

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

Magnus Ruud Kjaer^{1,2,3}, Umaer Hanif², Andreas Brink-Kjaer¹, Mads Olsen^{1,2,3}, Oliver Sum-Ping³, Oscar Carrillo³, Scott A. Sands⁴, Susan Redline⁴, Katie L. Stone^{5,6}, Poul Jennum^{2,7*}, Emmanuel Mignot^{3*}.

¹ Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark,

² Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, Glostrup, Denmark,

³ Stanford Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA

⁴ Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

⁵ San Francisco Coordinating Center, California Pacific Medical Center Research Institute, San Francisco, CA

⁶ Department of Epidemiology and Biostatistics, University of California, San Francisco

⁷ Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

* These authors contributed equally.

Corresponding authors:

Magnus Ruud Kjaer: magnusrk@stanford.edu

Emmanuel Mignot: mignot@stanford.edu

Supplementary information

Appendix

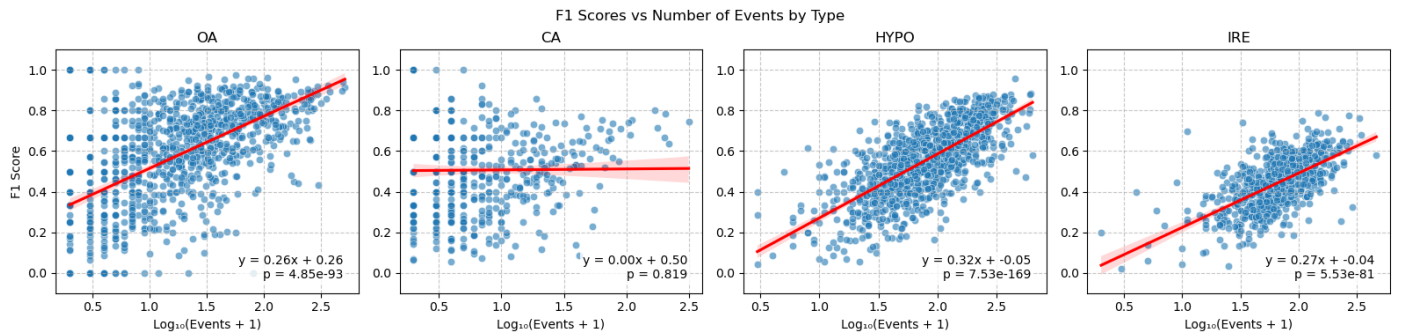
A0: Sleep disordered breathing event detection and sleep apnea diagnosis

To further explore ABED's ability to detect sleep-disordered breathing (SDB) events, we present average F1-score, precision, and recall across subjects in the test set from single-scored datasets (MESA, MrOS, WSC, CFS) in supplementary table 1. A similar trend is observed as seen in the per event metrics shown in Table 1: the detection for OA is strongest, followed by HYPO, CA, and finally, IRE. However, the overall performance metrics are generally lower than those in Table 1, motivating us to investigate how detection performance varies based on the number of annotated events in each PSG.

Supplementary Fig. 1 illustrates the relationship between F1-score and the number of events in the PSGs across different event types. For OA, HYPO, and IRE detection, a clear and statistically significant improvement in F1-scores is evident as the number of annotated events increases. Interestingly, the F1-score for CA detection appears unaffected by the number of annotated events in this category.

Supplementary table 1: The F1-score, precision, and recall for each event type (obstructive apnea (OA), central apnea (CA), hypopnea (HYPO), isolated respiratory event (IRE), and apnea event (OA or CA or HYPO)) averaged across subjects in the single scored test datasets (MESA, MrOS, WSC, CFS).

EVENT TYPE	METRIC		
	F1	Precision	Recall
OA	0.57 ± 0.27	0.53 ± 0.27	0.72 ± 0.29
CA	0.51 ± 0.22	0.30 ± 0.31	0.59 ± 0.35
HYPO	0.55 ± 0.17	0.57 ± 0.19	0.58 ± 0.21
IRE	0.44 ± 0.13	0.36 ± 0.17	0.59 ± 0.15
APNEA EVENT	0.69 ± 0.17	0.68 ± 0.20	0.74 ± 0.19



Supplementary figure 1: Scatterplots with regression lines show the relationship between the $\log_{10}$ -transformed number of events and F1-score for each event type. The number of events positively correlates with F1-score for OA, HYPO, and IRE, while CA detection does not demonstrate a significant relationship between the number of events and performance.

In supplementary table 2, we present ABED's diagnostic capabilities, specifically its correlation with the Apnea-Hypopnea Index (AHI) distributed across each cohort in the single-scored test set. The coefficient of determination (R^2) ranges from 0.826 to 0.871 across the four cohorts, with all showing similar Beta (β) values except for the CFS cohort, which exhibits a higher bias and a correspondingly lower Beta value. Regarding diagnostic accuracy and severity grouping accuracy, ABED performs best on the MESA and MrOS datasets but shows comparatively lower performance on the CFS and WSC cohorts.

Supplementary table 2: β is the slope of the regression line, and Intercept denotes the bias in AHI (estimated prediction at true AHI = 0). R^2 indicates the coefficient of determination.

COHORT	β	INTERCEPT	R^2	DIAGNOSIS ACCURACY	SEVERITY ACCURACY
MESA	0.94	1.58	85.7	90.4	79.1
MROS	0.88	2.08	87.1	89.0	80.2
CFS	1.07	-7.56	85.2	83.0	66.0
WSC	0.91	-1.54	82.6	83.6	65.6
ALL	0.90	0.84	84.4	87.5	75.3

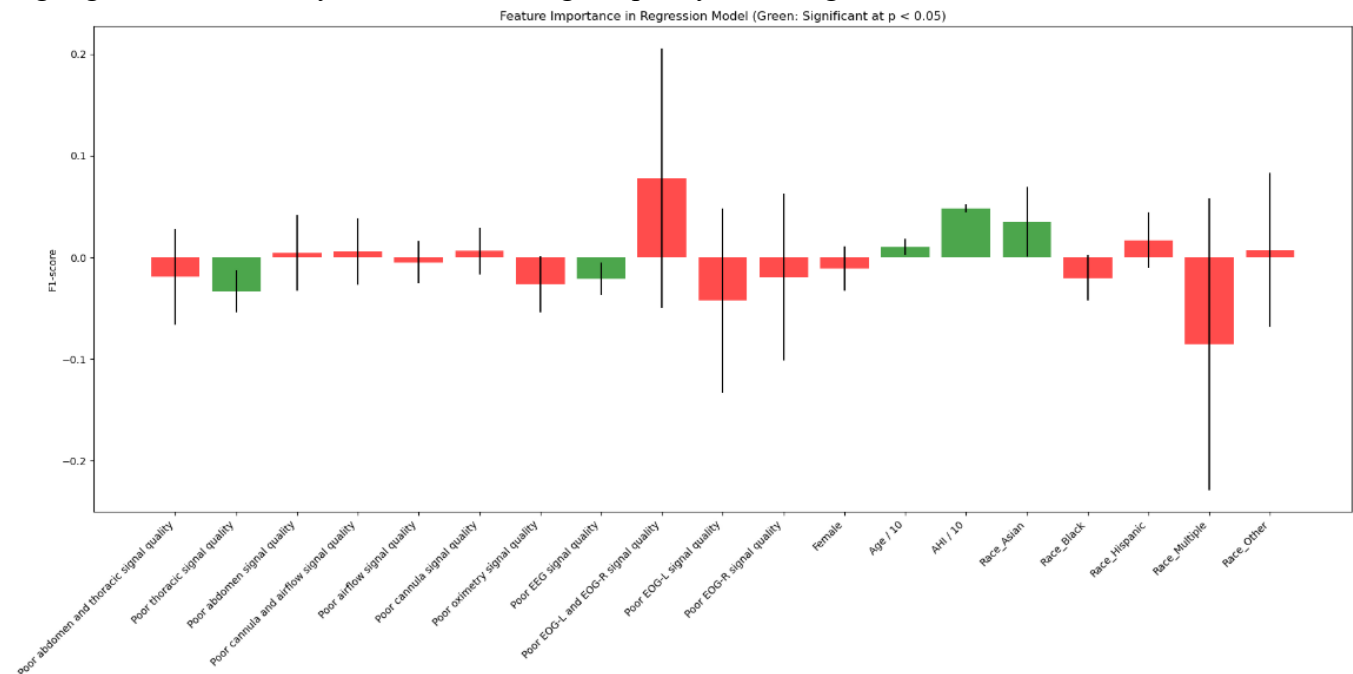
In supplementary table 3, we compare ABED's performance metrics to other state-of-the-art apnea detection models. We observe that other detectors often either show highly variable performance or lack transparency across all metrics and tasks. ABED generally ranks as the top performer or a close second, with comprehensive reporting across all relevant metrics.

*Supplementary table 3: Performance metrics describing the model's ability to diagnose sleep apnea (SA) by presence and severity based on the apnea-hypopnea index (AHI) and to detect events. PD: presence diagnosis accuracy, SA not present: $AHI < 15$ and SA is present: $AHI > 15$. SD: severity diagnosis accuracy, in groups: none: $AHI < 5$, mild: $5 < AHI < 15$, moderate $15 < AHI < 30$, severe $AHI > 30$. * Denotes event vs no event detection. Bold indicates the best performing model for the given metric.*

STUDY	AHI R ²	PD	SD	EVENT	MEAN (OA, CA, HYPO)	EPOCH
HAIDAR ET. AL. ¹	-	-	-	-	-	F1 = 0.83
STENKESTIE ET. AL. ²	-	78.8	56.8	-	-	F1 = 0.48
LAKHAN ET. AL. ³	-	79.6	63.7	-	-	-
YEO ET. AL. ⁴	-	83.6	-	-	-	F1 = 0.63*
BISWAL ET. AL. ⁵	82.8	-	87.0	-	-	-
OLESEN ET. AL. ⁶	77.0	-	-	F1 = 0.70*	-	-
NASSI ET. AL. ⁷	88.0	84.3	66.3	-	F1 = 0.43	-
ABED	84.4	87.5	75.3	F1 = 0.78*	F1 = 0.62	-

A1: signal quality

We examined how signal quality impacts the F1-score in detecting apnea events, using metrics from the National Sleep Research Resource (NSRR) on the MESA, MrOS, and CFS test datasets. Signal quality was classified as "Poor" based on the NSRR's lowest category. Our analysis found that thoracic and EEG signals marked as Poor significantly decreased the F1-score. In contrast, factors such as older age, a higher Apnea-Hypopnea Index (AHI), and identifying as Asian were associated with higher F1-scores. All significant factors had p-values below 0.05, indicating the model's robustness to variations in signal quality. These findings highlight ABED's ability to differentiate signal quality, allowing accurate event annotation.



Supplementary figure 2: Feature importance analysis for the regression model assessing the F1-score on MESA, MrOS, and CFS test datasets. Bars in green indicate features with significant impact on the F1-score, while red bars denote those without significant effect. Only thoracic and EEG signals classified as Poor significantly worsened the F1-score. Positive influences included factors like age, AHI, and Asian race. Baseline categories for the model include Caucasian for race, male for gender, and good signal quality for signal categories

Supplementary table 4: Stepwise linear regression was used to predict loop gain (LGN), pharyngeal muscle compensation (VCOMP), and arousal threshold (ART). Variables from five categories were used: demographic variables (BMI, age, sex), true apnea indexes, predicted apnea indexes, mean apnea probabilities multiplied by predicted apnea indexes, and the combination of mean probabilities and predicted apnea indexes. The base demographic variables were consistently forced into the models. The table displays selected variables and their corresponding multivariate effects, alongside the adjusted R^2 (R^2 adj) for each model. Bold indicates the best performing model.

VARIABLE		EFFECT	R2(ADJ)		
MODEL (LGN)					
BASE	BMI		0.088		
	Age				
	Sex				
BASE + TRUE	OAI _{true}		0.129		
BASE + PREDICTED	HI _{pred}		0.209		
	CAI _{pred}				
	NI _{pred}				
BASE + PROBABILITY X PREDICTED	OAI · $\overline{P(CA OA)}$		0.226		
	OAI · $\overline{P(HYPO OA)}$				
	NI · $\overline{P(IRE IRE)}$				
	CAI · $\overline{P(CA CA)}$				
BASE + PROBABILITY + PREDICTED	$\overline{P(CA IRE)}$		0.266		
	CAI _{pred}				
	$\overline{P(IRE HYPO)}$				
	NI _{pred}				
MODEL (VCOMP)					
BASE	BMI		-0.002		
	Age				
	Sex				
BASE + TRUE	-		-0.002		
BASE + PREDICTED	OAI _{pred}		0.069		
	NI _{pred}				
BASE + PROBABILITY X PREDICTED	NI · $\overline{P(IRE IRE)}$		0.083		
	OAI · $\overline{P(HYPO OA)}$				
BASE + PROBABILITY + PREDICTED	$\overline{P(HYPO OA)}$		0.078		
	NI _{pred}				
MODEL (ART)					
BASE	BMI		0.018		
	Age				
	Sex				
BASE + TRUE	HI _{true}		0.289		
	OAI _{true}				
BASE + PREDICTED	HI _{pred}		0.201		
	OAI _{pred}				
BASE + PROBABILITY X PREDICTED	HI · $\overline{P(IRE HYPO)}$		0.270		
	NI · $\overline{P(OA IRE)}$				
	HI · $\overline{P(HYPO HYPO)}$				
	OAI · $\overline{P(HYPO OA)}$				
BASE + PROBABILITY + PREDICTED	$\overline{P(IRE HYPO)}$		0.270		
	$\overline{P(OA IRE)}$				
	$\overline{P(HYPO HYPO)}$				
	$\overline{P(HYPO OA)}$				

A2: loop gain analysis on MESA cohort

In our analysis of the MESA cohort, we conducted loop gain assessments using stepwise linear regression. The base model consistently included BMI, age, and sex as variables, although their effects are not visualized across all models. Supplementary table 4 summarizes the multivariate models generated from five categories of variables: demographics, true apnea indexes, predicted apnea indexes, mean apnea probabilities multiplied by predicted apnea indexes, and mean apnea probabilities and predicted apnea indexes combined. The base demographic variables were forced into the stepwise regression process.

In supplementary table 4, we present the selected variables from each category, showing their multivariate effects and the adjusted R^2 (R^2 adj) achieved by the final models. The results indicate that the best-performing models are those incorporating the probabilities provided by ABED's scoring for both loop gain and pharyngeal muscle compensation. However, for predicting arousal threshold, the true annotations slightly outperformed the probability-based models.

A3: Extended methods

Pre-processing

An issue in the field of SDB detection is variability in filter setting which can distort the shape of breathing signals. Preprocessing was done on all PSGs to ensure conformity of input to ABED. For NF, OT, AbP and ThP (typical breathing channels), a 4th order digital Butterworth band-pass filter with a cut-off frequency of [0.1 – 4] Hz was applied before down sampling to 8 Hz. Modalities such as the blood oxygen saturation, wake and arousal probabilities (EEG, EOG, EMG through MAD, see Fig. 1a) were up sampled to 8 Hz. Frequency content of breathing channels in a PSG is thought not to surpass 4 Hz during sleep, therefore 8 Hz is optimal for lowest computational cost^{8,9}. Standardization was applied per PSG to all modalities apart from wake and arousal probabilities, which are already scaled from 0 to 1. The modalities are then stacked on top of each other such that each timestep has one datapoint per modality and this is eventually written into a memory mapped file.

Detection pipeline

ABED uses 240 second windows as input which overlap 50% with the following window until the end of the PSG recording. This is done so that as much context as possible can be used to create predictions. The ABED detection pipeline, for a window, consists of object detection, as shown in Fig. 1a, where 146 default events are made for each window. A default event is a bounding box placed in the 240 second window that gets an updated location and event type probabilities (OA, CA, HYPO, IRE and none-event probabilities scaled via a SoftMax function) after a forward pass in the convolutional neural (CNN) network.

During inference, non-maximum suppression, as used in¹⁰, was used to remove events that are not the highest probability in its location. Specifically, across different overlapping windows any event predicted having a high probability for multiple event types were classified following a simple set of 2 rules: (1) apneas (OA, CA) classification took precedence over hypopneas (HYPO) classification, (2) CA took precedence over OA. These rules are justified, as OA and CA are more severe versions of HYPO, and it is therefore likely that the HYPO probability is high for apnea events. Further, CAs are rarer and share similar traits compared to OA therefore actual CA events also have a high OA probability.

ABED's architecture is summarized in Fig. 1b. It contains three main groups of layers. The left grouping of layers contains the input (grey) data that consists of a 7x1920 matrix. This input is mixed in the channel mixing layer (orange) from its matrix shape into 8 channels using a 7x1 kernel, thus allowing ABED to mix all modalities into channels as necessary. Hereafter, a channel amplification layer (red) is used to reduce the time content of the features and increase the number of channels to fit into a flattened ResNet50¹¹ to extract features. In the middle grouping of layers, the bottleneck block is shown with three convolutional layers each with batch normalization¹² and ReLU¹³ between. The flattened ResNet50¹¹ feature extraction layers receive 64 feature channels each with 960 timesteps. The number of channels is doubled and a stride of 2 is applied 3 times along with an initial amplification of channels with a factor of 4 resulting in 2048 output feature channels each with 60 timesteps. These are then fed into a bi-directional LSTM layer which then finds information across the time axis in both directions for each channel. Hereafter, a dropout of 90% is applied¹⁴. Finally, in the right most grouping of layers, a classification layer and localization layer outputs the position and class probability for each of 146 bounding boxes. In the localization layer a kernel of size 1x60 fully connects each timestep per feature channel into the start and duration of 146 bounding boxes. Similarly, the four class probabilities for OA, CA, HYPO, IRE, and none are found by the classification layer with a kernel of 1x60 also fully connecting the time dimension into each of the 146 bounding boxes.

Training

ABED was trained using 5,456 PSGs belonging to the 4 cohorts mentioned above. A subset of the training data 546 PSGs was set aside for validation and 1,099 PSGs (MESA: 249, MrOS: 500, WSC: 250, CFS: 100) was set aside for final testing. Early stopping was used to attain the best possible validation-loss at any epoch. Classes that were less prevalent (OA, IRE and CA) were oversampled prior to training to resolve class imbalances (Table 3 illustrates why imbalance was an issue) allowing for a less biased training of highly prevalent events such as HYPO. Both windows including and excluding events were added during training of the model (40% not containing events, and 60% containing at least one event). Windows that included events were comprised of non-event segments and it was not necessary to increase the amount of un-labelled data as

the model would have much exposure of non-event data. Windows completely without events were chosen mainly to include wake stages as this was thought to make the model more robust to understanding that apneas do not occur during wake. Windows of 240 seconds in batches of 128 were used to train at each iteration. Here, the 240 second segments which contained events were placed randomly around the event but had to include the entire annotation. It was possible for other events to be present in the same window and by chance it was more likely that the HYPO event class was present. Therefore, the HYPO class was thought to be slightly overrepresented during training. ABED was implemented using the Python module PyTorch 1.7.1 and was trained using one NVIDIA V100 GPU. Hyperparameters were tuned slightly to ensure minimal overfitting towards the validation and test data.

Classification loss

Classification loss was calculated for two populations of default events in each 240 second window used for training of ABED. Default events containing true events (positive default events) and default events without annotated events (negative default events). Classification loss was calculated for each window, which contains a total of 146 default events with lengths varying from 10 seconds to 120 seconds incrementing by 10 seconds. The default events overlap 50% within length groups and completely between length groups. For both positive and negative default events, the classification loss was calculated for each individual default event using the cross-entropy loss function as written in Eq. 1.

$$C(p, q) = - \sum_{x \in X} p(x) \log q(x) \quad (1)$$

Here, $p(x)$ is the true probability (as a 1-hot vector where one class has a probability of 100% and the remaining have a probability of 0%) and $q(x)$ is ABED's predicted probability for each of the classes. x is each event class OA, CA, HYPO, IRE, or none. A softmax activation function has scaled the sum of ABED's predicted event probabilities to 1. $p(x)$ is 1 for the true class of either OA, CA, HYPO, or none, while it is 0 for the remaining classes. Hereby the classification loss is only perfect for the default event if it is predicted 100% accurately. The classification loss was calculated for each default event in the window and the summation of the events classification error defined the classifications loss for that window and can be written as written in Eq. 2.

$$clf_{loss}(\mathbf{P}, \mathbf{Q}) = \frac{1}{N_p} \sum_{p=1}^{N_p} C(p_p^+, q_p^+) + \sum_{n=1}^{N_n} C(p_n^-, q_n^-) \quad (2)$$

Here, $\mathbf{P}$ ($\mathbf{P} = [p_1, \dots, p_{146}]$, $\mathbf{P}^- = [p_1^-, \dots, p_{N_n}^-] \in \mathbf{P}$, $\mathbf{P}^+ = [p_1^+, \dots, p_{N_p}^+] \in \mathbf{P}$) are the true event probabilities for each default event and $\mathbf{Q}$ ($\mathbf{Q} = [q_1, \dots, q_{146}]$, $\mathbf{Q}^- = [q_1^-, \dots, q_{N_n}^-] \in \mathbf{Q}$, $\mathbf{Q}^+ = [q_1^+, \dots, q_{N_p}^+] \in \mathbf{Q}$) contains the predicted event probabilities for each default event, while superscript + are positive default events and superscript negative default events. N_p is the amount of positive default events in the window and N_n ($N_n = \min(10 - N_p, N_p \cdot 3)$) is the amount of worst negative default events in the window.

Localization loss

The localization loss was calculated for the positive default events using the Huber loss¹⁵. The Huber loss is a variation on the mean squared error loss (MSE), that is designed to weigh outliers less negatively than the MSE and its formulation can be seen in Eq. 3, where x_i represents predicted value, and y_i represents the true values. The absolute difference in x_i and y_i values are known as the residuals. In Eq. 4 the localization loss is defined for an entire 240 second window. Here, N_p defines the number of positive classifications, with predicted start and duration x_i and annotated start and duration y_i .

$$H_i = \begin{cases} \frac{1}{2} |x_i - y_i|^2, & \text{if } |x_i - y_i| \leq \delta \\ \delta \left(|x_i - y_i| - \frac{1}{2} \delta \right), & \text{otherwise} \end{cases} \quad (3)$$

$$loc_{loss}(\mathbf{x}, \mathbf{y}) = \frac{1}{N_p} \sum_{i=1}^{N_p} H_i(x_i, y_i) \quad (4)$$

During training of ABED, δ was set to 1, in this version of the Huber loss function the penalty of an event with residuals < 1 is scaled by half but retains it convex shape, while it is linear with residuals > 1 , still with higher penalties for higher residuals, but less so than compared to MSE. Residuals are presented in seconds during training.

Combined loss

Finally, the classification loss clf_{loss} and localization loss loc_{loss} is combined into the complete loss of a 240 second window as seen in Eq. 5.

$$loss = w \cdot clf_{loss} + loc_{loss} \quad (5)$$

Here, w is the weight applied to the classification loss for the model to change focus from localization and penalize wrong classification more. This is important as the Huber loss and cross entropy loss are not scaled to the importance in the object detection class. During experiments it was observed that the classification loss would be optimal later than the localization loss (i.e., the classification loss would be optimized one epoch later than the localization loss). This led to a combined loss, which was optimal at a time where the classification loss was not optimal. This did very little to harm the localization loss while improving the classification loss. The combination of these losses allow ABED to be optimized on localization and classification criteria simultaneously.

Network Optimization

Optimization was done on both localization and classification of each default event overlapping any annotated event and using the classification error of the worst performing default events that did not overlap with any annotated event. The Adaptive Moment Estimation (Adam) ¹⁶ algorithm was used as the optimizer during training with a learning rate of 10^{-3} , β_1 and β_2 equal 0.9 and 0.99, respectively.

Detection threshold

Thresholds were applied to the classification likelihoods of each default event to decide the presence of an event versus the none. We applied thresholds in the range of [0.51, 0.61, 0.71, 0.81, 0.91] and saved all outputs for the validation dataset. The thresholds' lowest point was chosen to be 51% certainty to ensure that a single default event would not predict multiple event types. The thresholds were then optimized based on the F1 for each individual class (e.g., OA vs CA vs HYPO vs IRE vs none) performance and used to run inference.

Evaluation

During evaluation of ABED, windows of 240 seconds were passed through ABED to find events. Each 240 second window was overlapping 50% with the previous window to get the context of each event object which is crucial for detection. This caused a possibility for events to be detected twice and to overlap. To ensure that only one detection was registered, only the highest classification probability was considered (e.g., the detected object that ABED was the most certain of). Eventually any detected events with a duration shorter than 8 seconds were removed as ≥ 10 sec duration is part of the definition of breathing events ¹⁷. On average, ABED annotated an entire PSG within 14 seconds, utilizing just one Tesla V100_SXM2 GPU and five cores of an Intel E5-2640v4 CPU. Evaluation of test data was done in such a way that any overlap in annotation and prediction resulted in a true prediction (i.e., IOU threshold = 0.0001). ABED was evaluated per event, per PSG, and with regards to AHI prediction and diagnosis, which was divided into several diagnosis classes:

(1) sleep apnea presence diagnosis:

- SA not present $AHI \leq 15$
- SA present $AHI > 15$

and

(2) sleep apnea severity diagnosis:

- No SA: $AHI < 5$
- Mild SA: $5 \leq AHI < 15$
- Moderate SA: $15 \leq AHI < 30$
- Severe SA: $30 \leq AHI$

The performance metrics chosen to evaluate ABED were precision, recall, F1, IOU, accuracy (also known as micro averaging), and coefficient of determination (R2) calculated as per equations 6 -11.

$$Precision = \frac{TP(x)}{TP(x) + FP(x)} \quad (6)$$

$$Recall = \frac{TP(x)}{TP(x) + FN(x)} \quad (7)$$

$$F_1(x) = \frac{Precision(x) \cdot Recall(x)}{Precision(x) + Recall(x)} \quad (8)$$

$$IOU(\mathbf{B}_{\text{pred}}, \mathbf{B}_{\text{true}}) = \frac{\mathbf{B}_{\text{pred}} \cap \mathbf{B}_{\text{true}}}{\mathbf{B}_{\text{pred}} \cup \mathbf{B}_{\text{true}}} \quad (9)$$

$$Accuracy = \frac{\sum_{x \in \mathbf{X}} TP(x)}{\sum_{x \in \mathbf{X}} TP(x) + FP(x)} \quad (10)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (11)$$

$$R_{adj}^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{(n - p - 1)} \right) \quad (12)$$

Here, TP is a true positive prediction (i.e., correctly classified), FP is a false positive prediction (i.e., wrongly classified), FN is a false negative prediction (i.e., wrongly missed). $\mathbf{B}_{\text{true}}$ is the true bounding box, $\mathbf{B}_{\text{pred}}$ is the predicted bounding box. x is an event class in $\mathbf{X}$, which is all event types (OA, CA, HYPO, IRE). y_i is the correct value, $\bar{y}$ is the mean of all correct values, f_i is the predicted value, n is the number of true and predicted values, and p is the number of predictors.

References

1. Haidar, R., McCloskey, S., Koprinska, I. & Jeffries, B. Convolutional Neural Networks on Multiple Respiratory Channels to Detect Hypopnea and Obstructive Apnea Events. *Proceedings of the International Joint Conference on Neural Networks* 2018-July, (2018).
2. Van Steenkiste, T., Groenendaal, W., Deschrijver, Di. & Dhaene, T. Automated Sleep Apnea Detection in Raw Respiratory Signals Using Long Short-Term Memory Neural Networks. *IEEE J Biomed Health Inform* 23, 2354–2364 (2019).
3. Lakhan, P., Ditthapron, A., Banluesombatkul, N. & Wilaiprasitporn, T. Deep Neural Networks with Weighted Averaged Overnight Airflow Features for Sleep Apnea-Hypopnea Severity Classification. *IEEE Region 10 Annual International Conference, Proceedings/TENCON* 2018-October, 441–445 (2018).
4. Yeo, M. *et al.* Respiratory Event Detection during Sleep Using Electrocardiogram and Respiratory Related Signals: Using Polysomnogram and Patch-Type Wearable Device Data. *IEEE J Biomed Health Inform* 26, 550–560 (2022).
5. Biswal, S. *et al.* Expert-level sleep scoring with deep neural networks. *Journal of the American Medical Informatics Association* 25, 1643–1650 (2018).
6. Zahid, A. N., Jennum, P., Mignot, E. & Sorensen, H. B. D. MSed: A Multi-Modal Sleep Event Detection Model for Clinical Sleep Analysis. *IEEE Trans Biomed Eng* 70, 2508–2518 (2023).
7. Nassi, T. E. *et al.* Automated Scoring of Respiratory Events in Sleep with a Single Effort Belt and Deep Neural Networks. *IEEE Trans Biomed Eng* 69, 2094 (2022).
8. Aittokallio, T., Nevalainen, O., Pursiheimo, U., Saaresranta, T. & Polo, O. Classification of Nasal Inspiratory Flow Shapes by Attributed Finite Automata. *Computers and Biomedical Research* 32, 34–55 (1999).
9. Aittokallio, T., Malminen, J. S., Pahikkala, T., Polo, O. & Nevalainen, O. S. Inspiratory flow shape clustering: An automated method to monitor upper airway performance during sleep. *Comput Methods Programs Biomed* 85, 8–18 (2007).
10. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. in 779–788 (2016).
11. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in 770–778 (2016).
12. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. in 448–456 (PMLR, 2015).
13. Nair, V. & Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. in.
14. Srivastava, N., Hinton, G., Krizhevsky, A. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 1929–1958 (2014).
15. Huber, P. J. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35, 73–101 (1964).
16. Kingma, D. P. & Ba, J. L. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2014).
17. Berry, R. B. *et al.* AASM | Scoring Manual Version 2.2 The AASM Manual for the Scoring of Sleep and Associated Events RULES, TERMINOLOGY AND TECHNICAL SPECIFICATIONS VERSION 2.2. (2015).