# Sleep Analysis Using Longitudinal Ear-EEG Recordings

Nannapas Banluesombatkul[1], Jesper Strøm[1], Maria Louise Stage Olsen[6,7], Kubra Kilic[6,7], Michael Sprehn[2], Mia Dyhr Thomsen[3], Martin Christian Hemmsen[3], Jonas Bloch Thorlund[4,5], Kaare Mikkelsen[1], Henrik Bjarke Vægter[6,7], and Preben Kidmose[1,†]

***Abstract*— Ear-EEG offers significant advantages for longitudinal sleep studies since it is less intrusive and more user-friendly compared to traditional scalp EEG. The feasibility of longitudinal sleep studies further relies on automated sleep analysis algorithms. This study proposes a systematic and robust procedure for sleep analysis with a key focus on identifying recordings with poor data quality. The method is based on the USleep sleep scoring model and leverages the model's confidence score as a measure of signal quality. This strategy is based on the observation that there is a high correlation between the sleep model's confidence score and Cohen's kappa between manual and model annotations. Notably, the procedure does not rely on manually labeled data or other manual steps. The procedure was evaluated on a novel dataset comprising 576 sleep recordings from 24 chronic pain patients. The procedure distinguished recordings with kappa values above and below 0.6 with an accuracy of 91.9%. Importantly, the exclusion criteria did not systematically eliminate recordings with poor sleep quality metrics, such as low sleep efficiency or frequent sleep stage transitions. Furthermore, the study highlights the benefits of multiple night sleep studies by visualizing the inter-night variability in each subject. In conclusion, the proposed procedure effectively excluded poor-quality recordings, enabling robust analysis of sleep patterns in patients.**

## I. INTRODUCTION

Human sleep exhibits significant night-to-night variability, influenced by factors such as stress [1] and environmental conditions [2]. This variability highlights the limitations of single-night sleep studies, as the selected night could be an outlier and fail to represent an individual's typical sleep patterns. In contrast, multi-night sleep studies capture night-to-night fluctuations, providing a more accurate characterization of sleep architecture and improving the diagnostic reliability of sleep-related conditions.

The gold-standard sleep assessment method, Polysomnography (PSG), is impractical for long-term use due to several limitations. It requires a complex setup, including precise scalp-EEG electrode placement, which is time-consuming and cumbersome. Additionally, scalp-EEG devices may cause discomfort and disrupt sleep. Alternatively, ear-EEG provides a discreet, less intrusive, and user-friendly solution, making it more suitable for long-term, at-home sleep monitoring [3].

Traditionally, sleep stages are scored for every 30 seconds by sleep experts, which is tedious and laborious. Automatic sleep scoring algorithms not only resolve this issue but also allow us to annotate sleep stages on other devices, including ear-EEG. The standard AASM rules for sleep staging were defined by characteristics of EEG signals recorded on the scalp. Previous studies have trained sleep stage classifiers using manual labels on scalp-EEG and showed the possibility of predicting sleep stages on ear-EEG signals with high agreement levels, mainly in young, healthy subjects [4–8].

The challenge of using devices for at-home sleep monitoring is the greater risk of poor data quality [9]. In the research setup, we can perform concurrent scalp-EEG and ear-EEG recordings. Then, we can calculate the kappa values between the model's prediction and manual scoring to validate the reliability of the recording devices and automatic sleep stage classifier. However, in actuality, especially for long-term monitoring, either scalp-EEG or manual scoring will not be available. Therefore, we need another way to assess if a given recording is trustworthy.

Previous research has shown a high correlation between the kappa value and confidence level of the sleep staging model [6]. However, both poor signal quality and disturbed sleep (poor sleep quality) can result in low kappa and low confidence scores. The clinical utility of the methods requires that we should not trust sleep scores from nights with poor signal quality, but we should keep nights with poor sleep quality because these are clinically important.

This study primarily aimed to validate the performance of automatic sleep staging using ear-EEG by comparing it to gold-standard, manually scored scalp-EEG in patients with chronic pain. Secondarily, we aimed to develop an automatic data processing pipeline to identify recordings with poor data quality without manual labels or scalp-EEG recordings. The main objective of the pipeline is to identify recordings with poor data quality with minimal bias towards poor sleep quality. Lastly, we illustrate the advantage of long-term sleep monitoring by exploring the inter-night variability.

[1]Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark.
[2]Department of Anesthesiology and Intensive Care Medicine, Respiration Center South, University Hospital Odense, Odense, Denmark.
[3]T&W Engineering, Allerød, Denmark.
[4]Center for Muscle and Joint Health, Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark.
[5]Research Unit for General Practice, Department of Public Health, University of Southern Denmark, Odense, Denmark.
[6]Pain Research Group, Pain Center South, Department of Anesthesiology and Intensive Care Medicine, University Hospital Odense, Odense, Denmark
[7]Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[†]Corresponding author: pki@ece.au.dk

Manuscript 825 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 8, 2025.

## II. DATASETS

### A. Data Information

The dataset from the Long-term Monitoring of Sleep with Ear-EEG in Patients with Chronic Pain study (NCT06368531) includes long-term sleep monitoring data from 24 chronic pain patients aged 18–58 years, including 20 females. Each subject was asked to wear the ear-EEG solution NeuroBuds from T&W Engineering, Denmark, while sleeping at home preferably 24 nights during a period of 5 weeks. The electrodes include e0, e1 (left ear) and e2, e3 (right ear). The dataset comprises 576 nights (11 to 33 nights per subject). Among these nights, each subject also participated in one PSG sleep assessment (referred to as the "PSG night") conducted at the Respiration Center at Odense University Hospital. During the PSG night, both ear-EEG signals and standard physiological signals—such as scalp-EEG (e.g., C3, C4, F3, F4, O1, O2, M1, M2), EOG (E1, E2), and others—were recorded simultaneously. Sleep stages were annotated in 30-second epochs by a trained sleep technician based solely on the standard PSG data.

### B. Data Preprocessing and Epochs Rejection Criteria

Both ear-EEG and scalp-EEG recordings were preprocessed as follows:

- All samples with NaN values were replaced with zero and marked as $nan$ for further use.
- The intervals where the signal maintained the same value for at least 1 second were identified as $flatline$.
- A zero-phase high-pass filter at 0.1 Hz and a notch filter at 50 Hz were applied to the signals.
- The signals were resampled to 128 Hz to ensure compatibility with the USleep architecture.
- Samples exceeding $\pm 100\mu V$ were marked as $over\_amplitude\_th$.
- Each recording was standardized per channel to have zero mean and unit variance.
- Samples exceeding $\pm$ 20 were marked as $over\_iqr$.
- The final $artifact$ markers were marked at any sample identified as $nan$, $flatline$, $over\_amplitude\_th$, or $over\_iqr$.
- Cross-referencing was calculated for scalp-EEG recordings during this step. For ear-EEG recordings, cross-referencing had already been performed in the original dataset.

Finally, "unscorable" epochs were identified separately for scalp-EEG and ear-EEG recordings. Epochs where more than 50% of the samples were marked as $artifact$ across all channels were considered unscorable. Additionally, epochs with $flatline$ periods exceeding 50% were also marked as unscorable for manual labels.

## III. METHODS

### A. Automatic Sleep Stage Classifier

USleep is a U-Net-based sleep stage classifier that has demonstrated efficiency and generalizability across multiple datasets [10]. In our previous study [11], the USleep model was pre-trained using scalp-EEG and EOG signals from 13 large datasets: ABC, CCSHS, CFS, CHAT, DCSM, HPAP, MESA, MROS, PHYS, SEDF SC, SEDF ST, SHHS, and SOF [12–22]. This pre-trained model, denoted as $\mathcal{M}_{\text{scalp}}$, was then used to predict sleep stages for each 30-second epoch from the scalp-EEG recordings in our dataset.

For ear-EEG recordings, we pre-trained another USleep model with single channel input (EEG only, denoted as $\mathcal{M}_{\text{scalp-EEG}}$) since the EOG was not recorded during the non-PSG nights. Given that mastoid electrodes (M1-M2) are located closer to the ears, they provide signals more similar to ear-EEG than other scalp-EEG electrodes. Furthermore, previous research has demonstrated the effectiveness of using mastoid channels for automatic sleep staging [7]. Therefore, we fine-tuned the $\mathcal{M}_{\text{scalp-EEG}}$ on M1-M2 on the same pre-training datasets when available. This model, denoted as $\mathcal{M}_{\text{ear}}$, was then used to predict the sleep stages for the ear-EEG recordings from our dataset.

### B. Model Prediction

For $\mathcal{M}_{\text{scalp}}$, one EEG channel and one EOG channel were used. We fed combinations of bi-polar EEG channels (C3-M2, C4-M1, F3-M2, F4-M1, O1-M2, and O2-M1) along with EOG channels (E1-M2 and E2-M1) into the model. Similarly, for $\mathcal{M}_{\text{ear}}$, bi-polar ear-EEG channels (e0-e2, e1-e3, e0-e3, and e1-e2) were used.

The final prediction was based on the ensemble of the individual models. More specifically, for $\mathcal{M}_{\text{scalp}}$, predictions from all scalp-EEG and EOG channel combinations were summed, followed by a softmax operation, and the class with the highest probability was chosen as the final prediction. For $\mathcal{M}_{\text{ear}}$, the same process was applied using predictions from all ear-EEG channels.

### C. Confidence Score as a Data Quality Measure

In longitudinal sleep recordings using ear-EEG, there is a certain likelihood of encountering recordings with poor signal quality. Such recordings can lead to inaccurate sleep scoring and may compromise the overall sleep assessment. Therefore, it is important to be able to identify poor-quality recordings to maintain the reliability of the sleep assessment. This assessment of the signal quality should be based on the ear-EEG signal itself.

For each 30-second epoch, the sleep scoring model generates five values representing the probability for each sleep stage. The more certain the model is about a sleep stage, the higher the probability of that stage. Therefore, the maximum probability can be interpreted as a measure of the model's confidence. Previous studies [6] have introduced the concept of using the confidence score as a measure of signal quality and found a high correspondence between kappa and median confidence score. This has been further investigated for predicting the sleep stage classification performance in [23]. This concept is further explored in this study.

A potential pitfall is that various forms of disturbed sleep may also lead to a decrease in confidence. Therefore, the challenge is to find a method to identify recordings with low

Manuscript 825 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 8, 2025.

signal quality while preserving recordings with disturbed or disrupted sleep.

The model confidence score is defined as the maximum probability on each epoch. The median of the scores across the entire night was then calculated to serve as the representative score for each night. For each 30-second epoch $i$, the confidence score $C_i$ is defined as:

$$C_i = \max(P_{i1}, P_{i2}, P_{i3}, P_{i4}, P_{i5}) \tag{1}$$

where $P_{ij}$ represents the probability of the $j$-th sleep stage for the $i$-th epoch.

The representative confidence score for the entire night ($C_{\text{night}}$) is calculated as:

$$C_{\text{night}} = \text{median}(C_1, C_2, \ldots, C_n) \tag{2}$$

where $n$ is the total number of epochs in the night.

However, the model is likely to be less confident during transition periods, as these may involve combinations of multiple sleep stages. As a result, poor sleep quality and increased sleep fragmentation can lead to a lower median confidence score for the night. To mitigate this bias, we selected only the epochs that were not adjacent to different stages. This approach allows us to obtain the confidence score only from the periods where the model should be confident, provided the signal quality is adequate.

### *D. A Systematic Recording Rejection Approach*

A threshold for the confidence score is necessary to distinguish between trustworthy and untrustworthy recordings, ensuring that only recordings with a confidence score above this threshold are included in further analysis. However, the optimal threshold is not universal and depends on the specific goals of the target application. If the high reliability of individual recordings is crucial for clinical assessment, a higher threshold should be set, though this may result in a higher number of rejected recordings. Conversely, if the focus is on having a larger number of nights for clinical analysis, a lower threshold may be more appropriate.

We propose a systematic approach to determine the optimal threshold using a performance metric ($PM$). As an example, we define the performance metric as:

$$PM(t) = \kappa_{P25}(t) \cdot \frac{N_{included}(t)}{N_{total}}, \tag{3}$$

in which $t$ represents the threshold, ranging from 0 to 1, and where recordings with a confidence score lower than the threshold will be rejected; $\kappa_{P25}(t)$ denotes the lower quartile of kappa of the included recordings; and $\frac{N_{included}(t)}{N_{total}}$ is the proportion of retained recordings. This PM balances the trade-off between the minimum acceptable kappa values and the proportion of recordings retained. The 25th percentile of kappa values was used as a measure of acceptable quality since it is less influenced by the excluded recordings.

The optimal threshold was determined as the threshold value that maximized the performance metric. Importantly, this threshold was determined using the PSG nights and subsequently applied to the entire dataset.

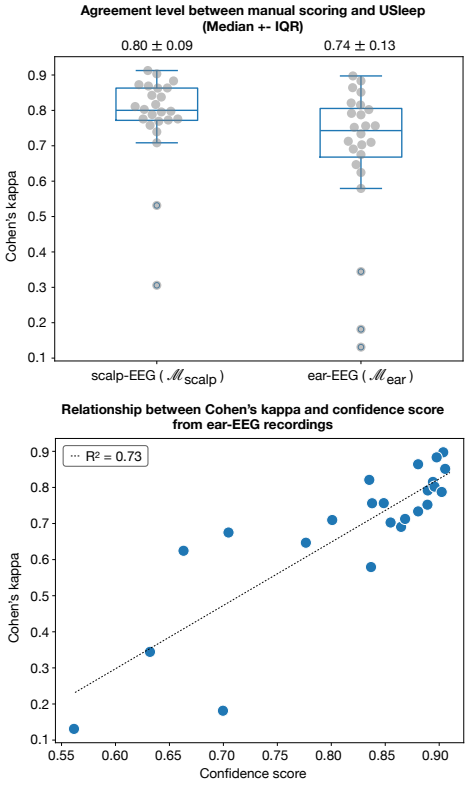


Fig. 1. Upper: Distribution of Cohen's kappa values comparing manual and automatic scoring from 24 PSG nights. Lower: Relationship between confidence score and Cohen's kappa for ear-EEG recordings, demonstrating a strong positive correlation.

### *E. Noise Simulation for Confidence Score Evaluation*

We artificially degraded data quality by adding pink noise at varying levels to the original signals. This experiment was designed to test the hypothesis that if the confidence score accurately reflects data quality, it should decrease as the noise level increases.

For each recording from the PSG nights, we first calculated the standard deviation ($\sigma_{\text{orig}}$) of the signal amplitude for each channel. White noise was then generated with a mean of 0 and a standard deviation of:

$$\sigma_S = S \cdot \sigma_{\text{orig}} \tag{4}$$

where $S$ values were set to 0.05, 0.1, 0.2, 0.5, and 1.

The pink noise was generated from white noise as follows:

1. Compute the Fast Fourier Transform (FFT) of the white noise:

$$X(f) = \mathcal{F}(\text{white\_noise}) \tag{5}$$

2. Apply a $1/f$ filter:

$$X_{\text{filtered}}(f) = X(f) \cdot H(f) \tag{6}$$

$$H(f) = \begin{cases} 0, & f < 0.1 \\ \frac{1}{f}, & 0.1 \leq f \leq 128 \end{cases} \tag{7}$$

Manuscript 825 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 8, 2025.

3. Obtain the time-domain pink noise by applying the inverse FFT:

$$\text{pink\_noise} = \mathcal{F}^{-1}(X_{\text{filtered}}(f)) \tag{8}$$

Finally, the pink noise was scaled to have zero mean and a standard deviation of $\sigma_S$, then added to the original signal.

Since noise was added separately to scalp-EEG and ear-EEG signals, and the $\kappa$ values from each were not intended for direct comparison in this experiment, we included all epochs scorable by both manual annotation and $\mathcal{M}_{\text{scalp}}$ for the scalp-EEG results, and by both manual annotation and $\mathcal{M}_{\text{ear}}$ for the ear-EEG results.

## IV. RESULTS

### A. Sleep Stage Classification Performance

Figure 1 presents the Cohen's kappa ($\kappa$) values between sleep stages scored by the sleep technician and those predicted by the USleep model on our dataset. As described in section II, epochs that were either unscorable by the technicians or unpredictable by the models were excluded for direct comparison between all methods.

The results demonstrate that most scalp-EEG recordings achieved high $\kappa$ values, with a median of 0.80, reaffirming the reliability of the pre-trained USleep model in predicting sleep stages from scalp-EEG data in this unseen cohort. For ear-EEG recordings, $\kappa$ values were slightly lower but still reflected a substantial agreement at a median of 0.74. These findings indicate that automatic sleep scoring is accurate on ear-EEG data, even in non-healthy populations, and underscore the high quality of the ear-EEG recordings.

### B. Confidence Score

The lower panel of Figure 1 shows a strong positive correlation between the model confidence score and $\kappa$. Specifically, in most cases, the confidence score decreased as $\kappa$ values declined. This finding suggests that the confidence score is a reliable proxy for assessing the quality of recordings without the requirement of manual annotations, providing a valuable metric for evaluating data usability.

### C. Effects of Noise-Added Data

After adding various levels of pink noise to the data, the model's predictions were evaluated against manual scoring from the original scalp-EEG data. The results are summarized in Figure 2. The number of sessions included at each noise level (indicated above each bar) varied because epochs meeting the rejection criteria (subsection II-B) were excluded, and in some recordings, no usable epochs remained for analysis. In total, the synthetic dataset comprises 131 scalp-EEG and 136 ear-EEG recordings, with 24 actual recordings for each method.

As illustrated in Figure 2 (upper), $\kappa$ values declined consistently as noise levels increased. The trend was observed across both scalp-EEG and ear-EEG recordings. This confirms that higher noise levels degrade signal quality and reduce the model's ability to predict sleep stages accurately.

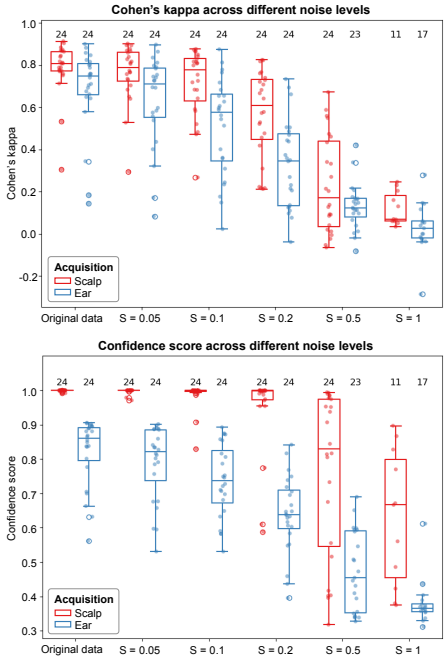


Fig. 2. Distribution of Cohen's kappa (top) and confidence score (bottom) across different noise levels. The numbers above the bars indicate the corresponding number of recordings. As the noise level increases, more epochs are rejected due to poor data quality. Consequently, some recordings have no remaining epochs, reducing the total number of recordings.

Similarly, Figure 2 (lower) illustrates a significant reduction in model confidence scores as noise levels increase. This consistent decline in confidence scores with deteriorating data quality highlights their utility as a robust metric for assessing signal quality.

However, the confidence scores for scalp-EEG and ear-EEG declined at different rates. While the confidence score for ear-EEG consistently decreased with increasing noise levels, scalp-EEG maintained a high confidence score until $S$ = 0.5. This suggests that the confidence score threshold (subsection IV-D) should be considered separately for each modality.

### D. Optimal Threshold for Recording Rejection

The lower panel of Figure 3 illustrates the relationship between confidence scores and $\kappa$ values for both actual and synthetic (noise-added) data, reaffirming the strong correlation between these two measures.

The results of our proposed procedure are displayed in the upper panel. The top figure highlights a clear trade-off between $\kappa$ values and the number of recordings included at each threshold. Our systematic approach identified a confidence score threshold of 0.75 as optimum, balancing this trade-off effectively. At this threshold, 36.8% of recordings were retained, achieving $\kappa_{P25}$ of 0.52, which resulted in the best performance metric ($PM$) of 0.19.

To evaluate the selected threshold, we applied it to all ear-EEG recordings depicted in the lowest panel of Figure 3. When we defined the recordings with $\kappa$ exceeding 0.6 as reliable recordings (due to substantial agreement between

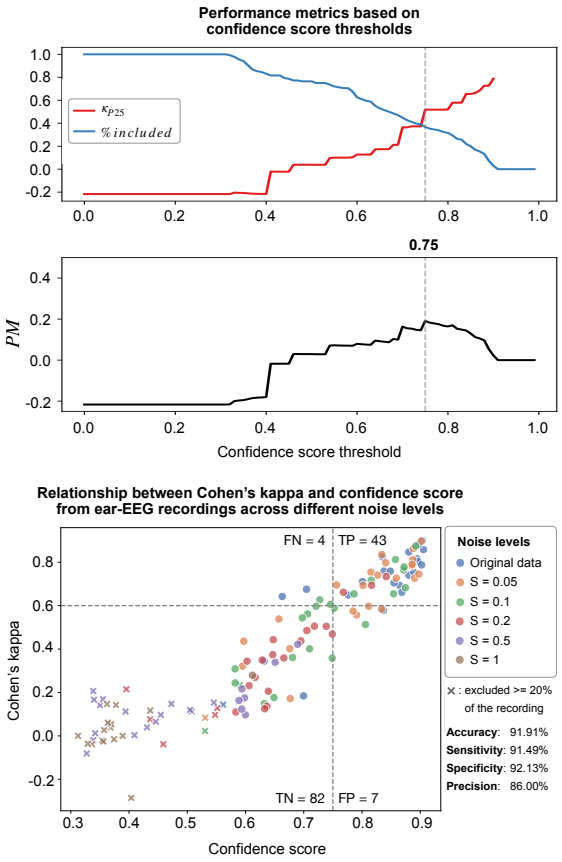


Fig. 3. Visualization of performance metrics across confidence score thresholds (upper) and the scatter plot, illustrating the correlation between $\kappa$ values and confidence scores for both actual and simulated data (lower). The gray dashed line in the upper panel indicates the threshold determined for the optimal performance metric, which is also applied to the lower figure.

model and manual scoring [24]), our procedure effectively identified them with an accuracy of 91.9%, a sensitivity of 91.5%, and a specificity of 92.1%. These results demonstrate the robustness of our approach in rejecting low-quality data.

### *E. Sleep Metrics Comparison between Scorers*

Sleep stage classification provides valuable insights into patients' sleep architecture, enabling the calculation of various sleep metrics. In this study, we focused on key metrics, including total sleep time (TST), sleep efficiency (SEFF), the proportion of REM versus NREM epochs per night (REM_NREMRATIO), and the number of sleep stage transitions from sleep onset to final awakening (STAGEC), as representative examples. These metrics were computed from manual scoring, $\mathcal{M}_{\text{scalp}}$, and $\mathcal{M}_{\text{ear}}$. Among the 24 PSG nights, the threshold identified five as poor-quality recordings, which were therefore excluded from this analysis.

A two-sided pairwise permutation test was performed to assess statistical differences between each pair of methods. As shown in Figure 4, the only significant difference between the model prediction and manual annotation was observed in the number of stage transitions. This suggests that both models tend to predict more fragmented sleep compared to manual scoring, which was also observed by the hypnogram

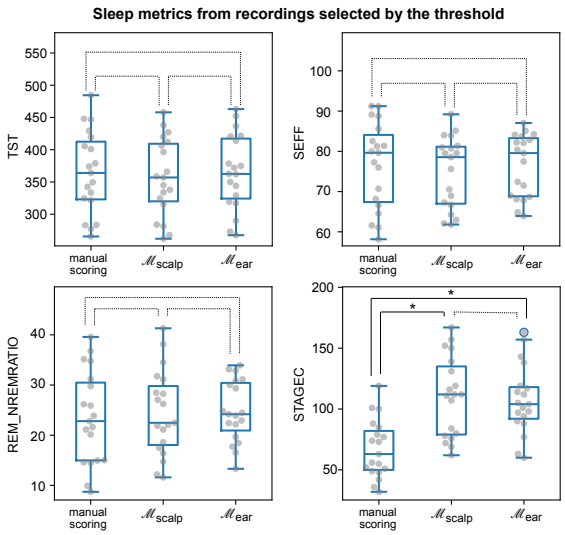


Fig. 4. Examples of sleep metrics, including Total Sleep Time (TST), Sleep Efficiency (SEFF), the ratio of REM to NREM periods (REM_NREMRATIO), and the number of stage transitions (STAGEC), derived from recordings selected by the threshold (19 out of 24 recordings). The solid line with an asterisk (*) indicates a significant difference between each pair ($p \leq 0.01$), while the dotted line represents an insignificant difference.

comparison. For TST, SEFF, and REM_NREMRATIO, sleep stage predictions from all methods yielded similar values.

### *F. Effects of Exclusion Criteria on Sleep Metrics*

The exclusion procedure consisted of two steps. First, we calculated the percentage of rejected epochs per electrode, as described in subsection II-B. Recordings with an average of rejected epochs exceeding 20% were excluded. Second, we excluded recordings with confidence scores below the predefined threshold of 0.75. Together, these steps led to the exclusion of 139 out of 576 ear-EEG recordings, representing 24.1% of the total dataset.

Figure 5 illustrates the effects of this procedure. The first four panels compare the sleep metrics before and after exclusion, while the lowest panel shows the percentage of excluded sessions for each subject. Each dot represents the sleep metric value from a recording. The results indicate that the exclusion procedure had minimal impact on the median sleep metrics for most subjects, as represented by the horizontal line within each group. However, for subjects with more than 50% of recordings removed, the changes were more pronounced. These findings demonstrate the procedure's effectiveness in retaining only reliable recordings without significantly altering subject-level sleep metrics.

### *G. The Advantages of Multiple-night Sleep Study*

Figure 5 reveals notable variations in sleep metrics across multiple nights for each subject. This highlights the value of multiple-night sleep studies in providing a more comprehensive understanding of the individual variations in sleep patterns. The dark stars, representing sleep metrics derived from the PSG nights, further demonstrate that a single PSG night may not accurately capture a subject's typical sleep

Manuscript 825 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 8, 2025.

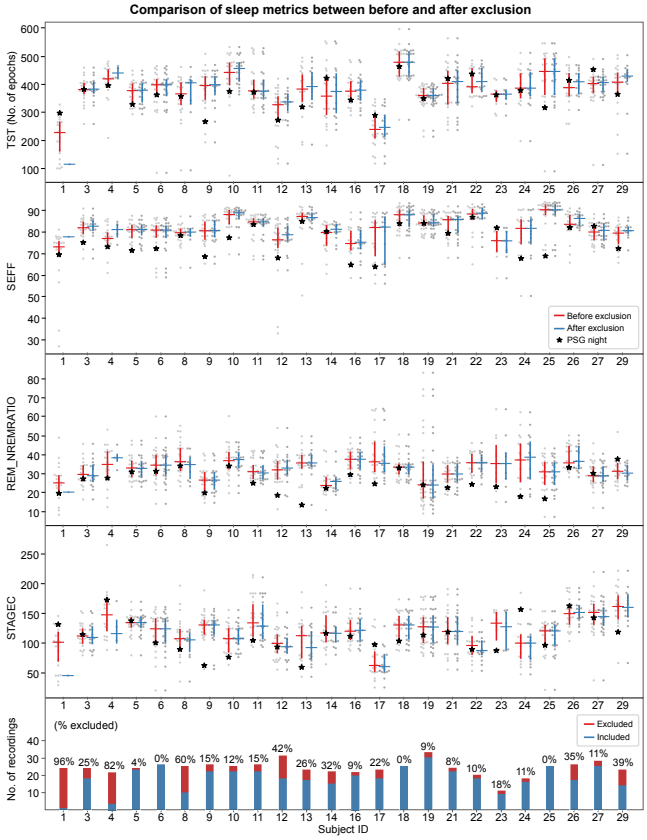


Fig. 5. The distribution of sleep metrics for each subject before and after applying the exclusion procedure, along with the proportion of included and excluded recordings per subject (the lowest panel). The sleep metrics from the PSG nights were calculated from ear-EEG recordings.

characteristics. These findings align with previous research [25] and emphasize the advantages of long-term sleep assessment, where ear-EEG devices can offer significant benefits.

### *H. Sleep Quality*

Our proposed procedure excludes poor-data-quality recordings with minimal bias on sleep quality for the following reasons. First, the model confidence score is not correlated with sleep quality, as indicated by the $R^2$ values of 0.01, 0.02, 0.03, and 0.13 between the confidence score and TST, SEFF, REM_NREMRATIO, and STAGEC, respectively. Second, we minimized the potential for this bias by using only non-transitional epochs. This ensures that higher transitional stages, which generally have lower values, do not influence the overall confidence score. Finally, Figure 5 shows that our procedure does not systematically reject nights with poor sleep quality, such as low TST, low SEFF, or high STAGEC.

## V. DISCUSSION

Since ear-EEG devices offer significant advantages for long-term sleep monitoring, particularly in at-home settings, they provide an opportunity for more comprehensive assessments of sleep behaviors. However, the trade-off is a potential decrease in data quality, necessitating a systematic approach to exclude unreliable recordings, especially in the absence of manual scoring. We proposed a procedure based on the model confidence score to address this issue and demonstrated its efficacy in excluding poor-quality recordings with minimal bias from poor sleep quality.

We defined reliable recordings based on $\kappa$ values, where a high agreement between the model and manual scoring indicates data reliability. Low $\kappa$ values can result from either poor data quality or model limitations on specific recordings. Regardless of the underlying cause, those predictions are inherently untrustworthy and should not be included in further analyses.

The confidence score for each epoch was derived from the maximum probability among the five classes. Therefore, it may overlook essential information information reflected in the prediction probability distribution. To address this, we also explored alternative measures such as prediction entropy and cosine similarity, which quantify the prediction's uncertainty and its similarity to manual-scoring consensus, respectively. However, when calculating performance metrics ($PM$) across thresholds and comparing between measures, the confidence score consistently yielded the highest Area Under the Curve. Moreover, after computing the median confidence score from the entire night, all measures were highly correlated. Overall, this indicated that the alternative measures did not contribute significantly to the identification of recordings with low data quality, and were thus left out of the procedure.

Interestingly, ear-EEG recordings from 5 out of 24 PSG nights were excluded by the confidence score despite typically being higher quality than home recordings. These rejected recordings had $\kappa$ of 0.13, 0.18, 0.34, 0.62, and 0.68, while the included sessions mostly got $\kappa > 0.64$, except for one recording with 0.58.

One of them clearly had poor data quality, with more than 30% of epochs rejected. The other two recordings, with 18–19% of epochs rejected, were excluded due to falling below the confidence score threshold. This suggests that, for sessions where the proportion of rejected epochs is borderline, the confidence score provides additional support in making exclusion decisions.

Another rejected PSG night had a moderate confidence score (0.7) but an exceptionally low $\kappa$ (0.18). While our procedure successfully excluded this recording, it was near the rejection threshold. A similarly low $\kappa$ was also observed between manual scoring and the model's prediction on the scalp-EEG data for this recording. To better understand this case, we conducted a post hoc analysis. We consulted the scorer, who noted that the recording was particularly challenging to score due to numerous transitions, including a prolonged period of transitioning between the Wake and N1 stages. Additionally, many N2 epochs resembled the REM stage but without muscle atonia.

This recording also demonstrates a scenario where the USleep model struggled to provide accurate predictions. Such outliers are inevitable in any dataset. However, they highlight the advantage of multiple-night sleep studies, where insights from other reliable nights can help offset their

impact. If these challenges arise from subject variability, incorporating personalized techniques could be a promising approach for future work.

The last rejected PSG night had a $\kappa$ of 0.68 but was excluded due to a confidence score of 0.70. The confidence score is relatively low because two out of four electrodes exhibited poor signal quality. This is an example of a recording that can achieve higher $\kappa$ when using one electrode instead of the ensemble method. Therefore, optimizing electrode weighting could be an interesting direction for future research.

Our approach paves the way for further analysis in the next step, where we can examine the longitudinal ear-EEG sleep recordings and explore parameters related to chronic pain patients. However, a limitation of this study is that we did not apply the pipeline to any other datasets. Therefore, future work should involve validating the pipeline on additional datasets to assess its generalizability and robustness across diverse populations.

## VI. CONCLUSION

The proposed procedure for sleep analysis was evaluated on a dataset comprising 576 sleep recordings from 24 chronic pain patients. A single night of concurrent PSG and ear-EEG was recorded from each patient, yielding a median Cohen's kappa of 0.74 between manual PSG scoring and model-based ear-EEG scoring. On this part of the data, the procedure distinguished recordings with kappa values above and below 0.6 with an accuracy of 91.9%. On the entire dataset, 139 out of 576 recordings (24.1%) were identified as having low signal quality, resulting in 437 (75.9%) nights with reliable hypnograms. Importantly, the procedure did not systematically reject recordings with sleep metrics indicating disturbed sleep, suggesting that the procedure identifies recordings with low data quality without discarding recordings with low sleep quality.

## REFERENCES


[1] E.-J. Kim and J. E. Dimsdale, "The effect of psychosocial stress on sleep: A review of polysomnographic evidence," *Behavioral sleep medicine*, vol. 5, no. 4, pp. 256–278, 2007.

[2] J.-H. Byun *et al.*, "The first night effect during polysomnography, and patients' estimates of sleep quality," *Psychiatry research*, vol. 274, pp. 27–29, 2019.

[3] K. B. Mikkelsen *et al.*, "Self-applied ear-eeg for sleep monitoring at home," in *2022 44th annual international conference of the IEEE engineering in Medicine & Biology Society (EMBC)*, IEEE, 2022, pp. 3135–3138.

[4] K. B. Mikkelsen *et al.*, "Automatic sleep staging using ear-eeg," *Biomedical engineering online*, vol. 16, pp. 1–15, 2017.

[5] Y. R. Tabar *et al.*, "At-home sleep monitoring using generic ear-eeg," *Frontiers in neuroscience*, vol. 17, p. 987 578, 2023.

[6] K. B. Mikkelsen *et al.*, "Accurate whole-night sleep monitoring with dry-contact ear-eeg," *Scientific reports*, vol. 9, no. 1, p. 16 824, 2019.

[7] K. B. Mikkelsen *et al.*, "Sleep monitoring using ear-centered setups: Investigating the influence from electrode configurations," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 5, pp. 1564–1572, 2021.

[8] Y. R. Tabar *et al.*, "Ear-eeg for sleep assessment: A comparison with actigraphy and psg," *Sleep and Breathing*, vol. 25, pp. 1693–1705, 2021.

[9] M. Braun *et al.*, "A systematic review on the technical feasibility of home-polysomnography for diagnosis of sleep disorders in adults," *Current Sleep Medicine Reports*, pp. 1–13, 2024.

[10] M. Perslev *et al.*, "U-sleep: Resilient high-frequency sleep staging," *NPJ digital medicine*, vol. 4, no. 1, p. 72, 2021.

[11] J. Strøm, A. L. Engholm, K. P. Lorenzen, and K. B. Mikkelsen, "Common sleep data pipeline for combined data sets," *Plos one*, vol. 19, no. 8, e0307202, 2024.

[12] J. P. Bakker *et al.*, "Gastric banding surgery versus continuous positive airway pressure for obstructive sleep apnea: A randomized controlled trial," *American journal of respiratory and critical care medicine*, vol. 197, no. 8, pp. 1080–1083, 2018.

[13] G.-Q. Zhang *et al.*, "The national sleep research resource: Towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 2018.

[14] C. L. Rosen *et al.*, "Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: Association with race and prematurity," *The Journal of pediatrics*, vol. 142, no. 4, pp. 383–389, 2003.

[15] S. Redline *et al.*, "The familial aggregation of obstructive sleep apnea.," *American journal of respiratory and critical care medicine*, vol. 151, no. 3, pp. 682–687, 1995.

[16] C. L. Marcus *et al.*, "A randomized trial of adenotonsillectomy for childhood sleep apnea," *New England Journal of Medicine*, vol. 368, no. 25, pp. 2366–2376, 2013.

[17] C. L. Rosen *et al.*, "A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: The homepap study," *Sleep*, vol. 35, no. 6, pp. 757–767, 2012.

[18] X. Chen *et al.*, "Racial/ethnic differences in sleep disturbances: The multi-ethnic study of atherosclerosis (mesa)," *Sleep*, vol. 38, no. 6, pp. 877–888, 2015.

[19] T. Blackwell *et al.*, "Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: The osteoporotic fractures in men sleep study," *Journal of the American Geriatrics Society*, vol. 59, no. 12, pp. 2217–2225, 2011.

[20] B. Kemp *et al.*, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the eeg," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.

[21] S. F. Quan *et al.*, "The sleep heart health study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[22] A. P. Spira *et al.*, "Sleep-disordered breathing and cognition in older women," *Journal of the American Geriatrics Society*, vol. 56, no. 1, pp. 45–50, 2008.

[23] K. B. Mikkelsen, Y. R. Tabar, and P. Kidmose, "Predicting sleep classification performance without labels," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2020, pp. 645–648.

[24] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[25] T. W. Kjaer *et al.*, "Repeated automatic sleep scoring based on ear-eeg is a valuable alternative to manually scored polysomnography," *PLOS Digital Health*, vol. 1, no. 10, e0000134, 2022.