

Expert-level probabilistic breathing event detector informs phenotyping of sleep apnea

Corresponding Author: Professor Emmanuel Mignot

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This study focuses on detecting sleep apnea-related respiratory events through the use of a deep learning approach and up to 7 PSG overnight signals (plus data from EEG, EOG, and EMG). The topic is of interest for sleep research, the objectives are well-proposed, and the manuscript is well-written. The authors establish that their approach is useful in sleep apnea detection, mainly because of its ability to properly differentiate the several types of respiratory events (including a new event type), thus covering a gap in the related literature. Despite the interest of the proposal, it is my opinion that several major and minor issues should be addressed prior to the publication of the manuscript:

Major issues:

1/ Apneic event length is 10 seconds or more. Therefore, the frequency associated with their occurrence is below 0.1 Hz. Accordingly, respiratory signals being filtered between 0.1 Hz and 4 Hz are losing very important information. Several studies report that the most useful sleep apnea-related information in respiratory signals is below 0.1 Hz. It is my opinion that filtering should be changed to between 0.025 Hz and 4 Hz to improve the performance of the model, which is high but not impressive.

2/ I understand the advantages of the current proposal regarding previous approaches. However, it is my opinion that the manuscript fails to compare the diagnostic performance (not the scoring performance) with that reached by many other proposals. There exist previous studies using thousands of subjects but only using single channels like SpO2 that reach diagnostic performances that cannot be seen as inferior to those reported here using 7 + 3 signals. As diagnosis is the major conceptual goal of these automated approaches, the authors should compare and discuss these results with other relevant proposals in this particular task.

3/ As there are deep learning solutions that do not require an equal sample rate for the signals, and 8 Hz is below the sampling rate recommended by the AASM for respiratory signals, this choice should be justified and acknowledged as a limitation of the study.

4/ I think that the choice of the particular deep learning architecture should be justified and discussed. The existence of literature pointing to CNNs as reaching higher performances than RNNs, even in tasks involving signal processing (time series), justifies clearly explaining the reasons behind the choice of the deep learning architecture (which in this case includes a combination of both approaches via ResNet).

5/ The intra-class correlation coefficient is better suited than Pearson's correlation to measure agreement between different scorers (with the predicted AHI being one of them).

Minor issues:

- "To simplify any annotated mixed apneas were labelled as OA due to there being an effort to breathe during the event." It is my opinion that this is not a simplification but a limitation (probably due to an unavoidable class imbalance concerning this type of event)

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The authors present an assessment of a new automatic sleep disordered breathing event detector (ABED) for identifying respiratory events in sleep disordered breathing. They report that generally the ABED performs as well as individual expert scores, particularly for obstructive apneic events. Data from a number of different, existing cohorts were used to train and validate the ABED, which is a strength of the study.

The potential of this is that the very rapid inference time (14 seconds per average PSG is cited) can increase efficiency and scope of sleep apnea testing.

A number of possible areas for clinical translation are given. I think the most attractive is in easily identifying endotypes of obstructive sleep apnea. Endotypic traits have been proposed for use in selecting patients that might benefit preferentially from a range of non-CPAP alternative therapies including oral appliances, hypoglossal nerve stimulation, supplemental oxygen, and pharmacotherapy. Given there is currently a one-size-fits-all management approach that predominates where CPAP is first line treatment for almost all patients. Accurately identifying patients likely to respond to alternatives to CPAP to avoid a trial and error approach to alternatives would be a major advance. Systems to identify endotypes from routine PSG signals with automated modelling have been developed and some are referenced here (21-25). However, the authors have not referenced further work that has demonstrated some of these systems are highly accurate in predicting, for example, oral appliance treatment outcomes in obstructive sleep apnea (see Dutta et al, J Clin Sleep Med 2022), or that combine clinical data and models from PSG data (see Eckert et al, Sleep 2024;47 (Suppl 1): A257) to identify successful candidates for Oral Appliance Therapy (and vice versa). The authors assert that the addition of "apnotyping" via probabilistic detection of event subtypes within each majority class could benefit sleep apnea endotyping. It remains an open question whether the ABED will be as effective as these methods in clinical practice as this should be acknowledged in the limitations. Further research is also needed to demonstrate if endotypes are useful in predicting which patients are at risk of adverse health consequences of OSA. This is beyond the scope of the current study.

The authors also do not reference previous work on the significance of isolated respiratory events without arousals or desaturations (IRE). Previous work has shown non-apneic REs without cortical arousal or desaturation are associated with significant respiratory and heart rate changes. And, that optimal CPAP and the reduction of resistive load are associated with the normalization of heart rate indicating potential clinical benefit (for example see Sankari et al, Sleep 2016 ;40:zsw024). Easy identification of these events would be useful but is not a novel finding.

There are a number of other limitations I feel need to be considered in addition to those given in the Discussion. As the authors point out in the Introduction, a number of novel metrics taken from the PSG have been identified that have shown to be better predictors of downstream consequences of OSA than the AHI. These include hypoxic burden and a number of metrics derived from heart rates. Further, the intensity of arousals from events are also likely to be clinically significant. It does not appear that the ABED identifies any of these metrics. Consequently, the value of the ABED in identifying which patients may benefit long-term from therapy is uncertain. Sleep studies also require scoring of sleep stages and identification of arousals due to limb movements or that occur spontaneously. I presume the ABED would need to be utilised in addition to automatic scoring software for these features.

Overall, the paper presents a potentially useful addition to automate the scoring of sleep-disordered breathing events that will likely be of value in sleep medicine.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have addressed all my concerns

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

No further comments

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

We thank the reviewers for their valuable suggestions. The following response will be formatted with Reviewer comments in blue and Responses in black.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This study focuses on detecting sleep apnea-related respiratory events through the use of a deep learning approach and up to 7 PSG overnight signals (plus data from EEG, EOG, and EMG). The topic is of interest for sleep research, the objectives are well-proposed, and the manuscript is well-written. The authors establish that their approach is useful in sleep apnea detection, mainly because of its ability to properly differentiate the several types of respiratory events (including a new event type), thus covering a gap in the related literature. Despite the interest of the proposal, it is my opinion that several major and minor issues should be addressed prior to the publication of the manuscript:

Major issues:

1/ Apneic event length is 10 seconds or more. Therefore, the frequency associated with their occurrence is below 0.1 Hz. Accordingly, respiratory signals being filtered between 0.1 Hz and 4 Hz are losing very important information. Several studies report that the most useful sleep apnea-related information in respiratory signals is below 0.1 Hz. It is my opinion that filtering should be changed to between 0.025 Hz and 4 Hz to improve the performance of the model, which is high but not impressive.

We thank the reviewer for this thoughtful and important comment regarding the low-frequency content of respiratory signals.

We would like to clarify that although apneic events typically last ≥ 10 seconds, their physiological signatures - particularly reductions in airflow or effort - are reflected in modulations of the signal envelope rather than the generation of new frequency components below 0.1 Hz. In our experience, these envelope changes remain detectable even after applying a 0.1 Hz high-pass filter, since the base breathing frequencies (typically $\sim 0.2\text{--}0.4$ Hz) are still present.

To empirically assess the reviewer's suggestion, we retrained and evaluated our model using a 0.025-4 Hz band-pass filter and compared the results to our original 0.1-4 Hz filtering. The results across four cohorts (MESA, MrOS, WSC, CFS) show minimal and inconsistent performance differences. As shown in rebuttal tables 1 and 2, there is no consistent improvement when including frequencies below 0.1 Hz; in fact, for some

event types (e.g., HYPO per event, OA per recording), performance is higher with the 0.1 Hz cutoff.

Rebuttal table 1: Per event across all PSGs:

F1 score	0.025 - 4 Hz	0.1 - 4 Hz
OA	0.73	0.71
CA	0.52	0.51
HYPO	0.55	0.65
IRE	0.53	0.47

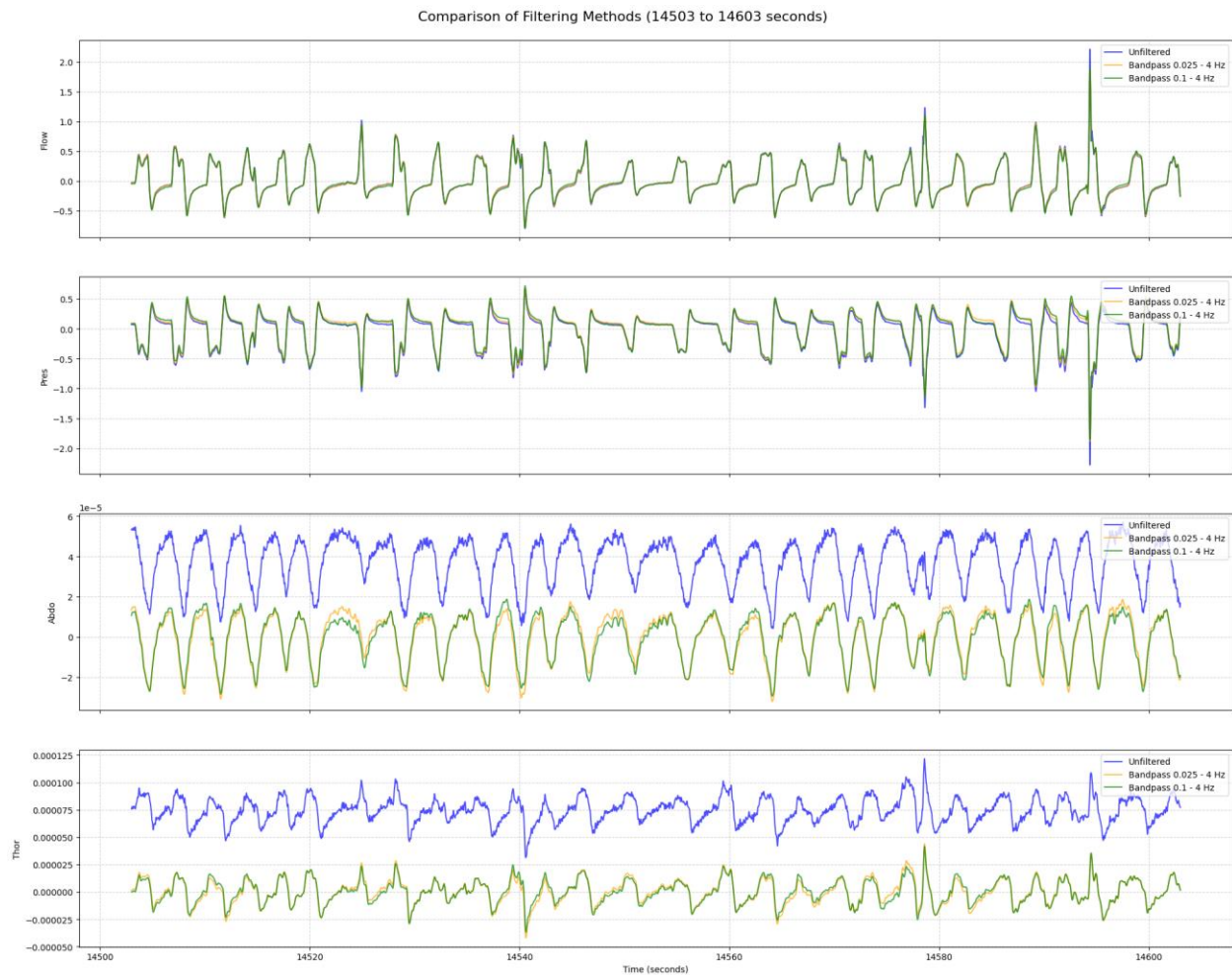
Rebuttal table 2: Per event mean by PSG:

F1 score	0.025 - 4 Hz	0.1 - 4 Hz
OA	0.52	0.57
CA	0.51	0.51
HYPO	0.60	0.55
IRE	0.49	0.44

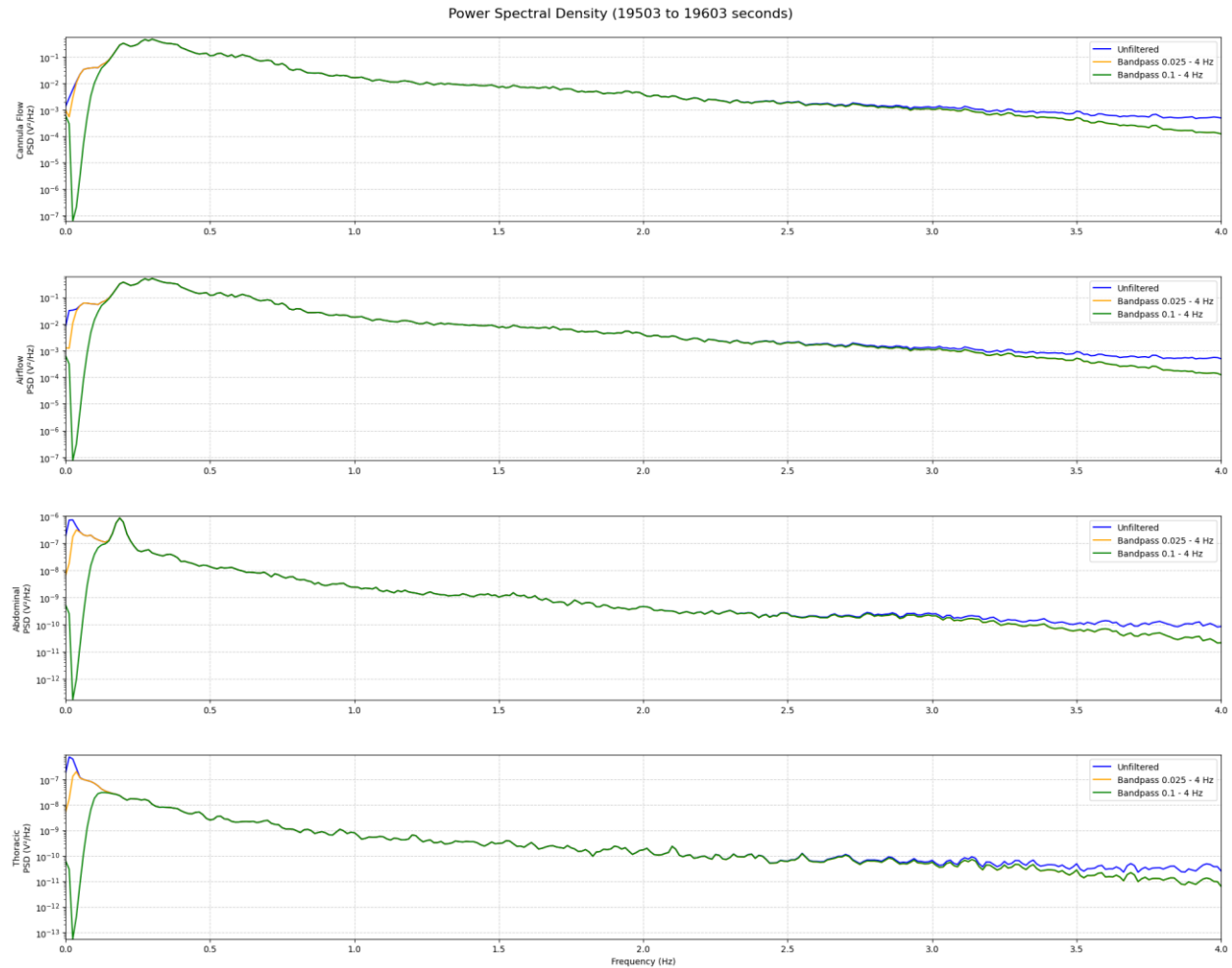
We also include Rebuttal Figure 1, which shows filtered (green = our chosen filter, orange = filter proposed by reviewer) vs. unfiltered (blue) examples of respiratory effort signals during two apneic events separated by recovery breaths.

The high similarity across traces between 0.025-4 Hz and 0.1-4 Hz filters suggests that the clinically relevant apnea signatures remain intact. In Rebuttal Figure 2 we see the

spectrogram of the recording and the attenuation of the signal below 0.1 Hz for the orange signal is prior to the first peak



Rebuttal Figure 1: Time series representation of respiratory signals with different filter settings. Blue: unfiltered, Green: original filtering (bandpass: 0.1 - 4hz), Orange: proposed changed filtering (bandpass: 0.025 - 4hz)



Rebuttal Figure 2: The power spectrum of signals with different filter settings. Blue: unfiltered, Green: original filtering (bandpass: 0.1 - 4hz), Orange: proposed changed filtering (bandpass: 0.025 - 4hz)

Finally, we note that the MrOS study applied a hardware high-pass filter at 0.15 Hz to most respiratory channels (except nasal flow), whereas MESA and CFS did not apply such filtering. To harmonize preprocessing across cohorts and avoid cohort-specific bias in model training, we chose a conservative high-pass cutoff of 0.1 Hz. This tradeoff supports generalization while still preserving the envelope information necessary for apnea detection and has minimal influence on results across populations.

2/ I understand the advantages of the current proposal regarding previous approaches. However, it is my opinion that the manuscript fails to compare the diagnostic performance (not the scoring performance) with that reached by many other proposals. There exist previous studies using thousands of subjects but only using single channels like SpO2 that reach diagnostic performances that cannot be seen as inferior to those

reported here using 7 + 3 signals. As diagnosis is the major conceptual goal of these automated approaches, the authors should compare and discuss these results with other relevant proposals in this particular task.

We thank the reviewer for raising this important point regarding diagnostic performance. We agree that comparing our model's diagnostic utility with simpler approaches is necessary to contextualize its relevance and impact.

Indeed, previous studies have demonstrated strong diagnostic performance using single-channel inputs such as SpO2 [<https://www.nature.com/articles/s41467-023-40604-3>], achieving high ICC and F1 metrics for AHI estimation at (ICC: 0.92, F1: 0.78), (ICC: 0.94, F1: 0.74), and (ICC: 0.94, F1: 0.75) for CFS, MrOS, and MESA, respectively. Comparatively, we achieve ICC = 0.92 and F1 = 0.74. These models benefit from ease of use and scalability, and we agree that their contributions are valuable - especially in population-level screening contexts.

In contrast, our approach is explicitly designed to go beyond estimation of AHI, which have been increasingly criticized as insufficient for characterizing apnea severity and guiding personalized treatment. Our model leverages multichannel PSG data to provide event-by-event classification, differentiation of apnea events (obstructive, central, hypopnea, and isolated respiratory events), and the ability to inform apnea endotyping.

Rather than only predicting whether a patient exceeds an AHI threshold, our model characterizes the temporal structure and type of respiratory disturbances, which are more closely associated with physiological burden and clinical outcomes. Several studies emphasize the importance of phenotyping sleep apnea patients beyond the AHI - for instance, identifying arousability, hypoxic burden, or event clustering - which are not easily captured by single-channel or threshold-based methods.

We have added a discussion of this comparison to the revised manuscript, highlighting the trade-offs between simplicity and diagnostic richness. Ultimately, we see our work as complementary: while single-channel systems are suitable for screening, detailed multi-channel models like ours offer higher resolution insights needed for personalized interventions and advancing our understanding of sleep apnea pathophysiology.

3/ As there are deep learning solutions that do not require an equal sample rate for the signals, and 8 Hz is below the sampling rate recommended by the AASM for respiratory signals, this choice should be justified and acknowledged as a limitation of the study.

We thank the reviewer for this insightful comment. We have now added both the rationale for our choice and a corresponding limitation statement in the revised manuscript.

Our choice of an 8 Hz sampling rate reflects a trade-off between computational efficiency and retaining signal information across a large, multi-cohort dataset. This sampling rate is sufficient to capture the primary respiratory dynamics relevant to sleep apnea detection, which predominantly occur in the 0.2-0.33 Hz range, corresponding to normal adult breathing rates ($\approx$ 12-20 breaths per minute)

[<https://pubmed.ncbi.nlm.nih.gov/8774258/>]. For thoracic, abdominal, and oral thermistor signals - where respiratory effort and airflow patterns are relatively slow - 8 Hz ensures adequate resolution while minimizing training costs. For nasal airflow, we note that although certain high-frequency components like snoring vibrations occur $>$ 100 Hz [<https://www.nature.com/articles/srep30559>], our approach does not rely on snoring acoustics, and the 8 Hz rate captures the signal envelope adequately for the purposes of apnea event classification.

However, we acknowledge the reviewer's point that the AASM recommends higher sampling rates ($\geq$ 25 Hz) for respiratory channels to ensure accurate waveform morphology, particularly in clinical diagnostic settings. While our approach performs well at 8 Hz, this limitation should be considered in contexts where finer temporal resolution may be required, especially for detecting subtle respiratory patterns or developing models for real-time clinical applications.

We now acknowledge this as a limitation in the manuscript and recommend that future work explore whether using higher sampling rates could further enhance model performance, particularly in prospective datasets.

4/ I think that the choice of the particular deep learning architecture should be justified and discussed. The existence of literature pointing to CNNs as reaching higher performances than RNNs, even in tasks involving signal processing (time series), justifies clearly explaining the reasons behind the choice of the deep learning architecture (which in this case includes a combination of both approaches via ResNet).

We thank the reviewer for the helpful comment encouraging a deeper justification of our model architecture which we have added to the Manuscript. The ABED architecture was designed to leverage both spatial and temporal signal characteristics through a hybrid structure that combines a convolutional feature extractor (a modified ResNet50) with a

temporal modeling layer (bi-directional LSTM). This design reflects the nature of the input data—multimodal time series from respiratory, oxygen saturation, and arousal-related signals—where both short-term local features and longer-range dependencies are important.

Convolutional neural networks (CNNs), particularly residual networks (ResNets), are known for their ability to efficiently learn spatial hierarchies in signal patterns and are less prone to overfitting compared to simpler CNNs or multilayer perceptrons (MLPs), especially when deep architectures are used thanks to residual connections [He et al., 2016, Ref. 69]. We selected ResNet due to its scalability, stable gradient flow, and demonstrated success in prior applications [He et al., 2016, Ref. 69].

While some studies suggest that CNNs alone can outperform recurrent models for time series tasks [Van Steenkiste et al., 2019, Ref. 31], these often rely on single-modality input or short temporal contexts. In contrast, sleep apnea involves complex temporal dependencies - such as the delayed relationship between respiratory flow reduction, oxygen desaturation, and arousal - that are difficult to capture without explicit sequence modeling. The distinction between a hypopnea and an isolated respiratory event (IRE) relies not just on morphology, but on when these features co-occur. Therefore, temporal modeling is critical.

To that end, we employed a bi-directional LSTM to capture both forward and backward temporal dependencies within 240-second windows, allowing the model to integrate context across time and across channels. This is essential in sleep breathing analysis, where the onset and consequence of an event may be separated by several tens of seconds. We also considered transformer-based models but found them less suitable for our mid-scale dataset due to their high parameter count and increased risk of overfitting without massive datasets.

5/ The intra-class correlation coefficient is better suited than Pearson's correlation to measure agreement between different scorers (with the predicted AHI being one of them).

We thank the reviewer for this insightful suggestion. We agree that the intra-class correlation coefficient provides a more appropriate measure of agreement between true and predicted AHI values as it also takes into account the bias of the model. We have calculated the ICC of true and predicted AHI and get $ICC = 0.92(95\% \text{ CI: } [0.90 \text{ } 0.93])$. We have now added these to the revised manuscript in Figure 1e.

Minor issues:

- "To simplify any annotated mixed apneas were labelled as OA due to there being an effort to breathe during the event." It is my opinion that this is not a simplification but a limitation (probably due to an unavoidable class imbalance concerning this type of event)

We thank the reviewer for this observation and agree that the handling of mixed apneas represents a limitation rather than a simplification. As noted, the primary reason for this decision was the significant class imbalance: mixed apneas constituted a very small fraction of events across all cohorts, making it challenging to train a robust classifier for this category. Mixed apneas were labeled as OA due to the partial presence of respiratory effort.

However, we acknowledge that mixed apneas may have distinct pathophysiological significance and that their labeling as OA reduces diagnostic granularity. We have updated the limitations section of the manuscript to reflect this point.

Reviewer #2 (Remarks to the Author):

We thank the reviewer for their insightful comments and suggestions.

The authors present an assessment of a new automatic sleep disordered breathing event detector (ABED) for identifying respiratory events in sleep disordered breathing. They report that generally the ABED performs as well as individual expert scores, particularly for obstructive apneic events. Data from a number of different, existing cohorts were used to train and validate the ABED, which is a strength of the study.

The potential of this is that the very rapid inference time (14 seconds per average PSG is cited) can increase efficiency and scope of sleep apnea testing.

A number of possible areas for clinical translation are given. I think the most attractive is in easily identifying endotypes of obstructive sleep apnea. Endotypic traits have been proposed for use in selecting patients that might benefit preferentially from a range of non-CPAP alternative therapies including oral appliances, hypoglossal nerve stimulation, supplemental oxygen, and pharmacotherapy. Given there is currently a one-size-fits-all management approach that predominates where CPAP is first line treatment for almost all patients. Accurately identifying patients likely to respond to alternatives to CPAP to avoid a trial and error approach to alternatives would be a major advance. Systems to identify endotypes from routine PSG signals with automated modelling have been developed and some are referenced here (21-25). However, the authors have not referenced further work that has demonstrated some of these systems

are highly accurate in predicting, for example, oral appliance treatment outcomes in obstructive sleep apnea (see Dutta et al, J Clin Sleep Med 2022), or that combine clinical data and models from PSG data (see Eckert et al, Sleep 2024;47 (Suppl 1): A257) to identify successful candidates for Oral Appliance Therapy (and vice versa).

We thank the reviewer for highlighting the clinical relevance of endotype-based treatment selection and for suggesting key references that strengthen the translational potential of our work. In our revised manuscript, we have incorporated the study by Dutta et al. (2022) [<https://jcsn.aasm.org/doi/full/10.5664/jcsn.9742>], which developed a physiological-based model using standard polysomnography and clinical data to predict oral appliance treatment outcomes in obstructive sleep apnea. This study reported a mean accuracy of $91\% \pm 8\%$ using 10-fold cross-validation and 100% accuracy in an independent validation set for predicting responders ($AHI < 5$ events/h) PMC. We have also referenced the work by Eckert et al. (2024) [https://academic.oup.com/sleep/article/47/Supplement_1/A257/7654650?login=true], which combines clinical data and PSG models to identify successful candidates for oral appliance therapy. These additions provide a more comprehensive overview of existing predictive models in the field.

The authors assert that the addition of “apnotyping” via probabilistic detection of event subtypes within each majority class could benefit sleep apnea endotyping. It remains an open question whether the ABED will be as effective as these methods in clinical practice as this should be acknowledged in the limitations. Further research is also needed to demonstrate if endotypes are useful in predicting which patients are at risk of adverse health consequences of OSA. This is beyond the scope of the current study.

We acknowledge that while the ABED demonstrates promising performance in detecting respiratory events and subtypes, its effectiveness in predicting treatment outcomes, such as response to oral appliance therapy, remains to be established. In the revised manuscript, we have added a statement in the limitations section to reflect this, noting that further research is needed to compare the ABED's predictive capabilities with existing models and to validate its clinical utility in diverse patient populations.

The authors also do not reference previous work on the significance of isolated respiratory events without arousals or desaturations (IRE). Previous work has shown non-apneic REs without cortical arousal or desaturation are associated with significant respiratory and heart rate changes. And, that optimal CPAP and the reduction of resistive load are associated with the normalization of heart rate indicating potential clinical benefit (for example see Sankari et al, Sleep 2016 ;40:zsw024). Easy identification of these events would be useful but is not a novel finding.

Thank you for highlighting this omission. We have now included a reference to the study by Sankari et al. (2016), which demonstrated that non-apneic respiratory events without cortical arousal or desaturation are associated with significant respiratory and heart rate changes, and that optimal CPAP therapy can normalize heart rate, indicating potential clinical benefit. This addition acknowledges the clinical significance of IREs and situates our work within the broader context of sleep-disordered breathing research.

There are a number of other limitations I feel need to be considered in addition to those given in the Discussion. As the authors point out in the Introduction, a number of novel metrics taken from the PSG have been identified that have shown to be better predictors of downstream consequences of OSA than the AHI. These include hypoxic burden and a number of metrics derived from heart rates. Further, the intensity of arousals from events are also likely to be clinically significant. It does not appear that the ABED identifies any of these metrics. Consequently, the value of the ABED in identifying which patients may benefit long-term from therapy is uncertain. Sleep studies also require scoring of sleep stages and identification of arousals due to limb movements or that occur spontaneously. I presume the ABED would need to be utilised in addition to automatic scoring software for these features. Overall, the paper presents a potentially useful addition to automate the scoring of sleep-disordered breathing events that will likely be of value in sleep medicine.

We appreciate the reviewer's insights into additional limitations. In the revised manuscript, we have expanded the limitations section to discuss the following points:

- Novel Metrics: Although ABED describes the overall apnotypes of subjects, we acknowledge that the ABED does not identify certain novel metrics derived from PSG, such as hypoxic burden and heart rate variability, which have been shown to be better predictors of downstream consequences of OSA than the AHI [<https://pmc.ncbi.nlm.nih.gov/articles/PMC8483223/>?, <https://academic.oup.com/eurheartj/article/40/14/1149/5146754>].
- Comprehensive Sleep Study Scoring: We recognize that sleep studies require scoring of sleep stages and identification of arousals due to limb movements or spontaneous occurrences. The ABED is designed to detect respiratory events and would need to be utilized in conjunction with automated scoring tools for full PSG interpretation.

These acknowledgments provide a more balanced perspective on the capabilities and current limitations of the ABED system.