





Article

SOUP: Sleep Data Copilot for Accurate Hypnogram Labeling

Marta Verona-Almeida ^{1,2} , Javier Mendez ¹, Rybel Wix ³ , José L. Ayala ^{4,5,*}  and Josué Pagán ^{2,5} 

- ¹ HTEC GmbH, Leipziger Str. 16, 82008 Unterhaching, Germany; marta.verona.almeida@htecgroup.com or marta.valmeida@alumnos.upm.es (M.V.-A.); javier.mendez.gomez@htecgroup.com (J.M.)
- ² Electronic Engineering Department, Universidad Politécnica de Madrid, Av. Complutense, 30, 28040 Madrid, Spain; j.pagan@upm.es
- ³ Department of Neurophysiology, Instituto de Investigacion Sanitaria Hospital Universitario de la Princesa, 28006 Madrid, Spain
- ⁴ Department of Computer Architecture and Automation, Universidad Complutense de Madrid, C/Profesor José García Santesmases, 9, 28040 Madrid, Spain
- ⁵ Center for Computational Simulation, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Boadilla del Monte, Spain
- * Correspondence: jayala@ucm.es

Abstract

Sleep analysis is crucial for diagnosing disorders and understanding physiological patterns. However, accurately labeling hypnograms is challenging due to significant interrater variability and resource constraints that limit the use of multiple experts. This study introduces a novel verification tool that assesses biomedical signals, including heart rate and activity, alongside labeled hypnograms against state-of-the-art conditions. The tool was developed to evaluate the quality and reliability of hypnogram annotations, providing feedback on the credibility of labels generated by automated methods and single expert annotations. It cross-references labeled data against physiological signals and identifies discrepancies or anomalies that may indicate errors in the labeling process. For validation, the tool was applied to the MESA dataset, a well-known collection of sleep data. Application of the tool demonstrated its ability to provide objective feedback on hypnogram labels and to identify anomalies in patient data, potentially assisting clinicians in refining their assessments. By offering a user-friendly interface and flexible design, this verification tool enhances the accuracy of sleep stage annotations and serves as a valuable resource for both clinical and research applications.



check for updates

Academic Editor: Gang Wei

Received: 13 November 2025

Revised: 30 November 2025

Accepted: 3 December 2025

Published: 8 December 2025

Citation: Verona-Almeida, M.; Mendez, J.; Wix, R.; Ayala, J.L.; Pagán, J. SOUP: Sleep Data Copilot for Accurate Hypnogram Labeling. *Appl. Sci.* **2025**, *15*, 12912. <https://doi.org/10.3390/app152412912>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: activity; heart rate; hypnogram; verification

1. Introduction

Sleep analysis plays a crucial role in understanding physiological patterns and diagnosing sleep disorders. Physiological signals, such as heart rate (HR), activity (ACC), body temperature, or electrodermal activity (EDA), provide valuable insights into sleep stages and overall sleep quality. These sleep stages are graphically represented in a hypnogram, which illustrates the different stages of sleep—wakefulness, light sleep (N1 and N2), deep sleep (N3), and REM sleep—over the course of a sleep period, showing transitions and durations for each stage.

Accurately labeling hypnograms is an important challenge that remains significant in sleep research and clinical practice. In fact, the reliability of the widely used American Academy of Sleep Medicine (AASM) guidelines for sleep staging has been questioned due to substantial variability in agreement between scorers [1–3]. Studies show that

Kappa values for inter-rater agreement on sleep stages vary widely, with particularly low agreement for the N1 stage ($k = 0.241, 0.412$). Furthermore, when scorers are from different medical centers, the Kappa value may be 0.58, but it can increase to 0.78 when they are from the same center [2]. To ensure certainty, the AASM recommends that ideally the sleep stage scoring should be performed by three different experts [4]. However, due to resource constraints, this rigorous labeling process may not always be feasible in practice. This, together with the non-validation of the data, can lead researchers to wrong conclusions without any clue of why this happens. This is particularly concerning in machine learning research, where data is gathered without verification, and the interpretability of results is often overlooked.

To address these challenges, several automated sleep staging tools have been developed. These tools aim to improve the consistency and efficiency of sleep stage labeling by leveraging machine learning algorithms and physiological signals. For example, Stephansen et al. [5] developed an automated sleep staging system that achieved human-level performance on the MASS dataset. Similarly, Fiorillo et al. [6,7] proposed a deep learning-based approach for sleep staging using EEG signals that outperformed traditional machine learning methods. Moreover, YASA [8] analyzed EEG, EOG, and EMG signals to perform sleep stage scoring, extracting sleep-related metrics, such as sleep efficiency or total sleep time. While these tools show promise, they still require careful verification to ensure their reliability and accuracy.

Some methods go beyond sleep-stage labeling by assessing confidence or uncertainty, typically using EEG, EOG, or EMG signals from PSG recordings. For example, aSAGA [9] is an automated sleep-staging model that identifies ambiguous regions in the hypnogram for expert review. U-PASS [10] incorporates uncertainty estimation to flag low-confidence epochs, allowing experts to review them and improve overall annotation reliability.

In contrast, SOUP provides a novel verification framework that quantitatively evaluates the alignment between hemodynamic signals and expected physiological patterns for each sleep stage. While other tools focus primarily on signal quality and stage consistency, SOUP adds a complementary perspective, providing an additional layer of data processing that directly assesses the reliability of hypnogram labels. It enables both single-patient and cohort-level assessments, allowing systematic detection of anomalous or inconsistent sleep stage annotations across datasets, as well as identifying potential physiological anomalies or health issues in individual patients. Clinicians and researchers can use this feedback to validate hypnograms, identify sleep stages requiring further review, and ensure higher-quality sleep stage data.

Designed for scalability, the tool allows non-programmers to easily incorporate new conditions or integrate additional biomedical signals for specific research or clinical needs. Its versatility enables application at both the individual patient level—supporting daily monitoring—and the cohort level—allowing large-scale sleep stage quality assessments and uncovering systematic inconsistencies across datasets.

The remainder of this paper is organized as follows: Section 2 begins with an overview of current research on patterns of relevant hemodynamical signals across sleep stages, followed by a presentation of the verification tool's framework and workflow. Section 3 details the results from applying the tool to the MESA dataset, and Section 4 discusses the implications of these findings and outlines potential future directions for this research. Finally, Section 5 provides the main conclusions of the study.

2. Materials and Methods

2.1. Biomedical Signal Characterization

The analysis of biomedical signals, such as heart rate (HR) and activity (ACC), can provide valuable insights into sleep stages and overall sleep quality. Several studies have investigated the relationship between these signals and sleep stages despite the absence of specialized verification tools. Using statistical tests and various analytical methods, these investigations have explored patterns within each stage of sleep.

- **Activity:** Activity is a widely researched signal to monitor sleep through wearable devices and non-invasive monitoring systems. According to Gaiduk et al. [11], activity levels in stage N1 are higher than in other sleep stages. Further research by Gaiduk et al. [12] indicates that the highest activity is observed during awakeness, while the NREM stages generally exhibit smaller and less frequent movements compared to awake states. Similarly, the study found that body movements are more frequent in epochs right before and after the REM stage. Finally, Ogilvie et al. [13] found that the awake and N1 stages exhibit the highest similarity in terms of activity levels.
- **Heart Rate:** Heart rate and its variability (HRV) between different stages of sleep have been the focus of numerous studies. When analyzing the signal in the time domain, Stein et al. [14] reported that heart rate increases during the REM stage and decreases during NREM, resulting in higher and less stable HR values during REM sleep compared to NREM [15]. Compared to wakefulness, Gaiduk et al. [12] found that transitioning from wakefulness to deeper sleep results in a decrease in heart rate, leading to a lower HRV during NREM compared to wakefulness [15]. However, heart rate variability in the frequency domain examines the power distribution across different frequency bands, particularly the low-frequency (LF, 0.04 to 0.15 Hz) and high-frequency (HF, 0.15 to 0.4 Hz) bands, as well as the LF:HF ratio. These three signals have been widely studied in terms of sleep stages. Firstly, the lowest LF power occurs during N3 sleep. Furthermore, Kryger et al. [15] established that LF power during NREM sleep is lower than during REM sleep and wakefulness. Within the same research [15], it was also revealed that the HF component reaches its highest levels during N3 sleep. Finally, in terms of LF:HF ratio, REM sleep has a higher ratio compared to NREM sleep [16,17]. In particular, the lowest ratio is found during the N3 stage of sleep [15,16,18].
- **Electrodermal activity (EDA):** EDA is a physiological indicator that reflects the activity of sweat glands, which are controlled by the autonomic nervous system. EDA can be decomposed into two components: the tonic component, which represents the skin conductance level (SCL), and the phasic component, which represents skin conductance responses (SCR). These signals are widely used in sleep research to assess variations in autonomic nervous system activity at different sleep stages. According to Herlan et al. [19], the EDA while awake is significantly higher than each of the other sleep stages. Herlan et al. also stated that SCR tends to show higher levels when individuals are awake. However, several studies [20,21] highlighted that higher SCR are found in N3 compared to the rest of the sleep stages. Finally, Richer et al. [22] found that SCL decreases during sleep.
- **Body temperature:** Skin temperature is a relevant physiological parameter that reflects the processes of body temperature regulation during sleep. Wei and Boger [23] found that wrist temperature increases before individuals fall asleep. In addition, their study reported a drop in wrist temperature in the morning when individuals wake up. Finally, according to Cambell and Murphy [24], women exhibit significantly higher temperatures throughout the night compared to men.

While these studies provide valuable insights into the relationships between biomedical signals and sleep stages, they often focus on specific signals or datasets, limiting their generalizability. Furthermore, most of these studies rely on manual sleep stage annotations, which can be subject to the same inconsistencies and reliability issues as the AASM guidelines.

2.2. Framework

This section introduces the developed framework, illustrated in Figure 1. The figure highlights three primary blocks that constitute the tool, along with their potential dependencies, described in this section. This framework is designed to validate the alignment of sleep stage labels with expected biomedical signal patterns, based on state-of-the-art conditions.

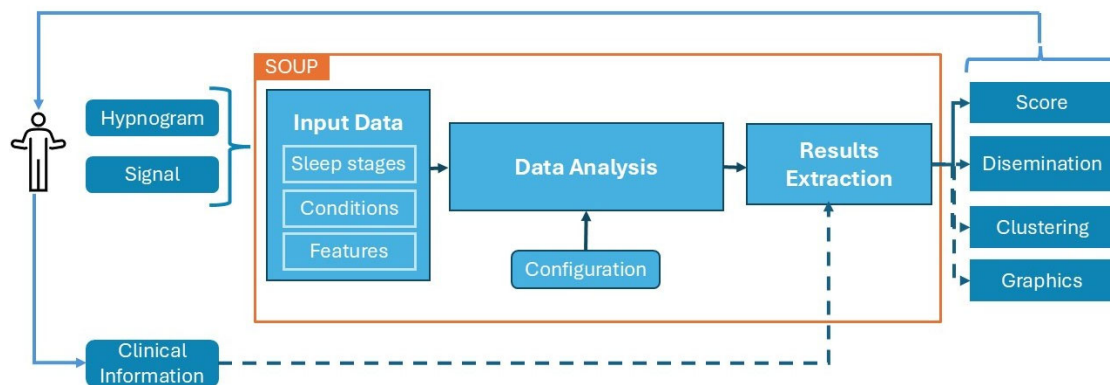


Figure 1. Overview of the proposed validation framework. Dashed lines indicate the functionalities that can be used just when a full dataset is analyzed. Full lines indicate both single-patient data and the full dataset. Arrows illustrate the data flow throughout the tool.

2.2.1. Input Data

As mentioned previously and shown in Figure 1, the tool requires a biomedical signal-hypnogram pair (e.g., activity-hypnogram) as input. For effective usage of the tool, the user should consider the following factors:

1. Synchronization of Features: All features must be synchronized, in 30 s epochs, for proper usage, ensuring consistency across the dataset.
2. Stage Labeling Consistency: Stage labeling must align with the specifications provided in the configuration file to avoid misinterpretation of results.
3. Clinical Dissemination: The user can perform a dissemination of the results based on clinical information. This should be a table containing a column named 'id', which matches the identifiers in the input data. Additionally, the remaining columns of this data frame will be automatically analyzed based on their values.

Following these guidelines ensures accurate analysis of results. Additional documentation on optimal tool usage, along with instructions for adding or modifying the state-of-the-art conditions to be evaluated, can be found in the SOUP opensource project on GitHub (<https://github.com/eesy-ai/soup>, accessed on 12 November 2025).

2.2.2. Data Analysis

In this section, the three main blocks of the data analysis functionality of this tool are explained. First, the biomedical conditions implemented, as specified in the configuration file, are detailed. Second, the analysis of those conditions is described, assessing whether the given labels align with expected physiological patterns.

1. Biomedical signal conditions: All the state-of-the-art conditions discussed in Section 2.1 regarding activity and heart rate have been implemented. This decision

was influenced by the prevalence of these parameters in public datasets containing biomedical signals and synchronized hypnograms [25,26]. In contrast, EDA and temperature are less commonly available, as noted in the mentioned databases. Therefore, SOUP currently does not include conditions for these signals. As a result, the tool includes implementations for 6 conditions to assess activity during sleep stages, 7 conditions for heart rate in the frequency domain (analyzing LF, LF:HF), and 6 conditions for heart rate in the time domain (HR, HRV). Table 1 shows a summary of the conditions implemented in this tool. Notably, the framework is flexible, allowing users to easily modify existing conditions or incorporate additional signals—such as EDA or temperature—for specific research or clinical applications.

- Value Analysis: An analysis based on absolute values is performed to determine if the conditions are met. This approach is useful because, in this type of dataset, the number of epochs in some sleep stages is reduced, limiting statistical power in the tests. All epochs along the night of the relevant sleep stages are analyzed for each specific condition. For conditions related to tendencies (i.e., feature values increasing or decreasing), a condition is considered fulfilled when at least 50% of epoch slots match the specified trend. This threshold was chosen as a balanced default to capture a majority trend while allowing for normal variability within a night. Importantly, this minimum percentage is configurable in the configuration file, enabling users to adjust the criteria based on cohort characteristics, signal quality, or research objectives.

Table 1. Summary of implemented conditions.

| Signal | Condition | Name |
|--------------|-------------------------------|--|
| ACC | [ACC] REM < Awake | NREM presents lower activity than awakesness |
| | [ACC] Awake > Sleep | Highest activity observed during awakesness |
| | [ACC] Awake \approx N1 | Awake and N1 stages exhibit similar activity levels |
| | [ACC] N1 > All Stages | Activity level in N1 is higher than in other sleep stages |
| | [ACC] NREM < (frequent) Awake | As sleep deepens, body activity becomes lower and less frequent |
| | [ACC] Higher Before/After REM | Movement tends to occur immediately before and after REM stage |
| HR (time) | [HR] NREM ↓ | HR decreases during NREM |
| | [HR] REM ↑ | HR increases during REM |
| | [HR] Awake > NREM | When transitioning from wakefulness to deeper sleep, HR decreases |
| | [HR] REM > NREM | HR increases with more variability in REM sleep |
| | [HRV] NREM < REM | HR is more stable during NREM than during REM |
| | [HRV] Awake > NREM | HRV during NREM is lower than during wakefulness |
| HR (freq) | [HF] N3 > NREM | Highest HF power during NREM is found in N3 sleep |
| | [LF] NREM < Awake | LF power is lower during NREM sleep compared to wakefulness |
| | [LF] NREM < REM | LF power is lower during NREM compared to REM |
| | [LF] N3 < All Stages | The lowest LF power is found in N3 |
| | [LF:HF] N3 < All Stages | The lowest LF:HF ratio is found in the N3 stage |
| | [LF:HF] REM < Awake | LF:HF ratio is higher during wakefulness compared to the REM stage |
| | [LF:HF] REM > NREM | LF:HF ratio during REM is higher than during NREM stages |

2.2.3. Result Extraction

This section describes the tool's outputs, including textual summaries, tables, and graphics. Additionally, a PDF containing all generated information is created to improve accessibility. These outputs offer comprehensive insights into the analyzed data, providing details for exploration and dissemination, and structured data for quantitative analysis. Healthcare professionals can use these outputs as recursive feedback during sleep studies to enhance the quality of the data.

1. **Score:** The final output of this tool represents the proportion of conditions that match expected patterns across patients. Therefore, a data table is generated, indicating if each condition is fulfilled on average during the night. This aims to allow users to perform in-depth post-analysis if desired.
2. **Dissemination:** When processing a complete dataset, the tool generates a comprehensive summary of conditions met across different sleep stages, facilitating the identification of inconsistencies. Additionally, it supports the integration of clinical metadata when available (see Section 3.2), enabling a deeper analysis. Beyond reporting the overall percentage of patients meeting each condition, the tool provides demographic breakdowns, such as condition fulfillment rates by sex, age group, or other relevant clinical factors. These percentages, collectively referred to as the score, are displayed for each patient. In full-dataset analyses, the score reflects the average fulfillment rate across the entire cohort, offering an intuitive measure of data consistency and quality.
3. **Graphics:** When a full dataset is included, several graphics will be displayed for a visual analysis of both the conditions and the patients. All these graphics are presented and analyzed in detail in the next section IV, where a usage example is described.
 - a. **Heatmap:** The heatmap displays patients as rows and conditions as columns. A green cell indicates that the patient fulfills the condition, while a red cell indicates that the condition is not fulfilled (Section 3.2.1).
 - b. **Bar plot:** This bar plot illustrates the percentage of patients fulfilling each condition, with each bar representing a different condition (Section 3.2.2).
4. **Clustering:** Patients are clustered using hierarchical clustering with a minimum distance of one, ensuring that only those who fulfill the exact same conditions are grouped together. This facilitates the identification of isolated patients and potential health-related conditions for further investigation. This approach can aid in clinical decision making and enables the analysis of cohorts of interest within a population (Section 3.2.1).

3. Results

SOUP can be used for the analysis of clinical datasets and the re-evaluation of existing data for a complete analysis, including the possibility of making new diagnoses. Additionally, it is a valuable tool for daily patient management and data analysis. In this section, both use cases will be demonstrated using the MESA dataset (ID 5003) [26], with access granted by the National Sleep Research Resource [27]. The objective is not to provide a new diagnosis, but to showcase the functionalities of the tool, aiming to detect any deviations or error trends in the labeling and analyze patient cohorts.

Finding a dataset that includes hypnograms, ECG, and activity data can be challenging, which makes the comprehensive nature of the MESA dataset particularly noteworthy. MESA dataset is a longitudinal study on cardiovascular disease and its risk factors, comprising data from 6.814 participants. Among them, 2.237 underwent a Sleep Exam with polysomnography, actigraphy (ACC), ECG monitoring, and a sleep questionnaire. After excluding participants with unsynchronized activity, missing values, or non-monotonic R points, the final cohort consisted of 1.308 patients (55.20% women, 44.80% men), aged 54–94 years (69 ± 9). Moreover, for clinical dissemination, only sleep-related diseases (sleep apnea, insomnia, and restless legs) have been considered. Those without sleep problems were considered healthy patients even when they may suffer from other health issues. Minimal processing was applied, primarily utilizing the pyHRV package (version 0.4.1) [28] to extract ECG temporal and frequency components.

This dataset serves as an ideal validation platform for testing the tool's capacity in analyzing biomedical signals (HR and ACC) and hypnograms across different populations. This dataset effectively showcases the practical applications and effectiveness of the developed tool.

3.1. Single Patient Analysis

To evaluate the tool's functionality for single-patient analysis, a 65-year-old white woman (ID 1102) was randomly selected. To simulate the feedback for healthcare professionals, the patient's hypnogram was deliberately modified based on interrater disagreement ratios reported by Deng et al. [29]: N2–N3 (22.09%), Wake–N1 (19.68%), and N1–N2 (18.75%). Accordingly, a proportion of epochs in the original hypnogram were randomly reclassified, mimicking typical interrater variability in sleep stage annotations: 22.09% of N2 to N3, 19.68% of Wake to N1, and 18.75% of N1 to N2. These modifications allow for assessing how discrepancies in sleep stage labeling influence the fulfillment of expected physiological conditions.

Table 2 presents the fulfillment percentages for the three signals in the modified hypnogram: Activity (ACC), Heart Rate in the time domain (HR Time), and Heart Rate in the frequency domain (HR Freq). Conditions associated with the N1 stage for ACC are entirely unmet (0%), whereas those for NREM and Awake are largely satisfied. In contrast, HR Time and HR Freq conditions generally demonstrate moderate to high fulfillment rates across sleep stages.

Table 2. Percentage of conditions fulfilled by biomedical signals and sleep stages in the modified hypnogram (ID 1102).

| Condition | Stage | Fulfillment Rate (%) |
|-----------|-------|----------------------|
| ACC | N1 | 0.00 |
| | NREM | 100.00 |
| | REM | 100.00 |
| | Awake | 75.00 |
| HR (time) | NREM | 80.00 |
| | REM | 66.67 |
| | Awake | 100.00 |
| HR (freq) | N3 | 33.33 |
| | NREM | 66.67 |
| | REM | 66.67 |
| | Awake | 100.00 |

At this stage, a healthcare professional might question why no N1-specific condition is met, prompting further investigation into potential misclassifications. This is particularly relevant, as N1 is one of the most challenging sleep stages to accurately classify.

A detailed analysis of the ACC conditions in Table 3 provides insights into when each condition is met. The results suggest that awake movement patterns differ from those in N1, and the highest movement levels during sleep do not occur in the N1 stage. This discrepancy may stem from the 18.75% of N1 epochs misclassified as N2, potentially disrupting expected movement patterns.

In practice, these outputs serve as a warning for healthcare professionals to review and, if necessary, correct the N1 stage annotations. In this scenario, restoring the original labels does not impact the HR-related conditions but significantly improves the ACC conditions fulfillment. Table 4 presents the status of the compliance of ACC conditions after the hypnogram correction. In particular, the table illustrates that all conditions are now consistently met on average.

Table 3. ACC conditions with modified hypnogram (ID 1102). ✓ Indicates that the condition was fulfilled and X otherwise.

| Condition | Fulfilled |
|-------------------------------|-----------|
| [ACC] NREM < Awake | ✓ |
| [ACC] Awake > Sleep | ✓ |
| [ACC] Awake ≈ N1 | X |
| [ACC] N1 > All Stages | X |
| [ACC] NREM < (frequent) Awake | ✓ |
| [ACC] Higher Before/After REM | ✓ |

Table 4. ACC conditions with corrected hypnogram (ID 1102). ✓ Indicates that the condition was fulfilled.

| Condition | Fulfilled |
|-------------------------------|-----------|
| [ACC] NREM < Awake | ✓ |
| [ACC] Awake > Sleep | ✓ |
| [ACC] Awake ≈ N1 | ✓ |
| [ACC] N1 > All Stages | ✓ |
| [ACC] NREM < (frequent) Awake | ✓ |
| [ACC] Higher Before/After REM | ✓ |

This analysis shows how the tool can identify discrepancies in sleep stage classification, providing healthcare professionals with insights to refine annotations and improve its accuracy.

3.2. Full Dataset Analysis

In this section, an analysis of the entire dataset is performed to demonstrate how to interpret the textual outputs and graphs generated by the tool. For each signal within the dataset, various produced outputs are presented and explained to provide a comprehensive understanding of all the output information.

3.2.1. Activity (ACC)

To analyze the ACC signal, Table 5 shows a partial summary of scores by patient, generated as a textual output by the tool. The data reveal an agreement score of 84.75% for the conditions. This demonstrates that the studied database generally aligns with state-of-the-art conditions.

Table 5. Summary of the scores for the activity signal.

| Patient ID | Fulfillment Rate (%) |
|------------|----------------------|
| ... | ... |
| 6751 | 100.00 |
| 6792 | 33.33 |
| 6796 | 100.00 |
| SCORE | 84.75 |

To analyze the specific conditions fulfilled, Figure 2 presents the heatmap displaying the fulfillment of conditions. Remarkably, the graph shows that four of the conditions have a high acceptance rate, while two conditions show some disagreement: [ACC] Awake ≈ N1 and [ACC] N1 > All Stages.

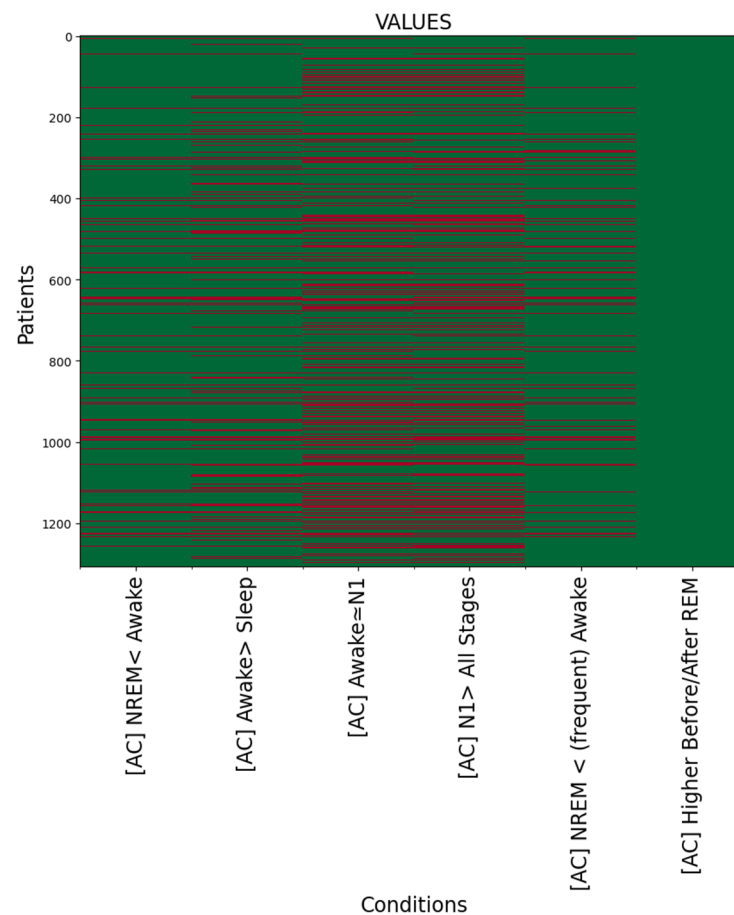


Figure 2. ACC: Heatmap indicating the fulfillment of activity conditions.

These findings, supported by the textual output, indicate that the two conflicting conditions are fulfilled in at least 70% of the patients, while the remaining conditions are satisfied by at least 80%. This demonstrates a strong overall agreement of the MESA dataset with the ACC conditions.

In addition, the clustering of patients who meet the same conditions is performed. In this analysis, the framework identifies 18 distinct clusters, with Table 6 presenting a partial output of the clustering dissemination based on clinical information. The largest cluster consists of 838 patients who fully meet all conditions, reflecting the dataset's overall trend. Within this group, 75.30% are healthy, 57.88% are female, and the average age is 70 years. In contrast, the smallest cluster includes only three individuals, where all conditions are met except for a discrepancy in [ACC] NREM < (frequent) Awake. However, a review of their clinical details revealed no clear pattern that explains this common misalignment. Additionally, the framework identifies a distinct subgroup, predominantly composed of women (8 women and 1 man), where only two conditions are met: [ACC] NREM < (frequent) Awake and [ACC] Higher Before/After REM. These patients are mostly healthy, with an average age of 72 ± 11 years, suggesting a subgroup of women whose ACC conditions deviate from the expected patterns. Further clinical analysis may help uncover underlying factors contributing to this variation, offering deeper insights into potential physiological or clinical differences within this group.

This analysis of the ACC signal shows a strong alignment of the database with state-of-the-art conditions in general terms and across all the studied sleep stages.

Table 6. Partial output of the clustering dissemination. ✓ Indicates that the condition was fulfilled and X otherwise.

| | | | | |
|--|--------|--------|-----|---------|
| Cluster ID | 1 | 2 | ... | 18 |
| Size | 3 | 838 | ... | 9 |
| Male (%) | 66.67 | 57.88 | ... | 88.89 |
| Female (%) | 33.33 | 42.12 | ... | 11.11 |
| Age (years) | 70 ± 7 | 69 ± 8 | ... | 72 ± 11 |
| Healthy (%) | 66.67 | 75.29 | ... | 88.89 |
| Insomnia (%) | 0 | 5.37 | ... | 0 |
| Sleep Apnea (%) | 33.33 | 15.87 | ... | 11.11 |
| Restless Legs (%) | 0 | 3.4 | ... | 0 |
| [AC] NREM < Awake | ✓ | ✓ | ... | X |
| [AC] Awake > Sleep | ✓ | ✓ | ... | X |
| [AC] Awake ≈ N1 | ✓ | ✓ | ... | X |
| [AC] N1 > All Stages | ✓ | ✓ | ... | X |
| [AC] NREM < (frequent) Awake | X | ✓ | ... | X |
| [AC] Post/Prev REM Higher | ✓ | ✓ | ... | ✓ |

3.2.2. HR—Time Domain

Considering HR with the time domain, the overall score is 73.90%. Consequently, the dataset meets the state-of-the-art conditions regarding HR in the time domain with acceptable precision.

To describe another output graphic of the tool, the analysis of specific condition fulfillment begins with an assessment of the barplots (see Figure 3), which show the fulfillment of each analyzed condition. This image shows that there are just two conditions that are fulfilled by less than 70% of the patients: [HR] NREM ↓ and [HRV] NREM < REM, fulfilled by 49.31 and 65.98%, respectively.

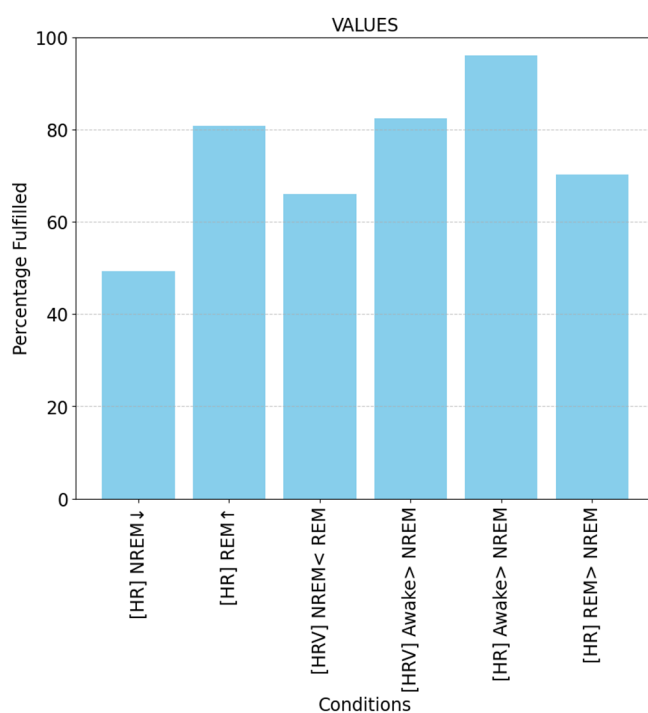


Figure 3. HR-time domain: Bar plot indicating the percentage of conditions met.

The clinical dissemination of [HRV] NREM < REM is presented in Table 7, showing that certain groups lower the overall condition fulfillment rate. Specifically, patients with apnea have a fulfillment rate of only 63.78%. Additionally, African American and White

patients register 63.55% and 63.29%, respectively. Furthermore, individuals older than 85 years exhibit the lowest fulfillment rate at 61.97%. These findings highlight specific demographics and conditions associated with lower fulfillment rates in the clinical data. Notably, healthy individuals show a scoring of 65.84%, lower than those with insomnia (71.64%) or restless legs syndrome (71.05%), suggesting that underlying cardiac issues may affect HRV despite the absence of sleep-related conditions.

Table 7. Clinical dissemination of [HRV] NREM < REM.

| Category | Group | Fulfillment Rate (%) |
|----------|-------------------------|----------------------|
| Overall | All | 65.98 |
| Disease | Healthy | 65.84 |
| | Insomnia | 71.64 |
| | Restless Legs Syndrome | 71.05 |
| | Sleep Apnea | 72.41 |
| Race | Black, African American | 63.55 |
| | Chinese American | 71.11 |
| | Hispanic | 70.75 |
| | White | 63.29 |
| Age | (54–64) | 67.72 |
| | (65–74) | 65.37 |
| | (75–84) | 64.91 |
| | (85–94) | 61.97 |
| Sex | Female | 64.88 |
| | Male | 67.52 |

A sensitivity analysis was performed by varying the configurable threshold for tendency conditions, which defaults to 50% and is explained in Section 2.2.2. Thresholds were tested from 10% to 90%. As shown in Figure 4, [HR] NREM ↓ drops sharply after 40%, indicating high sensitivity to stricter criteria. In contrast, [HR] REM ↑ declines more gradually, with a notable decrease only after 60%, suggesting greater robustness across the cohort. The overall score does not decrease dramatically because this parameter affects only two of the six conditions that contribute to the total score. At a 50% threshold, the overall score was 74.12%. [HR] NREM ↓ was fulfilled in 49.31% of epochs, and [HR] REM ↑ in 80.73%. These results highlight the importance of threshold selection and demonstrate the flexibility of SOUP, which allows users to adjust criteria according to cohort characteristics and study requirements.

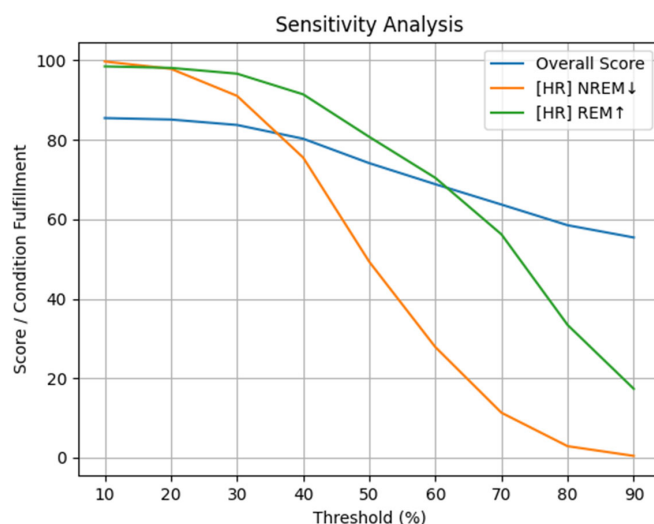


Figure 4. HR-time domain: Sensitivity analysis of condition fulfillment thresholds.

Temporal HR analysis confirms alignment with state-of-the-art conditions. However, specific demographic groups, such as patients with apnea and older adults, show lower scores, reflecting expected dysfunction of the autonomic nervous system.

3.2.3. HR—Frequency Domain

Analyzing HR within the frequency domain, the obtained score is 37.81%, which is considerably lower than the scores obtained for the previously examined signals.

To evaluate the fulfillment of each specific condition, Figure 5 shows the bar plot of the percentage of fulfillment for each condition. There it is shown that [LF:HF] REM > NREM and [LF:HF] REM < Awake conditions are met by 70.11% and 65.90% of the patients, respectively, while the rest of them are met by 30% or less. This highlights a weak alignment with the state-of-the-art outcomes.

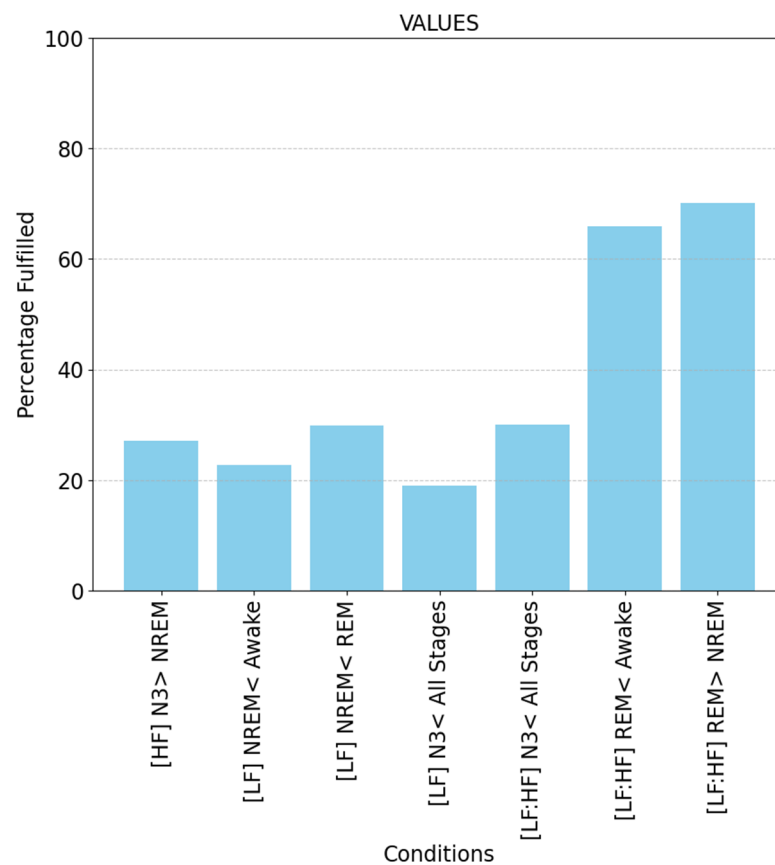


Figure 5. HR-frequency domain: Bar plot indicating the percentage of conditions met.

Table 8 shows the fulfillment rate of all conditions categorized by sleep stage. It is evident that the N3-related conditions are particularly problematic in this dataset, highlighting the difficulty of achieving consistent condition satisfaction during deep sleep.

Table 9 shows the clinical dissemination of the least fulfilled condition: [LF] N3 < All Stages. It shows that elderly individuals are less likely to meet this condition compared with younger individuals, revealing that age is a factor to consider when analyzing HR conditions in the frequency domain.

Analyzing the distribution of the sleep stages throughout the nights (Figure 6), it is evident that only 9.47% of the epochs are labeled as N3 sleep. According to Ashish K. Patel et al. [30], under optimal circumstances, 25% of sleep should be N3. Additionally, Dorffner et al. [31] noted a decrease of 1.7% per decade for males, which still exceeds the observed percentage in this dataset. This discrepancy could contribute to the mismatch among

the N3-related conditions, stemming from issues within the dataset, potential labeling inaccuracies, or the rigorous nature of the condition itself.

Table 8. Summary of HR conditions met (freq.) by sleep stage.

| Category | Group | Fulfillment Rate (%) |
|----------|-------------------------|----------------------|
| REM | [LF] NREM < Awake | 22.71 |
| | [LF] NREM < REM | 29.89 |
| | [LF:HF] REM < Awake | 65.90 |
| | [LF:HF] REM > NREM | 70.11 |
| NREM | [LF] NREM < Awake | 22.71 |
| | [LF] NREM < REM | 29.89 |
| | [LF:HF] REM > NREM | 70.11 |
| Awake | [LF] NREM < Awake | 22.71 |
| | [LF:HF] REM < Awake | 65.90 |
| N3 | [HF] N3 > NREM | 27.06 |
| | [LF] N3 < All Stages | 18.96 |
| | [LF:HF] N3 < All Stages | 30.05 |

Table 9. Clinical dissemination of [LF] N3 < All Stages.

| Category | Group | Fulfillment Rate (%) |
|----------|-------------------------|----------------------|
| Overall | All | 18.96 |
| Disease | Healthy | 17.78 |
| | Insomnia | 19.40 |
| | Restless Legs Syndrome | 21.05 |
| | Sleep Apnea | 24.49 |
| Race | Black, African American | 25.30 |
| | Chinese American | 14.07 |
| | Hispanic | 10.05 |
| | White | 21.61 |
| Age | (54–64) | 20.99 |
| | (65–74) | 18.54 |
| | (75–84) | 17.70 |
| | (85–94) | 12.68 |
| Sex | Female | 18.61 |
| | Male | 19.45 |

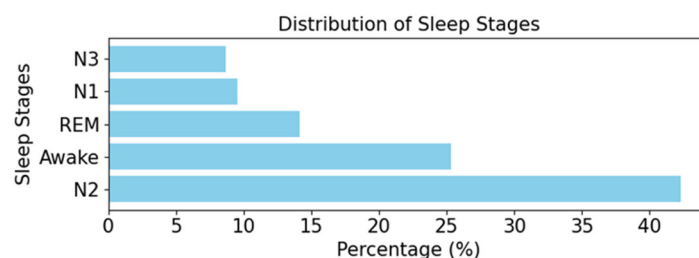


Figure 6. Percentage of epochs in each sleep stage in the dataset.

Another possible reason for these low scores could be the poor ECG signal quality observed in some intervals. Typically, resting HR ranges between 50 and 90 bpm [32]. However, this dataset includes intervals where HR spikes to 330 bpm.

These findings highlight the complexity of interpreting N3-related patterns in real-world, community-based datasets. While the tool can detect discrepancies between ex-

pected and observed patterns, such differences should be considered in the context of cohort demographics and signal quality, rather than automatically assumed to indicate mislabeling. Decisions regarding adjustment of expected N3 conditions for older populations would require specialized clinical evaluation, which is beyond the scope of this study.

4. Discussion

In the domain of sleep signal analysis, there is a wide range of specialized tools to facilitate the investigation of different biomedical signals. To begin with, pyActigraphy [33] is designed to process actigraphy data. This type of data, commonly obtained from wearable devices, provides valuable insight into activity patterns and circadian rhythms. Another notable tool is pyHRV [28], known for its effectiveness in analyzing heart rate variability. This aspect of physiological investigation provides an understanding of sleep quality and the dynamics of the autonomic nervous system, both of which are crucial in understanding the nature of restful sleep. Moreover, in the context of automatic classification of sleep stages, the tool Counting Sheep PSG [34] can be found. Designed to be EEGLAB-compatible, this software for MATLAB facilitates signal processing, visualization, event marking, and manual sleep stage scoring of polysomnography (PSG) data.

While these tools provide essential functionality for analyzing specific biomedical signals related to sleep, they fall short in offering a comprehensive framework for verifying the reliability of sleep stage annotations. Traditional tools typically focus on reclassifying sleep stages or processing signals, whereas our proposed tool, SOUP, addresses the critical need for validating the consistency of hypnogram–signal pairs against established state-of-the-art conditions. Unlike conventional tools, SOUP does not attempt to reclassify sleep stages but verifies the accuracy of provided annotations by comparing them with expected physiological patterns, such as activity and heart rate. This approach ensures the integrity and reliability of sleep stage annotations, which is crucial both for clinical practice and research. In this regard, SOUP represents a significant advancement over the state-of-the-art, filling an essential gap by systematically addressing the verification of sleep stage annotations and enhancing the reliability of sleep studies.

Beyond classical signal-processing tools, recent frameworks such as aSAGA or U-PASS introduce mechanisms to quantify uncertainty or ambiguity. As discussed in the Introduction, these methods help identify ambiguous or potentially unreliable epochs. However, their scope remains fundamentally different. In contrast to these approaches, which focus on model confidence or signal integrity, SOUP directly evaluates whether the physiological dynamics of the signals are consistent with the annotated sleep stages. This makes SOUP complementary rather than redundant: it provides a physiologically grounded validation step that current uncertainty-based or quality-metric tools do not offer.

SOUP acts as a “copilot” for sleep analysis, offering a systematic validation process for hypnogram labels. It is specifically designed to assess the alignment of sleep stage annotations with expected biomedical signal patterns. This approach makes SOUP a valuable asset for clinicians and researchers who rely on accurate sleep stage classification. Moreover, the tool is highly adaptable, capable of handling both individual and large-scale dataset analyses. It generates detailed reports that include textual summaries, tables, and graphics, making it accessible and valuable for both clinical practice and research.

From a clinical perspective, SOUP offers objective insights into sleep architecture, reducing interrater variability among sleep specialists and minimizing the time required for hypnogram generation. By enabling in-depth analyses of polysomnographic data, it enhances patient characterization, supporting more precise diagnostics and personalized interventions. Additionally, SOUP assists in identifying anomalies in sleep data, helping clinicians assess the clinical significance of deviations from established sleep patterns.

The tool is designed to be flexible, making it suitable for both research and clinical environments, independent of the source of hypnogram annotations, whether manual or automatic, or the number of annotators available. This flexibility is particularly valuable in clinical settings where only one specialist may be available for scoring, rather than the recommended three [4]. SOUP's scalability ensures that non-programmers can integrate new conditions and signals, thus expanding its applicability to various use cases and datasets.

To demonstrate its capabilities, the tool was applied to the MESA dataset, with two use cases: (i) a single-patient analysis, where errors in sleep stage classification were identified and corrected, and (ii) a full dataset evaluation, where systematic inconsistencies in sleep stage annotations were detected. The results showed that while activity and heart rate in the time domain generally aligned with expected conditions, compliance in the frequency domain was lower, particularly in N3 sleep. This discrepancy may indicate potential mislabeling or poor signal quality. It may also reflect cohort-specific characteristics, such as the advanced age of the MESA participants, whose physiology and sleep architecture can deviate from the normative patterns on which the conditions were originally defined. These considerations further reinforce the importance of verification tools like SOUP to enhance data integrity, while also highlighting the need to account for population-specific factors when interpreting condition fulfillment.

One of the key strengths of SOUP lies in its ability to analyze multiple signals and conditions simultaneously, ensuring the accuracy and reliability of sleep stage annotations. Furthermore, clinical professionals involved in this research, specifically neurophysiologists from the Sleep Unit at Hospital Universitario de la Princesa, have recognized SOUP as a valuable tool that can enhance clinical practice. By providing objective and systematic validation of sleep stage annotations, SOUP facilitates more consistent diagnostics and enables clinicians to efficiently identify anomalies or inconsistencies in polysomnographic data, ultimately contributing to improved patient care and optimized clinical workflows. Moreover, the structured reports generated by SOUP facilitate a deeper understanding of the data and provide a robust platform for targeted investigations into inconsistencies and anomalies. By providing both clinical and research benefits, SOUP offers a novel, efficient, and effective approach to improving the accuracy and reliability of sleep stage annotations, which is essential for advancing sleep medicine and related fields.

5. Conclusions

SOUP acts as a reliable copilot in sleep analysis by systematically verifying sleep stage annotations and highlighting potential inconsistencies. By improving the reliability and interpretability of hypnograms, it enhances the quality of sleep data and supports robust decision making in both clinical practice and research environments.

Author Contributions: Conceptualization, M.V.-A., J.M., R.W., J.L.A. and J.P.; Formal analysis, M.V.-A. and R.W.; methodology, M.V.-A.; software, M.V.-A.; validation, M.V.-A. and R.W.; investigation, M.V.-A.; visualization, M.V.-A.; writing—original draft preparation, M.V.-A.; writing—review and editing, J.M., J.L.A. and J.P.; supervision, J.M., J.L.A. and J.P.; project administration, J.L.A. and J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study were obtained from the Multi-Ethnic Study of Atherosclerosis (MESA) Sleep dataset, which is publicly available under registration through the National Sleep Research Resource.

Acknowledgments: The MESA Sleep study was funded by NIH-NHLBI (RO1 HL098433) and supported by NHLBI contracts HHSN268201500003I and N01-HC-95159 to N01-HC-95169 and NCATS agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. The National Sleep Research Resource was supported by NHLBI (R24 HL114473, 75N92019R002).

Conflicts of Interest: Authors Marta Verona-Almeida and Javier Mendez were employed by the company HTEC GmbH. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

| | |
|-----|------------------------|
| ACC | Activity |
| bpm | Beats per minute |
| ECG | Electrocardiogram |
| HR | Heart Rate |
| EDA | Electrodermal Activity |
| PSG | Polysomnography |

References

- Lee, Y.J.; Lee, J.Y.; Cho, J.H.; Choi, J.H. Interrater reliability of sleep stage scoring: A meta-analysis. *J. Clin. Sleep Med.* **2022**, *18*, 193–202. [[CrossRef](#)]
- Nikkonen, S.; Somaskandhan, P.; Korkalainen, H.; Kainulainen, S.; Terrill, P.I.; Gretarsdottir, H.; Sigurdardottir, S.; Olafsdottir, K.A.; Islind, A.S.; Óskarsdóttir, M.; et al. Multicentre sleep-stage scoring agreement in the Sleep Revolution Project. *J. Sleep Res.* **2023**, *33*, e13956. [[CrossRef](#)] [[PubMed](#)]
- Reinke, L.; van der Heide, E.M.; Fonseca, P.; Absalom, A.R.; Tulleken, J.E. Inter-rater disagreement in manual scoring of intensive care unit sleep data. *BMC Res. Notes.* **2025**, *18*, 138. [[CrossRef](#)]
- Hartley, S.; Goncalves, M.; Penzel, T.; Verbraecken, J.; Young, P. Revised European guidelines for the accreditation of sleep medicine centres. *J. Sleep Res.* **2024**, *33*, e14200. [[CrossRef](#)]
- Stephansen, J.B.; Olesen, A.N.; Olsen, M.; Ambati, A.; Leary, E.B.; Moore, H.E.; Carrillo, O.; Lin, L.; Han, F.; Yan, H.; et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat. Commun.* **2018**, *9*, 5229. [[CrossRef](#)] [[PubMed](#)]
- Fiorillo, L.; Favaro, P.; Faraci, F.D. DeepSleepNet-Lite: A simplified automatic sleep stage scoring model with uncertainty estimates. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 2076–2085. [[CrossRef](#)]
- Fiorillo, L.; Monachino, G.; van der Meer, J.; Pesce, M.; Warncke, J.D.; Schmidt, M.H.; Bassetti, C.L.; Tzovara, A.; Favaro, P.; Faraci, F.D. U-Sleep’s resilience to AASM guidelines. *npj Digit. Med.* **2023**, *6*, 33. [[CrossRef](#)] [[PubMed](#)]
- Vallat, R.; Walker, M.P. An open-source, high-performance tool for automated sleep staging. *eLife* **2021**, *10*, e70092. [[CrossRef](#)]
- Rusanen, M.; Jouan, G.; Huttunen, R.; Nikkonen, S.; Sigurðardóttir, S.; Töyräs, J.; Duce, B.; Myllymaa, S.; Arnardóttir, E.S.; Leppänen, T.; et al. Retrospective validation of automatic sleep analysis with grey areas model for human-in-the-loop scoring approach. *J. Sleep Res.* **2025**, *34*, e14362. [[CrossRef](#)]
- Heremans, E.R.; Seedat, N.; Buysse, B.; Testelmans, D.; van der Schaar, M.; De Vos, M. U-PASS: An uncertainty-guided deep learning pipeline for automated sleep staging. *Comput. Biol. Med.* **2024**, *171*, 108205. [[CrossRef](#)]
- Gaiduk, M.; Penzel, T.; Ortega, J.A.; Seepold, R. Automatic sleep stages classification using respiratory, heart rate and movement signals. *Physiol. Meas.* **2018**, *39*, 124008. [[CrossRef](#)] [[PubMed](#)]
- Gaiduk, M.; Perea, J.J.; Seepold, R.; Madrid, N.M.; Penzel, T.; Glos, M.; Ortega, J.A. Estimation of sleep stages analyzing respiratory and movement signals. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 505–514. [[CrossRef](#)] [[PubMed](#)]
- Ogilvie, R.D.; Wilkinson, R.T.; Allison, S. The detection of sleep onset: Behavioral, physiological, and subjective convergence. *Sleep* **1989**, *12*, 458–474. [[CrossRef](#)]
- Stein, P.K.; Pu, Y. Heart rate variability, sleep and sleep disorders. *Sleep Med. Rev.* **2012**, *16*, 47–66. [[CrossRef](#)]
- Kryger, M.H.; Roth, T.; Goldstein, C.A. *Principles and Practice of Sleep Medicine*, 7th ed.; Elsevier: Amsterdam, The Netherlands, 2022.
- Boudreau, P.; Yeh, W.H.; Dumont, G.A.; Boivin, D.B. Circadian variation of heart rate variability across sleep stages. *Sleep* **2013**, *36*, 1919–1928. [[CrossRef](#)]
- Scholz, U.J.; Bianchi, A.M.; Cerutti, S.; Kubicki, S. Vegetative background of sleep: Spectral analysis of the heart rate variability. *Physiol. Behav.* **1997**, *62*, 1037–1043. [[CrossRef](#)]

18. Stein, P.; Pu, Y. Cardiac autonomic function during different sleep stages in the elderly: Results from the Sleep Heart Health Study. *Sleep* **2008**, *31*, A98.
19. Herlan, A.; Ottenbacher, J.; Schneider, J.; Riemann, D.; Feige, B. Electrodermal activity patterns in sleep stages and their utility for sleep versus wake classification. *J. Sleep Res.* **2018**, *28*, e12694. [[CrossRef](#)]
20. Johnson, L.C.; Lubin, A. Spontaneous electrodermal activity during waking and sleeping. *Psychophysiology* **1966**, *3*, 8–17. [[CrossRef](#)]
21. Freixa i Baqué, E. Reliability of spontaneous electrodermal activity in humans as a function of sleep stages. *Biol. Psychol.* **1983**, *17*, 137–143. [[CrossRef](#)]
22. Richter, C.P. The significance of changes in the electrical resistance of the body during sleep. *Proc. Natl. Acad. Sci. USA* **1926**, *12*, 214–222. [[CrossRef](#)] [[PubMed](#)]
23. Wei, J.; Boger, J. Sleep detection for younger adults, healthy older adults, and older adults living with dementia using wrist temperature and actigraphy: Prototype testing and case study analysis. *JMIR mHealth uHealth* **2021**, *9*, e26462. [[CrossRef](#)] [[PubMed](#)]
24. Campbell, S.S.; Murphy, P.J. Relationships between sleep and body temperature in middle-aged and older subjects. *J. Am. Geriatr. Soc.* **1998**, *46*, 458–462. [[CrossRef](#)]
25. Quan, S.F.; Howard, B.V.; Iber, C.; Kiley, J.P.; Nieto, F.J.; O'Connor, G.T.; Rapoport, D.M.; Redline, S.; Robbins, J.; Samet, J.M.; et al. The Sleep Heart Health Study: Design, rationale, and methods. *Sleep* **1997**, *20*, 1077–1085. [[CrossRef](#)]
26. Chen, X.; Wang, R.; Zee, P.; Lutsey, P.L.; Javaheri, S.; Alcántara, C.; Jackson, C.L.; Williams, M.A.; Redline, S. Racial/ethnic differences in sleep disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep* **2015**, *38*, 877–888. [[CrossRef](#)]
27. Zhang, G.Q.; Cui, L.; Mueller, R.; Tao, S.; Kim, M.; Rueschman, M.; Mariani, S.; Mobley, D.; Redline, S. The National Sleep Research Resource: Towards a sleep data commons. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1351–1358. [[CrossRef](#)]
28. Gomes, P.; Margaritoff, P.; Silva, H. pyHRV: Development and evaluation of an open-source Python toolbox for heart rate variability (HRV). In Proceedings of the International Conference on Electrical, Electronic and Computing Engineering (IcETRAN), Veliko Gradište, Serbia, 3–6 June 2019; pp. 822–828.
29. Deng, S.; Zhang, X.; Zhang, Y.; Gao, H.; Chang, E.I.C.; Fan, Y.; Xu, Y. Interrater agreement between American and Chinese sleep centers according to the 2014 AASM standard. *Sleep Breath.* **2019**, *23*, 719–728. [[CrossRef](#)]
30. Patel, A.K.; Reddy, V.; Shumway, K.R.; Araujo, J.F. Physiology, sleep stages. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2023. Available online: <https://www.ncbi.nlm.nih.gov/books/NBK526132/> (accessed on 25 July 2023).
31. Dorffner, G.; Vitr, M.; Anderer, P. The effects of aging on sleep architecture in healthy subjects. In *GeNeDis 2014: Geriatrics*; Vlamos, P., Ed.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 93–100. [[CrossRef](#)]
32. Nanchen, D. Resting heart rate: What is normal? *Heart* **2018**, *104*, 1048–1049. [[CrossRef](#)]
33. Hammad, G.; Reyt, M.; Belyi, N.; Baillet, M.; Deantoni, M.; Lesoinne, A.; Muto, V.; Schmidt, C. pyActigraphy: Open-source Python package for actigraphy data visualization and analysis. *PLoS Comput. Biol.* **2021**, *17*, e1009514. [[CrossRef](#)] [[PubMed](#)]
34. Ray, L.B.; Baena, D.; Fogel, S. “Counting Sheep PSG”: EEGLAB-compatible open-source MATLAB software for signal processing, visualization, event marking and staging of polysomnographic data. *J. Neurosci. Methods* **2024**, *407*, 110162. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.