Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can eye movements flag sleep apnea? Single-channel electrooculography with external validation

Shrreya Jain

sdzhein.3m19@knmu.edu.ua

Kharkiv National Medical University

Research Article

Keywords: sleep-disordered breathing, obstructive sleep apnea, electrooculography, machine learning, wearable devices, polysomnography, external validation

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11812375/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Obstructive sleep apnea affects close to a billion adults, and most are never diagnosed. Wearable sleep devices already record eye movement (electrooculography, EOG) to score sleep, but whether a model built on this signal alone can detect breathing events in new patients has, to our knowledge, not been tested. We used two public polysomnography datasets: the UCD Sleep Apnea Database (25 adults, both eye channels) and a random sample of 100 adults from the 2018 PhysioNet/Computing in Cardiology Challenge (CinC; one eye channel). Each 30-second sleep epoch was labeled for apnea or hypopnea. Gradient-boosted models built on 19 spectral and waveform features per eye channel, with settings frozen before external testing, were compared with a model that knew only the sleep stage. Models were tested leave-one-patient-out within UCD, then trained on one dataset and tested on the other, in both directions.

Results

Within UCD, the two-eye model reached an area under the ROC curve (AUC) of 0.667 (95% CI 0.628–0.712), against 0.602 for sleep stage alone. A single eye reached 0.650. Trained on UCD and tested on the 100 CinC patients, the single-eye model reached the same AUC of 0.650 (0.628–0.672), against 0.609 for stage alone (difference 0.041, 95% CI 0.017–0.064). In the reverse direction the single eye reached 0.677 against 0.653 for stage alone, a smaller gain whose interval included zero. Left–right eye correlation and 16–30 Hz power were the most informative features. The signal was strongest in N2 and N3 sleep and weakest in N1, and within UCD it detected events with and without a scored arousal equally well. At the patient level, the eye signal was only weakly related to the apnea–hypopnea index in the external cohort ($\rho = 0.22$).

Conclusions

A single EOG channel carries a modest signal of apneas and hypopneas that adds to sleep staging and kept its accuracy in an independent cohort from another country. It cannot diagnose sleep apnea or estimate severity, but eye channels that are already part of wearable sleep devices could add information to home testing.

Background

Obstructive sleep apnea (OSA) is one of the most common chronic conditions in adults. A literature-based estimate put the number of adults aged 30 to 69 with OSA worldwide at about 936 million, with 425 million of them in the moderate-to-severe range [1]. Most of these people do not know they have it.

In a classic community study, 93% of women and 82% of men with moderate-to-severe sleep apnea syndrome had never been clinically diagnosed [2].

The reference test is still in-laboratory polysomnography, which needs a sleep bed, trained staff and manual scoring. Waiting lists are long, and patients with limited mobility or without a caregiver often cannot attend [3, 4]. Home sleep apnea tests have improved access, but they have their own gaps. Without a brain or eye signal they cannot tell when the patient is actually asleep, so the event rate is divided by recording time instead of sleep time, and the apnea–hypopnea index (AHI) tends to be underestimated [3].

This gap is one reason home devices are starting to add electrodes on the forehead and around the eyes. Wearable headbands such as the Dreem headband stage sleep about as well as individual human scorers [5]. More recently, a forehead patch with one electrode placed below the eye was added to a home apnea test to record a single combined EEG, EOG and EMG lead; it staged sleep with a kappa of about 0.70 against polysomnography in 106 patients with suspected OSA [6]. In these devices, the eye-movement channel (electrooculography, EOG) is used only to score sleep. Its respiratory content has not been studied.

There are physiological reasons to expect that it might carry some. Apneas and hypopneas often end in a brief arousal, with a burst of fast EEG activity and muscle tone that also reaches electrodes placed near the eyes [7]. The EOG is also contaminated by frontal EEG, so changes in sleep depth around events may show up in it. Single-channel EEG has already been used to detect apnea: a convolutional network trained on about 2,650 recordings reached an AUC of 0.80 within its own dataset, although performance dropped noticeably when it was tested on other datasets [8]. A PubMed search in November 2026 (electrooculography or EOG, combined with sleep apnea or sleep-disordered breathing terms and with detection or machine-learning terms; 124 records) found the EOG used for apnea-related work only together with other signals. A recent deep-learning model located apnea events from combined EEG and EOG and kept much of its accuracy in a second cohort [9], and other studies classified patients' sleep disorders from the EOG combined with EMG or EEG [10, 11]. In a robustness check, the EEG and EOG model of Elmi and Elmi lost 9.6% of its event-detection F1 score when given only the EOG at test time, without retraining [9]. This suggests that the EOG carries much of the respiratory information, but we found no study that built and externally validated a model on the EOG alone, or compared it with what sleep stage alone can explain.

In this study, we asked whether apneas and hypopneas can be picked up from eye-movement signals alone, whether this adds anything beyond knowing the sleep stage, and whether the result holds up in patients from a different hospital and country. We framed it as a feasibility study using public data, so that every step can be checked and repeated.

Methods

Study design and reporting

This was a retrospective diagnostic modeling study on two public, de-identified polysomnography datasets. It had one development cohort and one external cohort, and validation was run in both directions. Reporting follows the TRIPOD + AI statement [12].

Data sources

UCD Sleep Apnea Database (development cohort). This PhysioNet dataset contains 25 full overnight polysomnograms from adults referred to the sleep disorders clinic at St Vincent's University Hospital, Dublin, for suspected obstructive sleep apnea, central sleep apnea or primary snoring [13, 14]. Of the 25 patients, 21 were men; mean age was 50.0 ± 9.5 years (range 28–68) and mean body mass index 31.6 ± 4.0 kg/m². Each record includes left and right EOG channels, sleep stages scored in 30-second epochs according to Rechtschaffen and Kales (wake, REM, stages 1–4), and respiratory events scored by a sleep technologist with their onset, duration and whether they ended in an arousal. The 3,318 apneas and hypopneas comprised 1,433 obstructive, 1,076 central and 114 mixed hypopneas, and 216 obstructive, 343 central and 136 mixed apneas; 43% were therefore central. In all, 866 of the 3,318 events (26%) were scored with an arousal. Events marked as "possible" (19) and periodic-breathing segments (91) were not counted as events.

2018 PhysioNet/Computing in Cardiology Challenge (external cohort). This dataset comes from the sleep laboratory at Massachusetts General Hospital, USA [15]. Sleep stages were scored according to the American Academy of Sleep Medicine (AASM) manual, and arousals and respiratory events were annotated by certified technologists. From the 994 records in the public training set, we drew a random sample of 100 with a fixed seed (2018) before any analysis. These records carry a single EOG channel (E1–M2, 200 Hz), sleep stages in 30-second epochs (W, N1, N2, N3, R), and obstructive, central and mixed apneas and hypopneas with start and end times. Age and sex are not part of the public records and were not used.

Signal preparation and labels

The CinC signal was scaled using the units given in each record header and resampled from 200 Hz to 64 Hz with a polyphase filter, so that both datasets were analyzed at the same rate as the UCD EOG. Because every feature was later standardized within each recording, absolute amplitude units do not affect the models. Each recording was divided into 30-second epochs aligned with the sleep staging. Only epochs scored as sleep (REM, N1, N2 or N3) were used for training and testing. UCD stages 3 and 4 were merged into N3 in every analysis, and epochs scored as artifact or indeterminate, or lying before the first CinC stage annotation, were excluded.

An epoch was labeled positive if any part of it overlapped a scored apnea or hypopnea of any type, and negative otherwise. For each patient, we also calculated the AHI as the number of scored events divided by the hours of scored sleep. We used these values, rather than the AHI reported with each dataset, so that both cohorts were measured the same way.

Features

From each EOG channel in each epoch we extracted 19 features. After subtracting the epoch mean, the power spectrum was estimated with Welch's method (4-second windows). From it we took the log power and the relative power in six bands: 0.3–2, 2–4, 4–8, 8–12, 12–16 and 16–30 Hz (12 features). We added the log variance, Hjorth mobility and complexity [16], the log line length, kurtosis, skewness and the zero-crossing rate (7 features). When both eyes were available, we also calculated the Pearson correlation between the left and right channels in each epoch. During normal eye movements the two channels move in opposite directions, while signals that reach both electrodes equally, such as EEG or muscle activity, push them in the same direction.

Figure 1 shows where the EOG electrodes sit and what these signals look like around obstructive apneas in one patient from the development cohort.

Figure 1. What the model sees. (A) Standard EOG montage: E1 sits 1 cm below the outer corner of the left eye and E2 1 cm above the outer corner of the right eye, both referred to the mastoid (the AASM-recommended montage; the UCD montage is not documented). Besides eye movements, each electrode picks up frontal brain activity and muscle tone. (B) Ninety seconds of N2 sleep in UCD patient ucddb027 with four scored obstructive apneas (shaded). From top: airflow, shown for reference only and not used by the model; the left and right EOG signals; power in the 16–30 Hz band of the left EOG (2-second windows); and the correlation between the two eye channels (4-second windows). The bottom strip shows how the 30-second epochs were labeled. Arrows mark rises in fast activity after two of the apneas. This example is illustrative only; individual events vary, and the model's overall behavior is summarized in Figs. 2 and 3.

Every feature was standardized within each recording (z-score across all of that recording's epochs, including wake). This step uses no labels and removes differences in amplitude between people, electrodes and recording systems, but it does mean that a whole night has to be recorded before it can be analyzed. Because respiratory events cluster in time, we added temporal context for a fixed set of key features: their values in the two epochs before and after and a centered five-epoch rolling mean. In the two-eye model these were the 16–30 Hz and 8–12 Hz log power and log variance of each eye, the 0.3–2 Hz log power of the left eye and the left–right correlation (79 EOG inputs in total). In the single-eye models they were the left-eye versions of the same features (39 inputs). Missing neighbors at the start and end of a recording were set to zero. Sleep stage was coded as indicator variables for the current epoch and its two neighbors.

Models

We trained histogram-based gradient-boosted tree classifiers (scikit-learn 1.9.1) [17] with settings chosen during development on UCD without formal tuning and frozen before the external analysis: 120 iterations without early stopping, learning rate 0.08, at most 15 leaves per tree, balanced class weights and a fixed random seed. For each setting we compared three models: sleep stage only, EOG only, and EOG plus sleep stage. The stage-only model is the reference, because apneas are more common in some sleep stages than in others, so any model that could only recognize the stage would still score above chance.

Validation

We validated in three ways:

1. **Internal (UCD, both eyes)**: leave-one-patient-out cross-validation, in which each patient is predicted by a model trained on the other 24. The same procedure was repeated with left-eye features only, to check how much was lost by using a single channel.

2. **External (UCD → CinC)**: single-eye models trained on all 25 UCD patients and applied without change to the 100 CinC patients.

3. **Reverse external (CinC → UCD)**: single-eye models trained on all 100 CinC patients and applied to the 25 UCD patients.

No patient's data were used for both training and testing in any setting.

Statistical analysis

The main outcome was the epoch-level area under the receiver operating characteristic curve (AUC), pooled over all test epochs. Because a pooled AUC partly reflects differences between patients, we also report the median and interquartile range of the AUC calculated within each patient. At a probability threshold of 0.5 we report the Matthews correlation coefficient (MCC), sensitivity and specificity, and we assessed calibration with the calibration slope and the mean predicted probability against the observed event rate. At the patient level, we averaged each patient's predicted probabilities across their sleep epochs, compared this score with the AHI using Spearman's rank correlation, and calculated the AUC for identifying patients with AHI ≥ 15 events/hour. To check whether detection depended on arousal, we calculated the AUC separately for event epochs whose event was scored with an arousal and those without, each against all event-free epochs.

We report point estimates with 95% confidence intervals from a patient-level bootstrap (2,000 resamples), in which whole patients are resampled so that epochs from the same night stay together. Differences in AUC between models were estimated within the same bootstrap samples. These intervals treat the trained models as fixed, so they do not include the variation that would come from retraining on

a different set of patients. To see which kinds of features the model relied on, we used grouped permutation importance within the internal cross-validation: for each feature group (including its context versions), we shuffled its values across the test epochs five times and recorded the mean fall in AUC. Permutation importance understates groups that are correlated with other groups, so it shows which information the model used, not which information is unique.

Software and code

Analyses were done in Python 3 with NumPy, SciPy, pandas and scikit-learn. Signal extraction from the CinC files was run in Google Colaboratory. The code and the derived feature tables are available as described under Availability of data and materials.

Results

Cohorts

The two cohorts had a similar share of patients with moderate-to-severe apnea (AHI ≥ 15), but UCD had more severe cases (Table 1). Respiratory events overlapped 26% of sleep epochs in UCD and 22% in CinC.

Table 1
Characteristics of the two cohorts

Patients	**UCD (development)**	**CinC 2018 (external)**
	25	**100 (random sample of 994)**
Men, n	21	not available
Age, years, mean ± SD	50.0 ± 9.5	not available
Sleep scoring	Rechtschaffen and Kales	AASM
EOG channels	Left and right	E1–M2 only, 200 Hz (resampled to 64 Hz)
Scored sleep, hours, mean ± SD	5.4 ± 0.9	6.0 ± 1.1
Sleep epochs analyzed	16,067	72,219
Apneas and hypopneas scored	3,318	11,056
Sleep epochs with an event, %	25.6	21.5
AHI, events/h, median (IQR)	15.8 (12.0–33.6)	17.4 (8.0–25.5)
AHI range, events/h	2.2–94.8	0.7–82.4
AHI ≥ 15, n (%)	14 (56)	58 (58)
AHI ≥ 30, n (%)	8 (32)	10 (10)

AHI was recalculated from the scored events and hours of scored sleep in both cohorts; in UCD it agreed closely with the AHI reported with the dataset (Spearman ρ = 0.998). Age and sex are not part of the public CinC records. Of 91,971 CinC epochs, 2,610 before the first stage annotation and 17,142 scored as wake were excluded, leaving 72,219.

Performance in each validation setting

In every setting, the EOG models scored higher than the model that knew only the sleep stage (Table 2). The gain was statistically clear in the internal two-eye analysis and in the external test, and smaller and less certain in the internal single-eye analysis and the reverse test.

Table 2
Epoch-level and patient-level performance

Setting	Model	Epoch AUC (95% CI)	Per-patient AUC, median (IQR)	MCC	Sensitivity / specificity	Spearman ρ with AHI (95% CI)	AUC for AHI ≥ 15 (95% CI)
Internal, UCD (25 patients)	Sleep stage only	0.602 (0.545–0.666)	0.640 (0.571–0.706)	0.199	0.62 / 0.61	0.49 (0.07 to 0.80)	0.70 (0.47–0.89)
	EOG, both eyes	0.667 (0.628–0.712)	0.663 (0.612–0.740)	0.205	0.60 / 0.63	0.46 (0.06 to 0.72)	0.68 (0.43–0.90)
	Both eyes + stage	0.689 (0.651–0.729)	0.684 (0.637–0.740)	0.223	0.61 / 0.64	0.51 (0.12 to 0.78)	0.71 (0.49–0.92)
	EOG, left eye	0.650 (0.606–0.700)	0.680 (0.574–0.749)	0.177	0.58 / 0.62	0.33 (− 0.09 to 0.66)	0.62 (0.37–0.85)
	Left eye + stage	0.656 (0.612–0.706)	0.673 (0.576–0.733)	0.180	0.59 / 0.61	0.44 (0.04 to 0.75)	0.67 (0.44–0.87)
External, UCD → CinC (100 patients)	Sleep stage only	0.609 (0.586–0.632)	0.654 (0.545–0.716)	0.126	0.51 / 0.64	0.10 (− 0.10 to 0.31)	0.53 (0.42–0.65)
	EOG, left eye	0.650 (0.628–0.672)	0.679 (0.599–0.747)	0.174	0.58 / 0.63	0.22 (0.02 to 0.40)	0.64 (0.53–0.75)
	Left eye + stage	0.658 (0.637–0.679)	0.699 (0.610–0.755)	0.179	0.59 / 0.62	0.22 (0.02 to 0.40)	0.62 (0.51–0.73)
Reverse, CinC → UCD (25 patients)	Sleep stage only	0.653 (0.598–0.709)	0.634 (0.568–0.711)	0.167	0.40 / 0.77	0.42 (− 0.01 to 0.73)	0.67 (0.43–0.87)
	EOG, left eye	0.677 (0.624–0.735)	0.697 (0.635–0.778)	0.217	0.63 / 0.62	0.46 (0.03 to 0.77)	0.67 (0.42–0.89)
	Left eye + stage	0.693 (0.638–0.749)	0.692 (0.616–0.771)	0.233	0.64 / 0.62	0.49 (0.05 to 0.78)	0.69 (0.45–0.90)

AUC = area under the receiver operating characteristic curve; MCC = Matthews correlation coefficient; sensitivity, specificity and MCC at a probability threshold of 0.5. Per-patient AUC is calculated within

each patient (all patients had both event and event-free epochs). Confidence intervals are from a patient-level bootstrap (2,000 resamples). Gradient-boosted models with identical settings throughout.

Internal validation. Within UCD, the two-eye model improved on sleep stage alone by 0.065 AUC (95% CI 0.003 to 0.128), and the two eyes plus stage by 0.087 (0.034 to 0.138). With the left eye only, the gain was 0.048 (− 0.011 to 0.110), and 0.055 (0.005 to 0.106) with stage added. Measured within each patient, the EOG models again ranked above stage alone, though by less than the pooled AUC suggests.

External validation. Applied without change to the 100 CinC patients, the single-eye model trained on UCD reached an AUC of 0.650, the same as within UCD. It improved on stage alone by 0.041 (0.017 to 0.064), and the left eye plus stage by 0.049 (0.030 to 0.068). Stage alone transferred less well: its sensitivity and MCC fell more than those of the EOG models.

Reverse validation. Trained on the 100 CinC patients and tested on UCD, all three models scored higher than in the other settings, including stage alone (0.653). The single eye reached 0.677 and the left eye plus stage 0.693, but with only 25 test patients neither gain over stage alone was statistically clear (0.024, − 0.045 to 0.095; and 0.039, − 0.003 to 0.092). Calibration was close to ideal in this direction (slope 0.99–1.00). In the other settings, the EOG models had calibration slopes of 0.57–0.65 and over-predicted events on average (mean predicted probability 0.41–0.44 against an event rate of 0.22–0.26), as expected from training with balanced class weights.

Performance by sleep stage

The EOG signal was clearly stronger in N2 and N3 than in N1 (Fig. 2). N1 was close to chance in every setting (0.51–0.53), and REM was intermediate (0.58–0.60). N3 estimates were the least precise, because fewer than 10% of N3 epochs contained an event.

Computed from this study's leave-one-out and cross-dataset predictions; EOG-only gradient-boosted models

Figure 2. Epoch-level AUC of the EOG-only model within each sleep stage, with 95% bootstrap confidence intervals. Dark: internal leave-one-patient-out validation in UCD (both eyes); medium: trained on UCD, tested on CinC (left eye); light: trained on CinC, tested on UCD (left eye). The dashed line marks chance (AUC 0.5).

Which features mattered

In the internal two-eye model, two feature groups carried most of the information: the correlation between the left and right eye channels, and power in the 16–30 Hz band (Fig. 3). Power in the 8–12 Hz band contributed nothing that other features did not also carry. In N2 sleep, epochs with a respiratory event had a higher EOG amplitude, more 16–30 Hz power and a lower left–right correlation (median 0.31 vs. 0.51 in epochs without an event).

Computed from this study's grouped permutation importance, leave-one-patient-out in UCD (25 patients)

Figure 3. Grouped permutation importance in the internal two-eye EOG-only model: the fall in epoch-level AUC when all features in a group, including their temporal-context versions, were shuffled across test epochs (mean of five shuffles; base AUC 0.667). Groups that overlap with others, such as amplitude and Hjorth parameters, are understated by this method.

Patient-level results

The patients' average predicted probability was moderately related to their AHI within UCD (ρ = 0.46 for both eyes) and in the reverse test (0.46), but only weakly in the external test (0.22, 95% CI 0.02 to 0.40). For the EOG models, the AUC for identifying patients with AHI ≥ 15 ranged from 0.62 to 0.71 across settings, with wide confidence intervals. These models were built to flag epochs, not to estimate severity, and the patient-level results should be read in that light.

Events with and without arousal

The EOG models detected events that ended in a scored arousal and events that did not almost equally well: AUC 0.667 vs. 0.666 within UCD (1,153 and 2,954 event epochs), and 0.694 vs. 0.671 when the CinC-trained model was applied to UCD. This analysis was limited to UCD, because the CinC arousal annotations leave out most arousals linked to respiratory events. By contrast, the stage-only model picked up events with an arousal much better than those without (0.671 vs. 0.575 within UCD), possibly because arousals change the scored stage.

Discussion

In this study, eye-movement signals picked up apneas and hypopneas better than sleep stage alone, and a single eye channel kept its accuracy when moved to patients from a different hospital, country and scoring system. Trained on 25 patients in Dublin, it reached an AUC of 0.650 in 100 patients from Boston, the same as within the training cohort. The effect is modest, and the gain over sleep stage was not statistically clear in every setting: it was clear for both eyes within UCD and for one eye in the external test, but smaller and uncertain for one eye within UCD and in the reverse test. The direction of the effect, however, was the same in all four analyses.

What the eye channel may be picking up

The two most informative feature groups were the left–right correlation and power in the 16–30 Hz band. During events in N2, the EOG had a higher amplitude, more fast activity and a lower correlation between the two eyes. Because conjugate eye movements drive the two channels in opposite directions, while EEG and muscle activity tend to reach both electrodes alike, a lower correlation means that eye movement made up a larger share of the signal, or that the shared EEG or muscle signal fell. We cannot tell which from these data.

We expected the model to work mainly by detecting the arousal that often ends an event. It did not: events with and without a scored arousal were detected almost equally well in UCD, the cohort with

arousals linked to each event. The EOG may therefore be picking up changes during the event itself, or disturbances too brief or too small to be scored as arousals. The signal was strongest in N2 and N3, where the EOG is otherwise quiet, and close to chance in N1, where the signal is unsettled anyway. These explanations are hypotheses; testing them would need event-level analysis and simultaneous EEG and respiratory signals.

Comparison with other single-channel approaches

As far as we know, this is the first study to train and externally validate an apnea detector on the EOG alone, so there is no direct benchmark. The closest work used single-channel EEG. Barnes et al. trained a convolutional neural network on central EEG from about 2,650 recordings and reached an AUC of 0.80 within that dataset, but performance fell noticeably when the model was applied to other datasets, including UCD, which the authors attributed partly to differences in how events were annotated [8]. Their labels, model and validation design differ from ours, so the numbers cannot be compared directly. Elmi and Elmi located apnea events from combined EEG and EOG with a much larger deep-learning model and also tested it across cohorts [9]. In their robustness test, the model kept most of its performance with the EOG alone; together with our results, this suggests that a substantial part of what such models use comes from the EOG itself. The shared lesson is that single-channel methods lose accuracy when they move between laboratories, and that testing across datasets is the honest benchmark for any of them.

Why this could matter

The practical case for the EOG is that it is already being recorded. Sleep headbands and forehead patches are being designed to add sleep staging to home testing [5, 6]. If the same electrode carries information about respiratory events, it could add to home testing at no extra cost to the patient, for example by supporting events picked up by oximetry or airflow sensors. At this level of accuracy it is not a replacement for respiratory sensors, and we do not suggest using it to diagnose or rule out apnea.

The models were built to flag individual epochs, and they were not good at ranking patients by severity, particularly in the external cohort ($\rho = 0.22$). A patient-level score designed for that purpose, such as the number of detected events per hour of sleep, would need to be developed and tested separately. The models were also poorly calibrated outside the reverse test: they over-predicted events on average, as expected after training with balanced class weights, so their probabilities should not be read as event risks without recalibration.

Strengths and limitations

The main strength of this study is that it was tested outside the data it was built on, in both directions, with model settings frozen before external testing and no tuning on the test cohort. Both datasets are public and the code is available, so every result can be checked. Comparing every model with a sleep-stage-only baseline also guards against a common trap: an apnea detector that is really a sleep-stage detector.

There are several limitations. The development cohort was small (25 patients, 21 of them men), and the feature set and model settings were chosen on it, so the reverse test is not fully independent of the development data. Only one eye channel was available in the external cohort, so the two-eye model, including the left–right correlation, could not be tested externally. The two datasets used different staging and scoring rules, hypopnea definitions strongly affect event counts [18], and 43% of UCD events were central, so our labels cover sleep-disordered breathing in general rather than obstructive events alone. CinC arousal annotations exclude most respiratory-related arousals, so arousal-stratified results are reported for UCD only. Each recording was standardized to its own night, so the method in its current form needs a full night before it can be applied. Labels covered whole 30-second epochs, and an event that only touched an epoch counted the same as one that filled it. Our confidence intervals treat the trained models as fixed and do not include the variation that would come from retraining on other patients. We used hand-crafted features and a standard classifier rather than deep learning. Finally, both cohorts were patients referred for suspected sleep-disordered breathing and recorded in laboratories with wet electrodes, so results in a general population or with dry wearable electrodes may differ.

Future work should test the two-eye model in a larger external cohort with both EOG channels, such as the Sleep Heart Health Study, and examine whether adding the EOG to oximetry improves on oximetry alone. Event-level detection, normalization that works in real time, and a patient-level severity score would make the method easier to compare with standard scoring and closer to clinical use.

Conclusions

A single eye-movement channel carries a modest signal of apneas and hypopneas. It adds to sleep staging and kept its accuracy in an independent cohort from another country. It is far too weak to diagnose sleep apnea or estimate its severity on its own. Because eye electrodes are already part of wearable sleep devices, it is worth testing whether this information can make home testing more accurate at no extra cost to the patient.

Abbreviations

AASM
American Academy of Sleep Medicine
AHI
apnea–hypopnea index
AUC
area under the receiver operating characteristic curve
CI
confidence interval
CinC
Computing in Cardiology
EEG

electroencephalography
EMG
electromyography
EOG
electrooculography
MCC
Matthews correlation coefficient
OSA
obstructive sleep apnea
REM
rapid eye movement
UCD
University College Dublin.

Declarations

Ethics approval and consent to participate: Not applicable. This study used only public, fully de-identified data from PhysioNet. Data collection for both datasets was approved by the original institutions, as described in the source publications [13, 15].

Consent for publication: Not applicable.

Availability of data and materials: The UCD Sleep Apnea Database is available at physionet.org/content/ucddb/1.0.0. The 2018 PhysioNet/Computing in Cardiology Challenge data are available at physionet.org/content/challenge-2018/1.0.0. The analysis code and the derived feature tables are provided as Additional file 1, and the full validation tables, including per-patient results, as Additional file 2.

Competing interests: The author declares no competing interests.

Funding: None.

Authors' contributions: SJ conceived the study, obtained and processed the data, carried out the analysis, interpreted the results and wrote the manuscript. The author read and approved the final manuscript.

Use of artificial intelligence: Claude (Anthropic) was used to help write and debug the analysis code, to edit the language of the manuscript, and to produce Figure 1. Figure 1 was generated with plotting code written with Claude's help: panel A is a simple schematic of standard electrode placement, and panel B plots unaltered recordings from the UCD database. No generative image model was used, and the figure contains no synthetic or altered data. The author checked the code and its output, reviewed Figure 1 against the underlying recordings, verified all references, and takes full responsibility for the content.

Acknowledgements: The author thanks the teams who collected the UCD and CinC 2018 datasets and made them freely available through PhysioNet.

References

1. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MSM, Morrell MJ et al (2019) Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med 7(8):687–698. 10.1016/S2213-2600(19)30198-5
2. Young T, Evans L, Finn L, Palta M (1997) Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women. Sleep 20(9):705–706. 10.1093/sleep/20.9.705
3. Steinberg R, Spector AR, McVeigh T, Fudim M (2025) Home sleep apnoea testing: advances, challenges and considerations in heart failure. Card Fail Rev 11:e29. 10.15420/cfr.2025.29
4. Kapur VK, Auckley DH, Chowdhuri S, Kuhlmann DC, Mehra R, Ramar K et al (2017) Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline. J Clin Sleep Med 13(3):479–504. 10.5664/jcsm.6506
5. Arnal PJ, Thorey V, Debellemaniere E, Ballard ME, Bou Hernandez A, Guillot A et al (2020) The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. Sleep 43(11):zsaa097. 10.1093/sleep/zsaa097
6. Massie F, Vits S, Verbraecken J, Bergmann J (2025) The evaluation of a novel single-lead biopotential device for home sleep testing. Sleep 48(4):zsae248. 10.1093/sleep/zsae248
7. Eckert DJ, Younes MK (2014) Arousal from sleep: implications for obstructive sleep apnea pathogenesis and treatment. J Appl Physiol 116(3):302–313. 10.1152/japplphysiol.00649.2013
8. Barnes LD, Lee K, Kempa-Liehr AW, Hallum LE (2022) Detection of sleep apnea from single-channel electroencephalogram (EEG) using an explainable convolutional neural network (CNN). PLoS ONE 17(9):e0272167. 10.1371/journal.pone.0272167
9. Elmi Z, Elmi S (2026) Anchor-free temporal localization of apnea events from EEG/EOG with state-space models. Physiol Meas 47(8):085013. 10.1088/1361-6579/ae8ca8
10. Sharma M, Darji J, Thakrar M, Acharya UR (2022) Automated identification of sleep disorders using wavelet-based features extracted from electrooculogram and electromyogram signals. Comput Biol Med 143:105224. 10.1016/j.compbiomed.2022.105224
11. Jain R, Ganesan RA (2025) Effective diagnosis of various sleep disorders by LEE classifier: LightGBM-EOG-EEG. IEEE J Biomed Health Inf 29(4):2581–2588. 10.1109/JBHI.2024.3524079
12. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B et al (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. 10.1136/bmj-2023-078378
13. St Vincent's University Hospital / University College Dublin Sleep Apnea Database (version 1.0.0). PhysioNet. 10.13026/C26C7D

14. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23):e215–e220. 10.1161/01.CIR.101.23.e215
15. Ghassemi MM, Moody BE, Lehman LW, Song C, Li Q, Sun H et al (2018) You snooze, you win: the PhysioNet/Computing in Cardiology Challenge 2018. Comput Cardiol 45:1–4. 10.22489/CinC.2018.049
16. Hjorth B (1970) EEG analysis based on time domain properties. Electroencephalogr Clin Neurophysiol 29(3):306–310. 10.1016/0013-4694(70)90143-4
17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
18. Ruehland WR, Rochford PD, O'Donoghue FJ, Pierce RJ, Singh P, Thornton AT (2009) The new AASM criteria for scoring hypopneas: impact on the apnea hypopnea index. Sleep 32(2):150–157. 10.1093/sleep/32.2.150

Figures

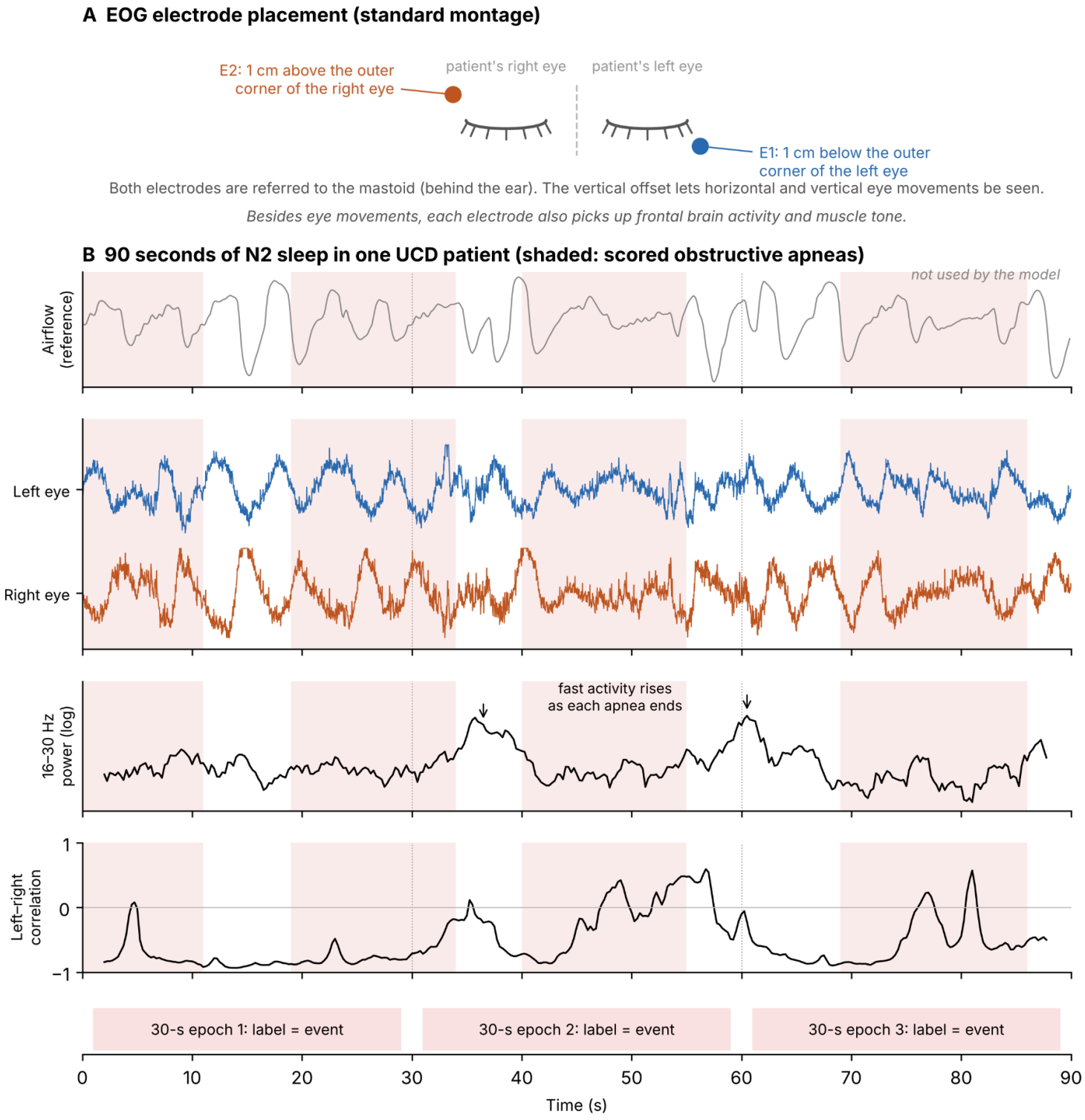

Figure 1

What the model sees. (A) Standard EOG montage: E1 sits 1 cm below the outer corner of the left eye and E2 1 cm above the outer corner of the right eye, both referred to the mastoid (the AASM-recommended montage; the UCD montage is not documented). Besides eye movements, each electrode picks up frontal brain activity and muscle tone. (B) Ninety seconds of N2 sleep in UCD patient ucddb027 with four scored obstructive apneas (shaded). From top: airflow, shown for reference only and not used by the

model; the left and right EOG signals; power in the 16–30 Hz band of the left EOG (2-second windows); and the correlation between the two eye channels (4-second windows). The bottom strip shows how the 30-second epochs were labeled. Arrows mark rises in fast activity after two of the apneas. This example is illustrative only; individual events vary, and the model's overall behavior is summarized in Figures 2 and 3.

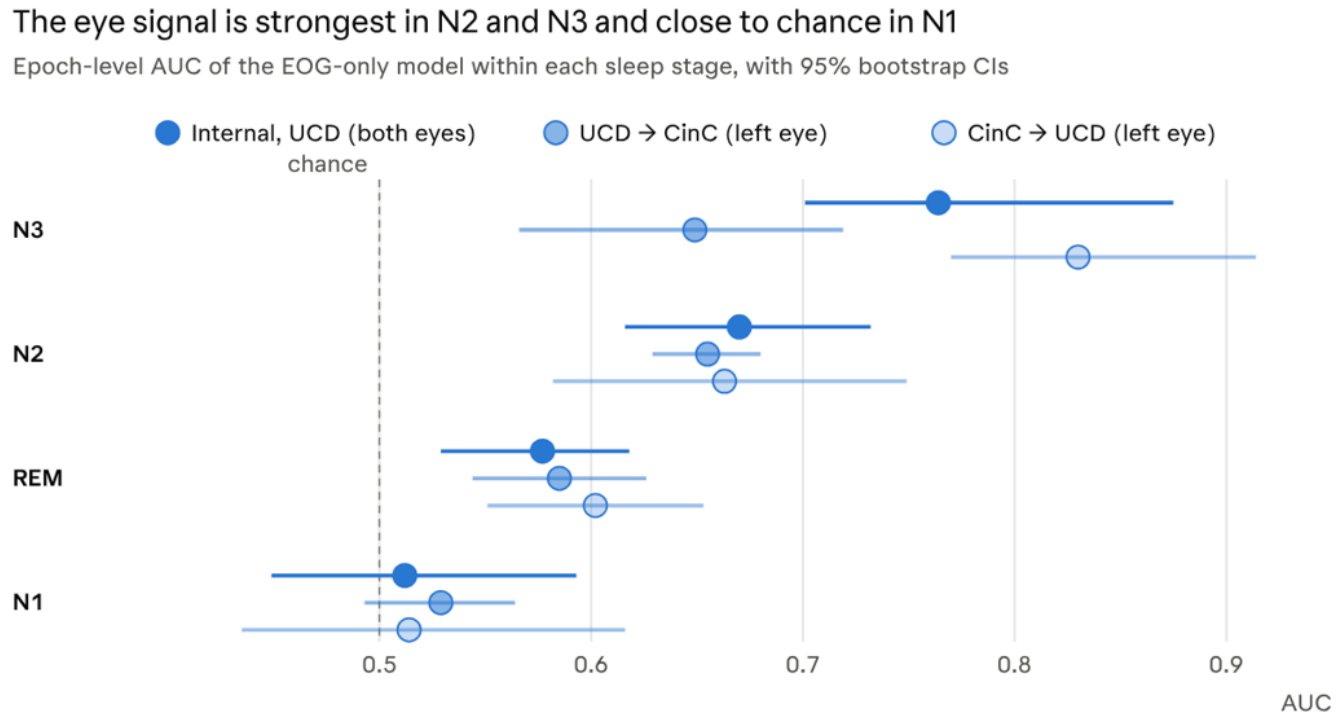

Figure 2

Epoch-level AUC of the EOG-only model within each sleep stage, with 95% bootstrap confidence intervals. Dark: internal leave-one-patient-out validation in UCD (both eyes); medium: trained on UCD, tested on CinC (left eye); light: trained on CinC, tested on UCD (left eye). The dashed line marks chance (AUC 0.5).

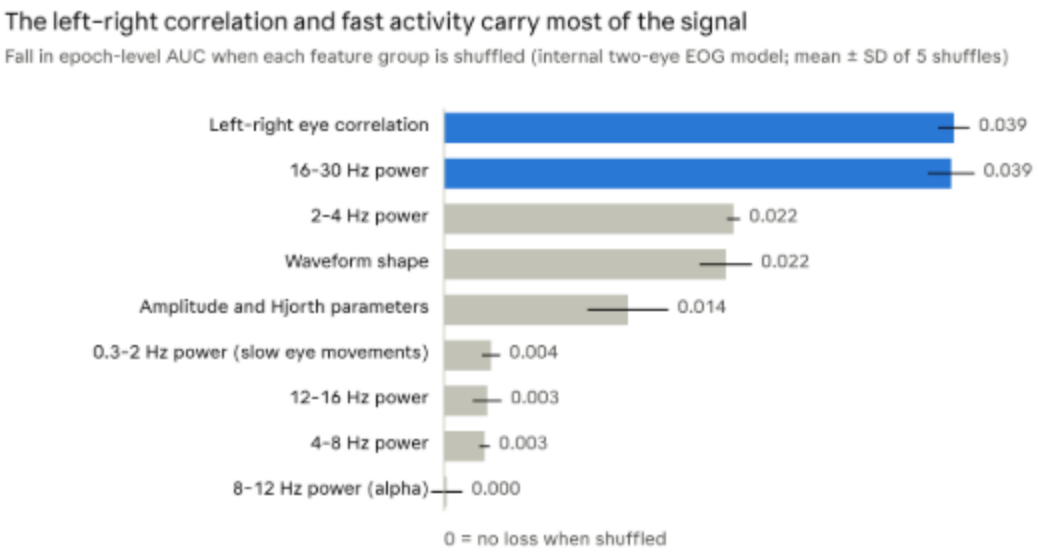

Figure 3

Grouped permutation importance in the internal two-eye EOG-only model: the fall in epoch-level AUC when all features in a group, including their temporal-context versions, were shuffled across test epochs (mean of five shuffles; base AUC 0.667). Groups that overlap with others, such as amplitude and Hjorth parameters, are understated by this method.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1code.zip
- Additionalfile2validationtables.xlsx