Supplementary

Contents

Datasets

Supplementary Note 1. Korea Image-based Sleep Study (KISS) Dataset

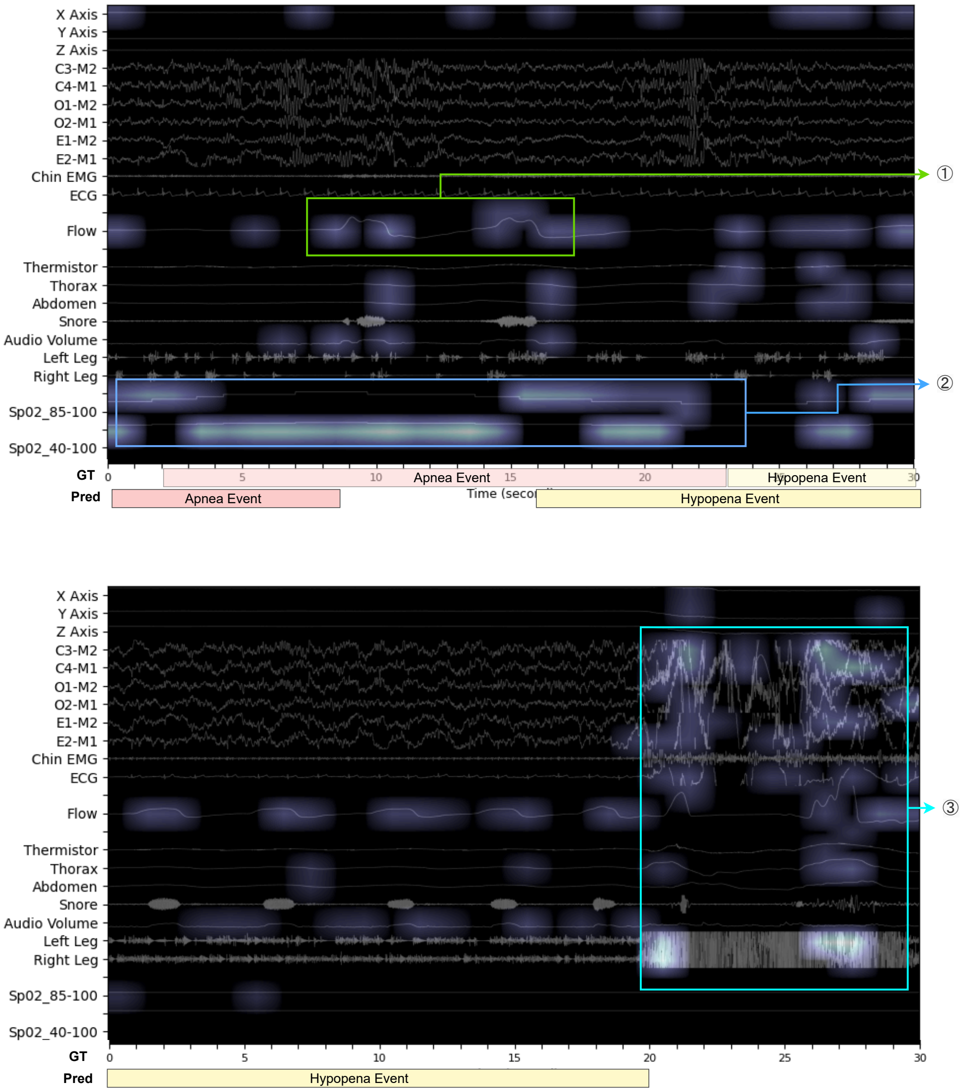
KISS dataset is a standardized PSG image-based dataset constructed by[1]. The dataset was collected from four different sleep centers using two distinct PSG devices: Embla (Natus Medical, San Carlos, CA, USA) and NOX (Nox Medical, Reykjavik, Iceland). It comprises a total of 10,253 PSG records collected between 2013 and 2020. The scoring was performed following the American Academy of Sleep Medicine (AASM) guidelines[2–4], version 2.6. Each record includes 21 different biosignals: Acceleration (x, y, and z-axis), EEG (C3M2, C4M1, O1M2, O2M1), EOG (E1M2, E2M1), chin EMG, ECG, airflow, oronasal thermistor, thoracic movement, abdominal movement, snoring, audio volume, leg EMG (left and right leg), and oxygen saturation (85100%, 40100%). Before converting the raw PSG data into the standardized images, the data underwent preprocessing, including the application of a fourth-order Butterworth low-pass filter, a high-pass filter, and a band-pass filter. The cutoff frequencies for these filters were selected according to the AASM scoring manual. For this study, two subsets each from a single center and a single PSG device were used. The full dataset of KISS does not publicly available because of legal restrictions imposed by Korean government in relation to the Person Information Protection Act. However, if some investigators wish to use it, they could access it after obtaining the relevant permit from the Korean National Information Society Agency in AI Hub[5].

Supplementary Note 2. Sleep Heart Health Study Visit 2 (SHHS-2) Dataset

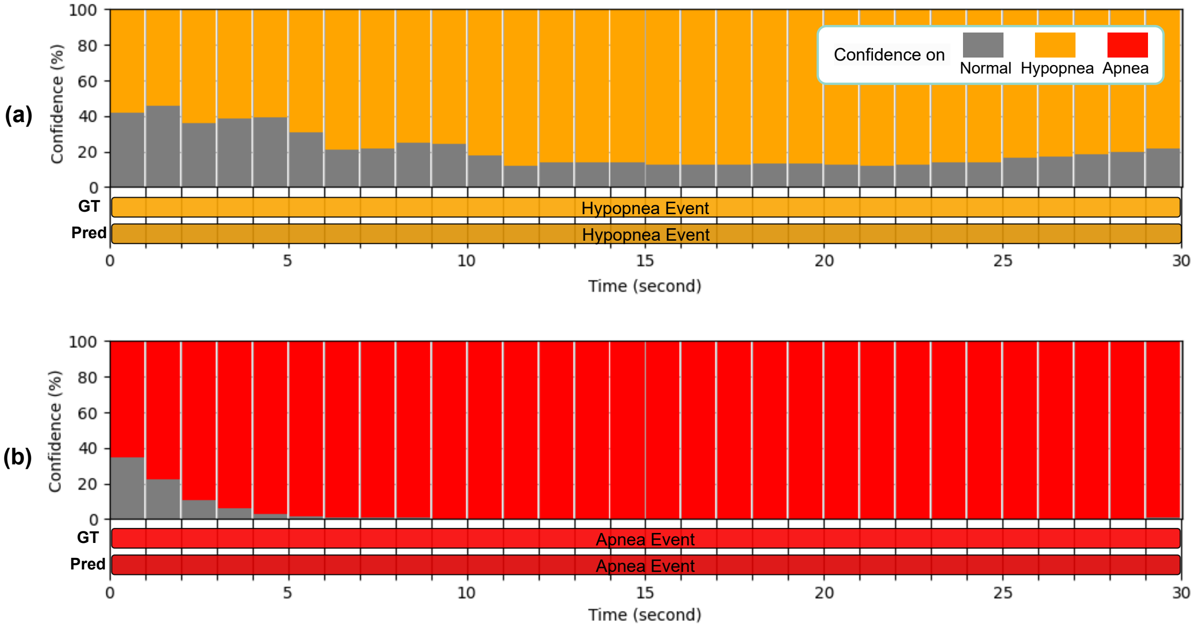
The Sleep Heart Health Study (SHHS) is a multicenter cohort study conducted by the National Heart, Lung, and Blood Institute[6]. The study was conducted over two visits, and the SHHS-2 dataset was constructed from the second visit from 2001 to 2003. The SHHS-2 dataset collected a total of 3,295 recordings, with 2,651 currently available. This dataset includes biosignal recordings of EEG (C3/A2 and C4/A1), EOG (right and left), EMG, thoracic and abdominal excursions, nasal-oral airflow, oxygen saturation, ECG, and body position. For this study, the raw signals were converted into standardized PSG images. Because there are fewer signal channels in SHHS-2 compared to the KISS dataset, we duplicated similar signals as suggested in[1], maintaining consistent standardized dataset form.

Results

Model Explainability



Supplementary Figure 1. Attention score visualization for wrong predictions. In the upper example, *VOSA* misclassified both the apnea and hypopnea event timings according to the ground-truth labels. However, the attention scores indicate that *VOSA* accurately identified normal breathing (②) and SpO$_2$ change (③). This suggests a possible labeling error in the ground truth, and that *VOSA* may have made a more clinically accurate prediction. In the lower example, noise appears to have been introduced in the data (③). Although no event was annotated, *VOSA* predicted a hypopnea event and focused attention on the noisy region. This suggests that the model may have interpreted the noise as an arousal-related pattern.



Supplementary Figure 2. Confidence score visualization for additional cases. Consecutive confidence scores for each class is visualized as different colors. Each box corresponds to the confidence score for a single second, with red, orange, and gray indicating apnea, hypopnea, and normal classes, respectively. Samples (a) and (b) are epochs entirely composed of hypopnea and apnea events, respectively. The overall confidence score for the hypopnea epoch is lower compared to the apnea epoch.

Automatic PSG Report Generation

Supplementary Note 3. Automatic PSG Report Generation

To present detailed results of automatic PSG report generation, we sequentially provide additional examples of VOSA-generated reports across diverse OSA severity levels, scatter plots comparing each estimated statistic, and event-level evaluation of the generated report.

The event-level report includes event number, event label, start and end time, corresponding epochs, and event duration. A representative case from the KISS is shown in Supplementary Table 1, comparing the AI-generated report (using *VOSA* and SleepXViT) with the manually annotated report. *VOSA* estimated an AHI of 13.4, closely matching the reference value of 14.0, while SleepXViT achieved a macro F1-score of 90.7% in sleep staging. Both models produced temporally consistent and accurate annotations across diverse event types, including N1 and N2 sleep stages, hypopneas, and obstructive apneas. Sleep stages exhibited exact alignment with expert annotations, while the apnea event showed only a minimal deviation by 1 second in duration (22.0 vs. 23.0 seconds) and 1-2 seconds in onset and offset timing. Even for the more ambiguous hypopnea class, the model correctly localized the event to the appropriate epochs, with a modest overestimation in duration (44.0 vs. 31.0 seconds). Discrepancies in event numbering stem from the exclusion of non-respiratory events (e.g., arousals, body position changes) in the generated report. However, the sequence and content of the sleep stages and respiratory events remained coherent and accurate.

VOSA-SleepXViT Generated Polysomnography Report

Phone: (XXX) XXX-XXXX

Patient number	A2019-NX-01-1605	**Age**	18	**Acquisition**	XXXXX
Started	08/12/19 at 10:13:00 PM	**Sex**	Female	**Type**	Adult
Stopped	08/13/19 at 05:50:49 AM	**BMI**	17.2	**Duration**	457.8

Sleep Data

Lights Out:	10:13:00 PM	Sleep Onset:	10:40:30 PM
Lights On:	XX:XX:XX AM	Sleep Efficiency:	81.0 %
Total Recording Time:	457.8 min	Sleep Latency (from Lights Off):	27.5 min
Total Sleep Time (TST):	350.5 min	REM Latency (from Sleep Onset):	145.0 min
Total Wake Time (TWT):	87.0 min	Wake After Sleep Onset (WASO):	59.5 min

Sleep Stage

Stage	**Latency**	**Latency**	**Duration**	**/ TST**	**/ TIB**
	(out)	**(onset)**	**(min)**	**(%)**	**(%)**
N1	27.5	0.0	31.0	8.9	6.8
N2	32.0	4.5	141.0	40.2	30.8
N3	37.0	9.5	121.5	34.7	26.5
REM	172.5	145.0	77.5	22.1	16.9

AHI

	REM	**NREM**	**TST**
AHI	1.5	0.4	0.6

Respiratory Data

	Apnea	**Hypopnea**	**A+H**
Number	0	4	4
Mean Dur (sec)	0.0	22.8	22.8
Max Dur (sec)	0.0	27.0	27.0
Total Dur (min)	0.0	1.5	1.5
% of TST	0.0	0.4	0.3
Index (#/TST)	0.0	0.6	0.6
Index (REM)	0.0	1.5	1.5
Index (NREM)	0.0	0.4	0.4

Manually Annotated Polysomnography Report

Phone: (XXX) XXX-XXXX

Patient number	A2019-NX-01-1605	**Age**	18	**Acquisition**	XXXXX
Started	08/12/19 at 10:13:00 PM	**Sex**	Female	**Type**	Adult
Stopped	08/13/19 at 05:50:49 AM	**BMI**	17.2	**Duration**	457.8

Sleep Data

Lights Out:	10:13:00 PM	Sleep Onset:	10:41:00 PM
Lights On:	XX:XX:XX AM	Sleep Efficiency:	80.4 %
Total Recording Time:	458.0 min	Sleep Latency (from Lights Off):	28.0 min
Total Sleep Time (TST):	368.5 min	REM Latency (from Sleep Onset):	145.5 min
Total Wake Time:	89.8 min	Wake After Sleep Onset (WASO):	61.8 min

Sleep Stage

Stage	**Latency**	**Latency**	**Duration**	**/ TST**	**/ TIB**
	(out)	**(onset)**	**(min)**	**(%)**	**(%)**
N1	27.5	0.0	32.0	8.7	7.0
N2	32.0	4.5	149.0	40.4	32.5
N3	37.0	9.5	125.5	34.1	27.4
REM	173.0	145.5	62.0	16.8	13.5

AHI

	REM	**NREM**	**TST**
AHI	1.0	0.6	0.7

Respiratory Data

	Apnea	**Hypopnea**	**A+H**
Number	0	4	4
Mean Dur (sec)	0.0	17.1	17.1
Max Dur (sec)	0.0	19.2	19.2
Total Dur (min)	0.0	1.1	1.1
% of TST	0.0	0.3	0.3
Index (#/TST)	0.0	0.7	0.7
Index (REM)	0.0	1.0	1.0
Index (NREM)	0.0	0.6	0.6

Note. Latency (out/onset): from lights out/sleep onset.
A+H: Apnea+Hypopnea

Supplementary Figure 3. Automaticly generated PSG Summary Report (without OSA). AI-generated results (top) were obtained by integrating *VOSA* for respiratory event detection and SleepXViT for sleep staging. Statistics exclusively computed from *VOSA* are highlighted in **green**, while metrics that rely on combined outputs from both models are highlighted in **blue**. Manually annotated results are shown below for comparison. Unavailable data (e.g., anonymized information) is represented as XX.

VOSA-SleepXViT Generated Polysomnography Report

Phone: (XXX) XXX-XXXX

Patient number	A2018-NX-01-0289	**Age**	28	**Acquisition**	XXXXX
Started	10/19/18 at 10:33:31 PM	**Sex**	Male	**Type**	Adult
Stopped	10/20/18 at 05:11:00 AM	**BMI**	26.4	**Duration**	397.5

Sleep Data

Lights Out:	10:34:00 PM	Sleep Onset:	10:35:30 PM
Lights On:	XX:XX:XX AM	Sleep Efficiency:	94.6 %
Total Recording Time:	397.5min	Sleep Latency (from Lights Off):	1.5 min
Total Sleep Time (TST):	375.5 min	REM Latency (from Sleep Onset):	47.0 min
Total Wake Time (TWT):	21.5 min	Wake After Sleep Onset (WASO):	20.0 min

Sleep Stage

Stage	**Latency (out)**	**Latency (onset)**	**Duration (min)**	**/ TST (%)**	**/ TIB (%)**
N1	1.5	0.0	27.5	7.3	6.9
N2	6.0	4.5	155.5	41.4	39.2
N3	17.5	16.0	57.5	15.3	14.5
REM	48.5	47.0	135.0	36.0	34.0

AHI

	REM	**NREM**	**TST**
AHI	6.2	14.7	11.8

Respiratory Data

	Apnea	**Hypopnea**	**A+H**
Number	0	74	74
Mean Dur (sec)	0.0	21.9	21.9
Max Dur (sec)	0.0	44.0	44.0
Total Dur (min)	0.0	27.0	27.0
% of TST	0.0	7.2	7.2
Index (#/TST)	0.0	11.8	11.8
Index (REM)	0.0	6.2	6.2
Index (NREM)	0.0	14.7	14.7

Manually Annotated Polysomnography Report

Phone: (XXX) XXX-XXXX

Patient number	A2018-NX-01-0289	**Age**	28	**Acquisition**	XXXXX
Started	10/19/18 at 10:33:31 PM	**Sex**	Male	**Type**	Adult
Stopped	10/20/18 at 05:11:00 AM	**BMI**	26.4	**Duration**	397.5

Sleep Data

Lights Out:	10:34:00 PM	Sleep Onset:	10:35:00 PM
Lights On:	XX:XX:XX AM	Sleep Efficiency:	94.7 %
Total Recording Time:	397.5min	Sleep Latency (from Lights Off):	1.0 min
Total Sleep Time (TST):	376.0 min	REM Latency (from Sleep Onset):	49.5 min
Total Wake Time (TWT):	21.0 min	Wake After Sleep Onset (WASO):	19.5 min

Sleep Stage

Stage	**Latency (out)**	**Latency (onset)**	**Duration (min)**	**/ TST (%)**	**/ TIB (%)**
N1	1.0	0.0	45.5	12.1	11.5
N2	6.0	5.0	157.0	41.8	39.5
N3	16.5	15.5	59.0	15.7	14.9
REM	49.5	48.5	114.5	30.5	28.8

AHI

	REM	**NREM**	**TST**
AHI	8.9	12.6	11.5

Respiratory Data

	Apnea	**Hypopnea**	**A+H**
Number	0	73	73
Mean Dur (sec)	0.0	26.4	26.4
Max Dur (sec)	0.0	76.6	76.6
Total Dur (min)	0.0	31.9	31.9
% of TST	0.0	8.5	8.5
Index (#/TST)	0.0	11.5	11.5
Index (REM)	0.0	8.9	8.9
Index (NREM)	0.0	12.6	12.6

Note. Latency (out/onset): from lights out/sleep onset.
A+H: Apnea+Hypopnea

Supplementary Figure 4. Automaticlly generated PSG Summary Report (Mild OSA). AI-generated results (top) were obtained by integrating *VOSA* for respiratory event detection and SleepXViT for sleep staging. Statistics exclusively computed from *VOSA* are highlighted in **green**, while metrics that rely on combined outputs from both models are highlighted in **blue**. Manually annotated results are shown below for comparison. Unavailable data (e.g., anonymized information) is represented as XX.

VOSA-SleepXViT Generated Polysomnography Report

Phone: (XXX) XXX-XXXX

Patient number	C2018-EM-01-0056	**Age**	46	**Acquisition**	XXXXX
Started	07/20/18 at 09:37:20 PM	**Sex**	Male	**Type**	Adult
Stopped	07/21/18 at 06:03:14 AM	**BMI**	27.6	**Duration**	505.1

Sleep Data

Lights Out:	09:37:20 PM	Sleep Onset:	09:40:50 PM
Lights On:	XX:XX:XX AM	Sleep Efficiency:	59.2 %
Total Recording Time:	505.1 min	Sleep Latency (from Lights Off):	3.5 min
Total Sleep Time (TST):	299.5 min	REM Latency (from Sleep Onset):	416.5 min
Total Wake Time (TWT):	206.5 min	Wake After Sleep Onset (WASO):	203.0 min

Sleep Stage

Stage	**Latency**	**Latency**	**Duration**	**/ TST**	**/ TIB**
	(out)	**(onset)**	**(min)**	**(%)**	**(%)**
N1	3.5	0.0	76.5	25.5	15.1
N2	8.0	4.5	148.5	49.6	29.4
N3	117.0	113.5	54.0	18.0	10.7
REM	420.0	416.5	20.5	6.9	20.5

AHI

	REM	**NREM**	**TST**
AHI	5.9	23.4	22.4

Respiratory Data

	Apnea	**Hypopnea**	**A+H**
Number	56	56	112
Mean Dur (sec)	17.5	14.8	16.2
Max Dur (sec)	30.0	26.0	30.0
Total Dur (min)	16.4	13.8	30.2
% of TST	5.5	4.6	10.1
Index (#/TST)	11.2	11.2	22.4
Index (REM)	0.0	5.9	5.9
Index (NREM)	11.8	14.7	23.4

Manually Annotated Polysomnography Report

Phone: (XXX) XXX-XXXX

Patient number	C2018-EM-01-0056	**Age**	46	**Acquisition**	XXXXX
Started	07/20/18 at 09:37:20 PM	**Sex**	Male	**Type**	Adult
Stopped	07/21/18 at 06:03:14 AM	**BMI**	27.6	**Duration**	505.1

Sleep Data

Lights Out:	09:37:20 PM	Sleep Onset:	09:44:50 PM
Lights On:	XX:XX:XX AM	Sleep Efficiency:	59.7 %
Total Recording Time:	505.1 min	Sleep Latency (from Lights Off):	7.5 min
Total Sleep Time (TST):	301.9 min	REM Latency (from Sleep Onset):	412.5 min
Total Wake Time (TWT):	204.0 min	Wake After Sleep Onset (WASO):	196.5 min

Sleep Stage

Stage	**Latency**	**Latency**	**Duration**	**/ TST**	**/ TIB**
	(out)	**(onset)**	**(min)**	**(%)**	**(%)**
N1	7.5	0.0	72.5	24.0	14.3
N2	8.0	0.5	155.0	51.4	30.6
N3	112.0	104.5	54.0	17.9	10.7
REM	420.0	412.5	20.5	6.8	4.1

AHI

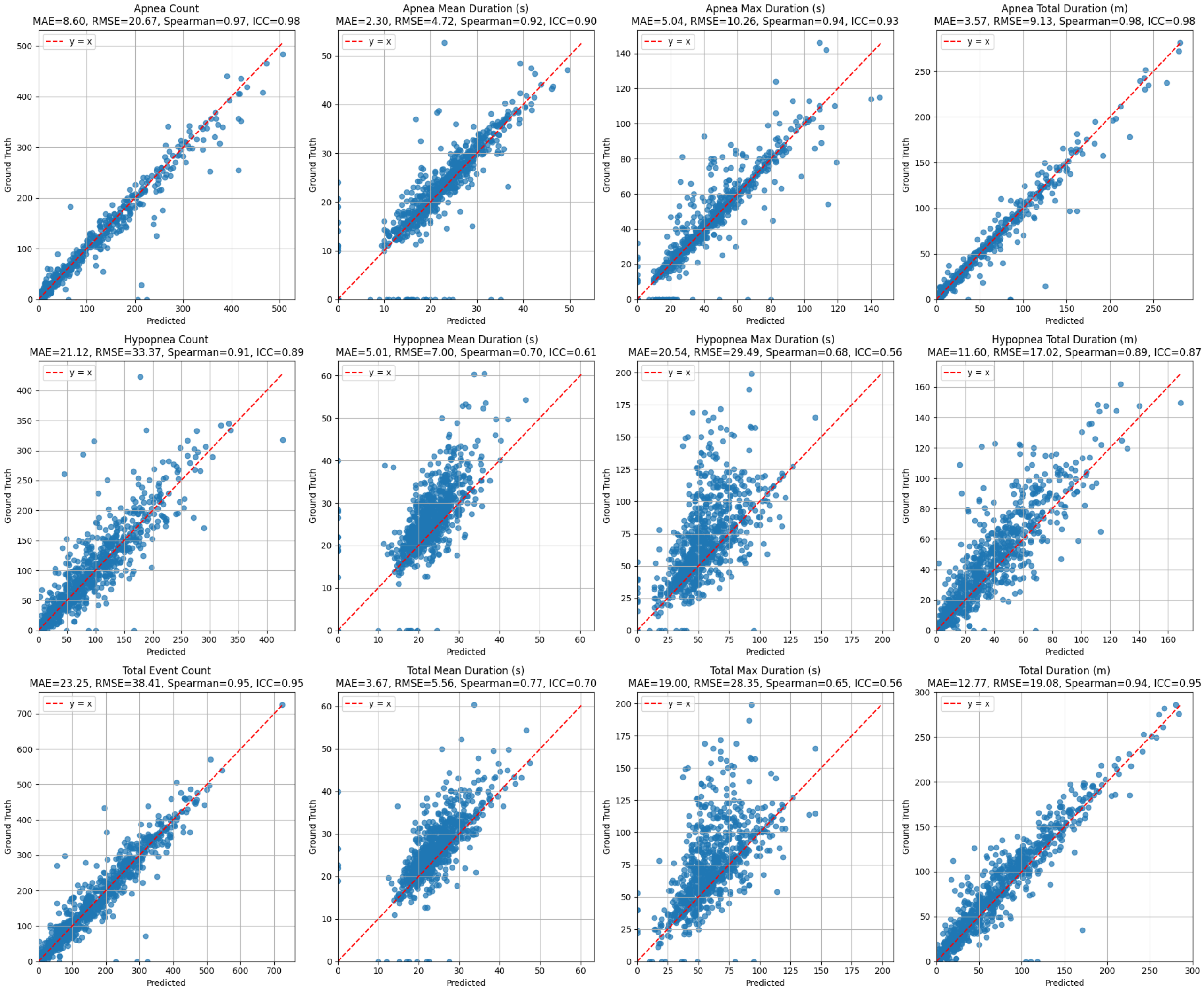
	REM	**NREM**	**TST**
AHI	14.6	23.7	23.1

Respiratory Data

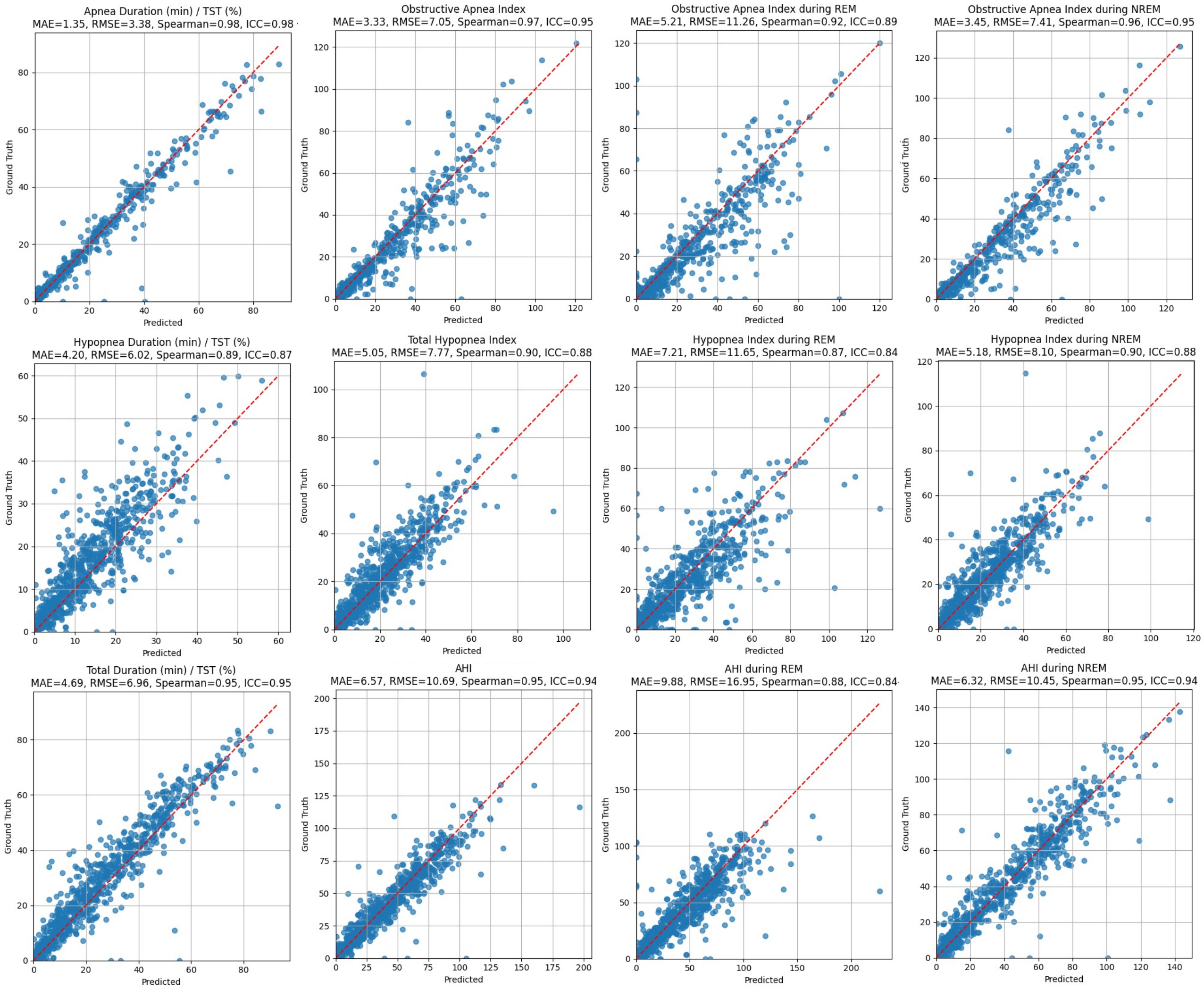
	Apnea	**Hypopnea**	**A+H**
Number	46	68	114
Mean Dur (sec)	19.9	19.3	20.3
Max Dur (sec)	34.7	32.0	34.7
Total Dur (min)	16.3	22.3	38.6
% of TST	5.4	7.3	12.8
Index (#/TST)	9.5	13.5	23.1
Index (REM)	0.0	14.6	14.6
Index (NREM)	10.2	13.4	23.7

Note. Latency (out/onset): from lights out/sleep onset.
A+H: Apnea+Hypopnea

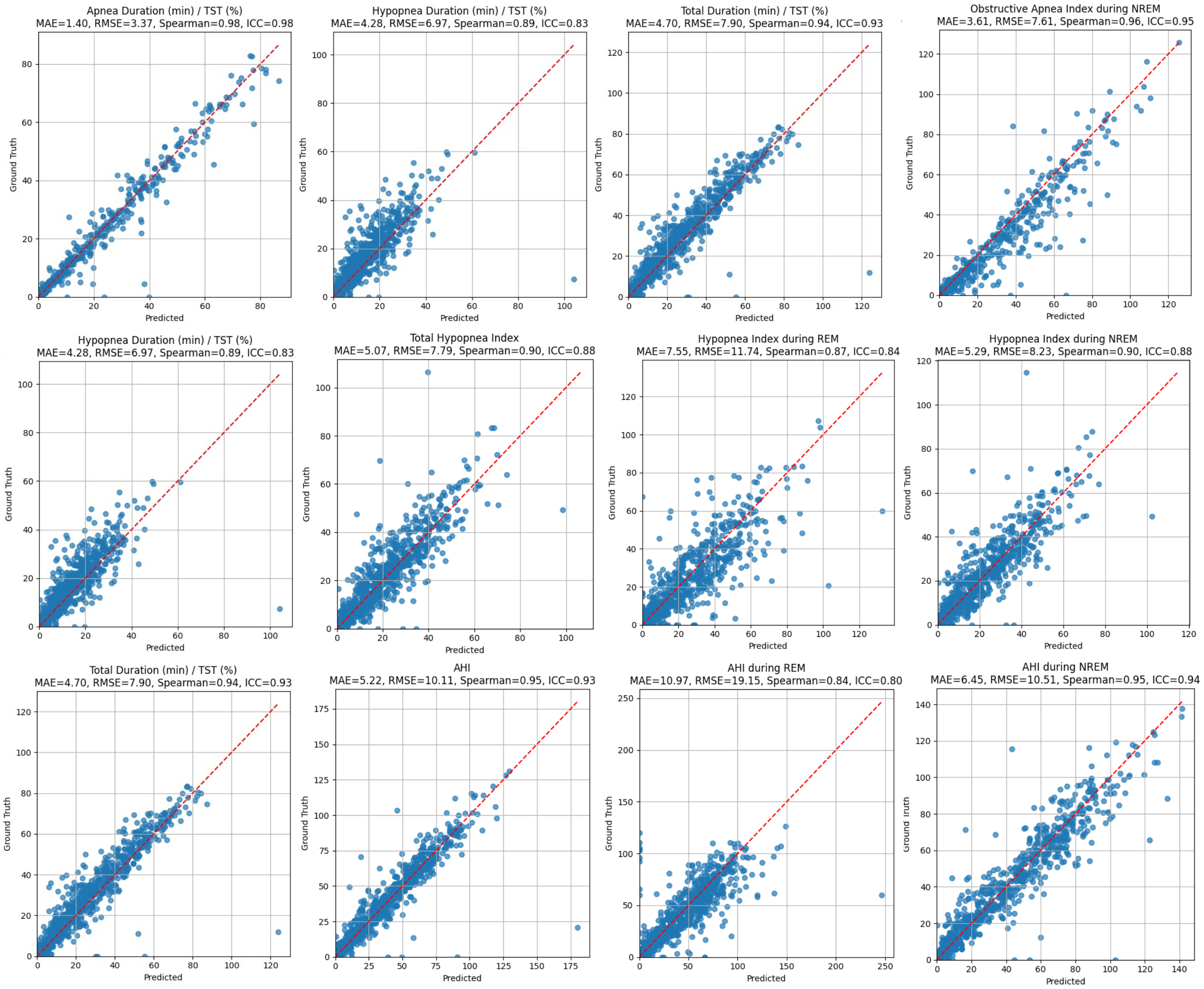
Supplementary Figure 5. Automaticlly generated PSG Summary Report (Moderate OSA). AI-generated results (top) were obtained by integrating *VOSA* for respiratory event detection and SleepXViT for sleep staging. Statistics exclusively computed from *VOSA* are highlighted in **green**, while metrics that rely on combined outputs from both models are highlighted in **blue**. Manually annotated results are shown below for comparison. Unavailable data (e.g., anonymized information) is represented as XX.



Supplementary Figure 6. Scatter plots for VOSA-exclusively generated PSG report statistics Each scatter plot illustrates the predicted versus ground-truth values for each PSG report statistic. The red dashed line indicates the ideal line ($y = x$). The first, second, and third rows correspond to statistics for apnea, hypopnea, and the combination of apnea and hypopnea (A+H), respectively.



Supplementary Figure 7. **Scatter plots for VOSA (w/ GT) generated PSG report statistics** Each scatter plot illustrates the predicted versus ground-truth values for each PSG report statistic. The values are estimated by integrating *VOSA*'s respiratory event predictions with ground-truth sleep stage annotations. The red dashed line indicates the ideal line ($y = x$). The first, second, and third rows correspond to statistics for apnea, hypopnea, and the combination of apnea and hypopnea (A+H), respectively.



Supplementary Figure 8. Scatter plots for VOSA (w/ SleepXViT) generated PSG report statistics Each scatter plot illustrates the predicted versus ground-truth values for each PSG report statistic. The values are estimated by integrating *VOSA*'s respiratory event predictions with SleepXViT's sleep stage predictions. The red dashed line indicates the ideal line ($y = x$). The first, second, and third rows correspond to statistics for apnea, hypopnea, and the combination of apnea and hypopnea (A+H), respectively.

Supplementary Table 1. Event-level comparison of AI-generated (VOSA-SleepXViT) and manually annotated PSG report (Patient no: C2020-EM-01-0133, Predicted AHI: 13.4 (Ground-Truth: 14.0), SleepXViT MF1: 90.7%)

Generated Report (*VOSA*-SleepXViT)		**Manually annotated Report**	
Event no.	**Event Description**	**Event no.**	**Event Description**
394	**Event Label: N1** Start Time: 2020/01/08 12:56:30 AM End Time: 2020/01/08 12:57:00 AM Start Epoch: 378, End Epoch: 379 Duration: 30.0 s	485	**Event Label: N1** Start Time: 2020/01/08 12:56:30 AM End Time: 2020/01/08 12:57:00 AM Start Epoch: 378, End Epoch: 379 Duration: 30.0 s
395	**Event Label: Hypopnea** Start Time: 2020/01/08 12:56:52 AM End Time: 2020/01/08 12:57:36 AM Start Epoch: 378, End Epoch: 380 Duration: 44.0 s	486	**Event Label: Hypopnea** Start Time: 2020/01/08 12:56:57 AM End Time: 2020/01/08 12:57:28 AM Start Epoch: 378, End Epoch: 379 Duration: 31.0 s
396	**Event Label: N2** Start Time: 2020/01/08 12:57:00 AM End Time: 2020/01/08 12:57:30 AM Start Epoch: 379, End Epoch: 380 Duration: 30.0 s	487	**Event Label: N2** Start Time: 2020/01/08 12:57:00 AM End Time: 2020/01/08 12:57:30 AM Start Epoch: 379, End Epoch: 380 Duration: 30.0 s
689	**Event Label: N1** Start Time: 2020/01/08 03:14:30 AM End Time: 2020/01/08 03:15:00 AM Start Epoch: 654, End Epoch: 655 Duration: 30.0 s	842	**Event Label: N1** Start Time: 2020/01/08 03:14:30 AM End Time: 2020/01/08 03:15:00 AM Start Epoch: 654, End Epoch: 655 Duration: 30.0 s
690	**Event Label: Apnea Obstructive** Start Time: 2020/01/08 03:14:39 AM End Time: 2020/01/08 03:15:01 AM Start Epoch: 654, End Epoch: 655 Duration: 22.0 s	843	**Event Label: Apnea Obstructive** Start Time: 2020/01/08 03:14:37 AM End Time: 2020/01/08 03:15:00 AM Start Epoch: 654, End Epoch: 655 Duration: 23.0 s
691	**Event Label: Wake** Start Time: 2020/01/08 03:15:00 AM End Time: 2020/01/08 03:15:30 AM Start Epoch: 655, End Epoch: 656 Duration: 30.0 s	846	**Event Label: Wake** Start Time: 2020/01/08 03:15:00 AM End Time: 2020/01/08 03:15:30 AM Start Epoch: 655, End Epoch: 656 Duration: 30.0 s

Ablation Study

Supplementary Note 4. Ablation Study

An ablation study was conducted on the KISS's per-second classification task to validate each component of *VOSA*'s architecture. Comparing the first and second rows, the stage 1 alone demonstrated substantial performance, however, stage 2 further improved the model, suggesting that *VOSA* effectively captures longer temporal dependencies through multi-epoch input. As shown in third and fourth rows, removing either of *VOSA*'s transformer blocks reduced performance compared to the second row, with a larger drop excluding the time-domain attention block. This underscores the importance of sequential pattern learning in OSA detection. In the last row, we applied data balancing, as employed in DRIVEN[?] by undersampling normal data, to avoid overfitting to the majority class. It resulted in a significant performance drop indicating that *VOSA*'s transformer architecture effectively captures general and comprehensive information from a data-driven approach using the entire dataset.

Supplementary Table 2. Ablation Study of *VOSA* architecture on KISS

Training Stage*		Transformer Block†		Data Balancing	Performance			
1	2	Time	Signal		Acc	MF1	Sens	wF1
✓	✓	✓	✓		**90.6**	**82.6**	**82.0**	**90.4**
✓		✓	✓		89.6	80.0	78.9	89.2
✓		✓			88.4	78.2	77.3	88.0
✓			✓		85.2	68.8	67.7	83.8
✓		✓	✓	✓	78.4	72.8	80.5	80.9

The checkmark (✓) indicates the inclusion of corresponding components.

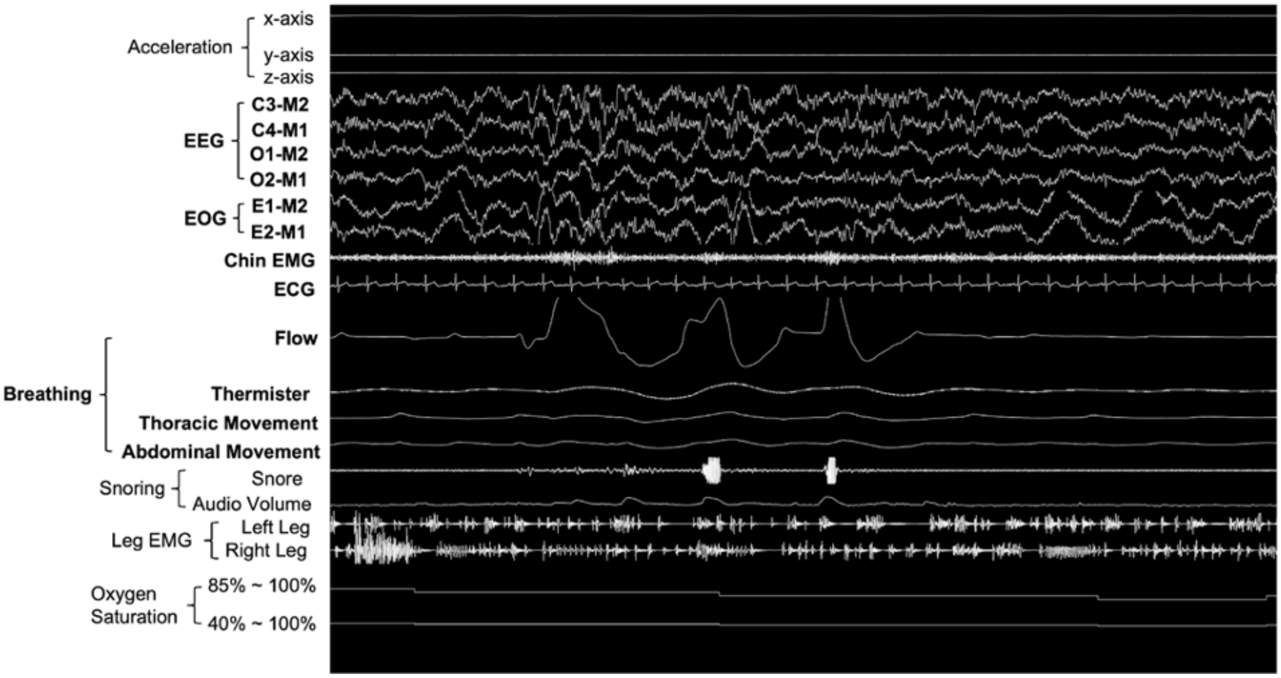
Each row represents an independent experiment, with the first row presenting *VOSA* s final per-second performance.

*Stages 1 and 2 incorporate single epoch and multiple epochs as inputs, respectively.

†Time and Signal indicates *VOSA*'s two transformer blocks, designed to learn patterns in the time and signal domains, respectively.

Definition of abbreivations: Acc = Accuracy, MF1 = Macro F1 score, Sens = Sensitivity, wF1 = weigthed F1-score.

Discussion

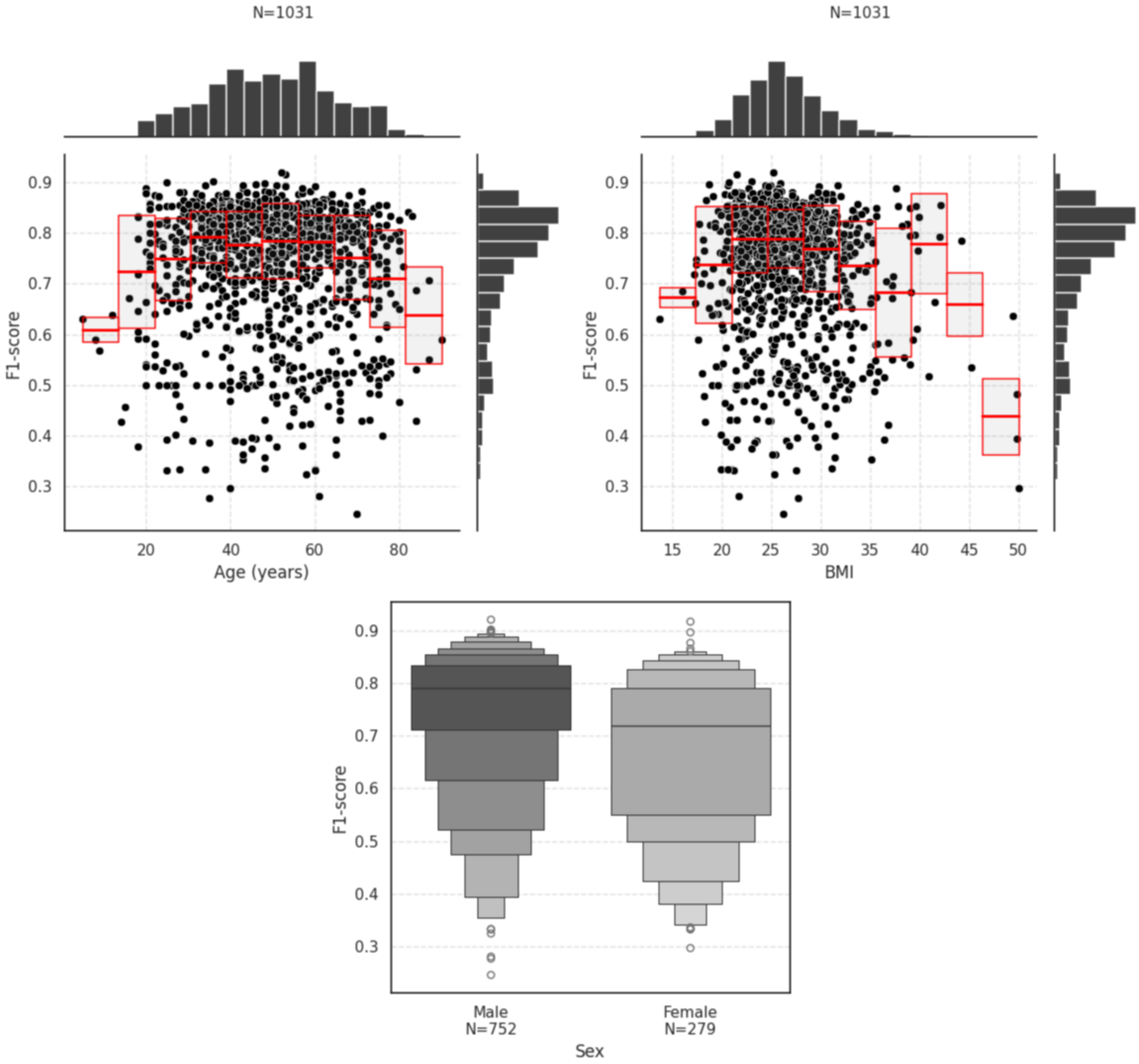


Supplementary Figure 9. A sample of standardized PSG image data from KISS dataset. Each image represents 30 seconds, encompassing 21 biosignal channels. The image is rendered as a high-resolution PNG file with 1920×1080 resolution. The X-axis represents the time, while the Y-axis corresponds to different biosignal channels.

Supplementary Note 5. Demographic Bias

We investigated potential demographic biases in *VOSA* by analyzing the effects of age, BMI, sex, and dataset origin. Supplementary Figure 10 visualizes the distribution of mean F1 scores across individual PSG records for the per-second event classification task as a function of age, BMI, and sex. These plots illustrate correlations rather than causal relationships. The analysis included 1,031 PSG records from the test sets of KISS and SHHS-2, comprising 770 and 261 records, respectively, where demographic information were available.

To quantify the impact of demographic variables, we employed a beta regression model predicting mean F1 scores based on four covariates: age, BMI, sex, and dataset origin. The estimated model coefficients were as follows: age 0.001 ± 0.001(95% CI, $z = 1.061, p = 0.288$), BMI -0.016 ± 0.004(95% CI, $z = -4.136, p < 0.001$), and sex 0.366 ± 0.040(95% CI, $z = 9.176, p < 0.001$). These results indicate that model performance (F1 score) decreases with increasing BMI and for female subjects. Both BMI and sex had statistically significant effects ($p < 0.001$), while age did not show a significant impact. The beta regression analysis was conducted using Python 3.8 and the statsmodels v0.14.1 package.



Supplementary Figure 10. **Correlations of *VOSA* performance and demographic variables.** Each plot presents the mean F1-scores per subject for age, BMI, and sex, positioned in the upper left, upper right, and bottom, respectively. In the age and BMI plots, red boxes represent bins divided into 10 equal intervals, with red center lines indicating medians, while the lower and upper red lines denote the interquartile range. In the sex plot, the central black line represents the median, with other lines depicting various quantiles. The outliers are displayed as round-shaped points and the k-depth parameter is set to 5.

Method

Supplementary Note 6. Apnea-Hypopnea Index (AHI)

Apnea-Hypopnea Index (AHI)[2–4], the main clinical index in the diagnosis of OSA, was calculated for per-patient OSA severity classification. AHI measures the number of apnea and hypopnea events per hour during sleep, formulated as (1)

$$\text{Apnea-Hypopnea Index (AHI)} = \frac{\text{Number of apnea events + hypopnea events}}{\text{Total Sleep Time}} \tag{1}$$

OSA severity is categorized as follows: normal (AHI < 5), mild ($5 \leq$ AHI < 15), moderate ($15 \leq$ AHI < 30), and severe (AHI ≥ 30). In this work, per-second event labels were aggregated to determine the number of events, counting consecutive event labels over 10 seconds as a single event, while the total sleep time was obtained from the PSG reports.

References

1. Jeong, J. *et al.* Standardized image-based polysomnography database and deep learning algorithm for sleep-stage classification. *Sleep* zsad242 (2023).

2. Berry, R. B. *et al.* The aasm manual for the scoring of sleep and associated events: Rules, terminology, and technical specifications. In *American Academy of Sleep Medicine*, 176 (Darien, Illinois, 2012).

3. Berry, R. B. *et al.* The aasm manual for the scoring of sleep and associated events: Rules, terminology, and technical specifications, version 2.2. In *American Academy of Sleep Medicine* (American Academy of Sleep Medicine, Darien, Illinois, 2015).

4. Berry, R. B., Brooks, R., Gamaldo, C. E. *et al.* The aasm manual for the scoring of sleep and associated events: Rules, terminology, and technical specifications. In *American Academy of Sleep Medicine* (American Academy of Sleep Medicine, Darien, Illinois, 2017).

5. AIHub. Image of sleep quality assessment and sleep disorder diagnosis. In *AIHub Data Repository* (2021).

6. Redline, S. *et al.* Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. *Sleep* **21**, 759–767 (1998).