67.07%**	**71.26%**	**68.03%**	**7.33%**	**75.58%**	**39.57%**
	Duration window	Removed outlier	68.71%	65.36%	68.71%	64.80%	69.78%	72.90%	33.46%
Random Forest	None	None	71.36%	67.72%	71.36%	66.46%	73.71%	73.40%	41.52%
	Duration window	Replaced outlier	71.28%	67.54%	71.28%	66.02%	74.32%	73.21%	41.82%
	Isolation Forest	**Removed outlier**	**72.73%**	**67.97%**	**72.73%**	**68.19%**	**73.48%**	**73.46%**	**40.84%**
	Duration window	Removed outlier	71.28%	67.84%	71.28%	66.35%	15.90%	74.83%	42.75%
XGBoost	None	None	68.67%	66.46%	68.67%	66.00%	62.62%	74.13%	36.06%
	Duration window	Replaced outlier	68.17%	66.04%	68.17%	65.48%	65.06%	73.99%	33.87%
	Isolation Forest	**Removed outlier**	**70.36%**	**66.72%**	**70.36%**	**67.51%**	**10.04%**	**76.79%**	**37.51%**
	Duration window	Removed outlier	68.15%	65.87%	68.15%	65.44%	62.10%	73.75%	34.85%
LSTM	**None**	**None**	**73.25%**	**72.65%**	**73.36%**	**70.91%**	**58.21%**	**78.02%**	**51.13%**
	Duration window	Replaced outlier	72.80%	71.84%	72.80%	69.99%	57.94%	77.38%	50.58%
	Isolation Forest	Removed outlier	19.29%	38.73%	19.28%	20.26%	10.75%	78.38%	50.73%
	Duration window	Removed outlier	71.34%	70.76%	71.34%	68.70%	59.19%	77.63%	52.68%

noisy samples. However, no single strategy consistently improves all models, as noise reduction effectiveness depends on both the algorithm and data.

While our framework enhances robustness, future work should explore advanced noise detection methods and validate them on larger, more diverse datasets. One possible direction is on explore advanced noise detection methods using deep learning, such as autoencoders or attention-based models, to better identify mislabeled or ambiguous segments in both PSG and wearable data.

Acknowledgments

Part of the results presented in this work were obtained through the project "Center of Excellence in Artificial Intelligence - AI4Wellness", funded by Samsung Eletrônica da Amazonia Ltda., in accordance with the Brazilian Informatics Law No. 8.248/91.

References

Birrer, V., Elgendi, M., Lambercy, O., and Menon, C. (2024). Evaluating reliability in wearable devices for sleep staging. *NPJ Digital Medicine*, 7(1):74.

Chaparro-Vargas, R. and Cvetkovic, D. (2013). A single-trial toolbox for advanced sleep polysomnographic preprocessing. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5829–5832. IEEE.

Charlton, P. H., Marozas, V., Mejía-Mejía, E., Kyriacou, P. A., and Mant, J. (2025). Determinants of photoplethysmography signal quality at the wrist. *PLOS Digital Health*, 4(6):e0000585.

Chen, X., Xu, X., Liu, A., Lee, S., Chen, X., Zhang, X., McKeown, M. J., and Wang, Z. J. (2019). Removal of muscle artifacts from the eeg: A review and recommendations. *IEEE Sensors Journal*, 19(14):5353–5368.

Chriskos, P., Frantzidis, C. A., Gkivogkli, P. T., Bamidis, P. D., and Kourtidou-Papadeli, C. (2018). Achieving accurate automatic sleep staging on manually pre-processed eeg data through synchronization feature extraction and graph metrics. *Frontiers in human neuroscience*, 12:110.

Correa, M. A. G. and Leber, E. L. (2011). Noise removal from eeg signals in polisomnographic records applying adaptive filters in cascade. *Adaptive filtering applications*, 34:1–26.

de Zambotti, M., Goldstein, C., Cook, J., Menghini, L., Altini, M., Cheng, P., and Robillard, R. (2024). State of the science and recommendations for using wearable technology in sleep and circadian research. *Sleep*, 47(4):zsad325.

Fedorin, I. and Slyusarenko, K. (2021). Consumer smartwatches as a portable psg: Lstm based neural networks for a sleep-related physiological parameters estimation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 849–452.

Gunter, K. M., Brink-Kjær, A., Mignot, E., Sørensen, H. B., During, E., and Jennum, P. (2023). Svit: A spectral vision transformer for the detection of rem sleep behavior disorder. *IEEE Journal of Biomedical and Health Informatics*, 27(9):4285–4292.

Huang, X., Shirahama, K., Irshad, M. T., Nisar, M. A., Piet, A., and Grzegorzek, M. (2023). Sleep stage classification in children using self-attention and gaussian noise data augmentation. *Sensors*, 23(7).

Kim, J., Lee, J., and Shin, M. (2017). Sleep stage classification based on noise-reduced fractal property of heart rate variability. *Procedia Computer Science*, 116:435–440. Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017).

Metsis, V., Schizas, I. D., and Marshall, G. (2015). Real-time subspace denoising of polysomnographic data. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 1–4.

Permana, Z. Z. R., Sari, R. I., Febriani, N. S., and Setiawan, A. W. (2023). Effect of smote for sleep stages classification using decision tree, k-nearest neighbor and random forest. In *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 1–6.

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and De Vos, M. (2019). Joint classification and prediction cnn framework for automatic sleep stage classification. *IEEE Transactions on Biomedical Engineering*, 66(5):1285–1296.

Sekkal, R. N., Bereksi-Reguig, F., Ruiz-Fernandez, D., Dib, N., and Sekkal, S. (2022). Automatic sleep stage classification: From classical machine learning methods to deep learning. *Biomedical Signal Processing and Control*, 77:103751.

Walch, O., Huang, Y., Forger, D., and Goldstein, C. (2019). Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*, 42(12):zsz180.

Wang, Q., Zhao, D., Wang, Y., et al. (2019). Ensemble learning algorithm based on multi-parameters for sleep staging. *Medical & Biological Engineering & Computing*, 57(8):1693–1707.