*medicina*

MDPI

*Review*

# Artificial Intelligence-Based Detection of Obstructive Sleep Apnea Using Acoustic Signals: Insights from Recent Studies and Future Directions

**Peizheng Wang [1], Arnab Majumdar [1] and Cheng-Yu Tsai [2,3,4,5,6,*]**

[1] Department of Civil and Environmental Engineering, Imperial College London, Exhibition Road, London SW7 2AZ, UK; pw220@ic.ac.uk (P.W.); a.majumdar@imperial.ac.uk (A.M.)
[2] Research Center of Sleep Medicine, College of Medicine, Taipei Medical University, Taipei 11031, Taiwan
[3] TMU Research Center of Artificial Intelligence in Medicine and Health, Taipei Medical University, Taipei 11031, Taiwan
[4] School of Respiratory Therapy, College of Medicine, Taipei Medical University, Taipei 11031, Taiwan
[5] Division of Pulmonary Medicine, Department of Internal Medicine, Shuang Ho Hospital, Taipei Medical University, New Taipei City 23561, Taiwan
[6] School of Biomedical Engineering, College of Biomedical Engineering, Taipei Medical University, Taipei 11031, Taiwan
* Correspondence: momo_cyt@tmu.edu.tw

## Abstract

*Background and Objectives*: Acoustic signals offer accessible inputs for artificial intelligence (AI) methods for obstructive sleep apnea (OSA) assessment. This narrative review examines how acquisition, modeling and output construction affect screening, event detection and severity estimation. *Materials and Methods*: Studies published or available online from 2022 to 29 September 2026 were identified through iterative retrieval, citation checking and supplementary PubMed and Europe PMC searches. Clinical evidence and within-study design comparisons were synthesized qualitatively, with structured appraisal of study quality and applicability. *Results*: Published meta-analyses reported approximately 90% sensitivity and 87% specificity for overnight audio and 83% for both measures for awake speech (rounded pooled estimates). Evaluation design and disease prevalence were associated with overnight-audio sensitivity; all speech studies used professional microphones in hospitals. A prospective home study testing a previously developed model in a separate cohort obtained complete recordings in 101/117 attempts and reported severe-OSA sensitivity/specificity of 93.3/94.4% on iOS. A multicenter snoring study obtained usable recordings in 121/152 recruits; a severe-OSA area under the receiver operating characteristic curve of 0.964 coexisted with apnea–hypopnea index agreement limits of −41.25 to +23 events/h. Component comparisons linked performance to features, fusion and patient-output construction, although most were internally validated. *Conclusions*: Acoustic AI may support screening and referral, but routine clinical value depends on recording completion, performance in the intended population and added benefit over accessible clinical tests. Narrative selection and heterogeneous validation preclude a ranking of clinical effectiveness.

**Keywords:** obstructive sleep apnea; artificial intelligence; acoustic signals; model design; speech; respiratory sounds; clinical validation

Check for updates

Academic Editor: Stefanie Krick

Received: 30 September 2026
Revised: 22 October 2026
Accepted: 30 October 2026
Published: 4 November 2026

**Copyright:** © 2026 by the authors. Published by MDPI on behalf of the Lithuanian University of Health Sciences. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Obstructive sleep apnea (OSA) involves recurrent upper-airway obstruction during sleep and is associated with impaired daytime function and adverse health outcomes. Clinical assessment aims to identify patients who need sleep testing and to determine the severity of their breathing disturbances. Polysomnography (PSG) provides detailed physiological measurements, but the equipment, staff, and overnight recording involved can limit access and repeated assessment [1,2]. Technically adequate home sleep apnea testing (HSAT) is an established diagnostic option for appropriate uncomplicated adults; clinical prediction tools alone do not replace PSG or HSAT [3]. Acoustic assessment has attracted interest because microphones are widely available and can record signals related to speech and breathing. Understanding what these recordings can reveal, and how reliably they can be analyzed, is important for developing practical screening and monitoring methods.

Acoustic OSA research includes speech, awake breathing sounds, simulated snoring, and spontaneous breathing sounds and snoring during sleep [1,2]. These signals are collected under different conditions and answer different questions. Speech recordings sample voice production and vocal-tract properties that may be associated with OSA risk, whereas overnight recordings can capture changes around breathing disturbances. Artificial intelligence (AI), including conventional machine learning and deep learning, is used to relate these signals to OSA labels. Conventional models use selected acoustic features; deep networks can learn features from waveforms or spectrograms. Recent studies have explored pretrained speech models, respiratory acoustic models, clinical information, and sequences of sound segments [4–8]. Some methods have also been tested prospectively during home recording [9]. These developments make it useful to examine how the choice of signal, training method, and model design affects the resulting prediction.

Cao et al. reviewed speech- and snoring-based detection in the literature through 2022 [1]; TaghiBeyglou et al. focused on speech-based severity estimation [10]; and Alqudah et al. examined OSA detection during wakefulness across several measurement approaches [2]. Recent meta-analyses quantify diagnostic performance from sleep audio and awake speech [11,12]. Pillion et al. additionally reviewed 15 audio-based studies of moderate-to-severe OSA [13]. These reviews establish the clinical rationale and summarize the available methods or diagnostic accuracy. Building on these reviews, the present synthesis connects recent model-design choices to the complete assessment process across awake and overnight acoustic tasks. It addresses three practical gaps in interpreting this literature: how explicit, task-trained, and pretrained representations are combined; how supervision and clinical information shape learning; and how local predictions become a patient-level screening decision or severity estimate. Each comparison is interpreted alongside patient separation, validation setting, and the clinical target. Relating performance to these stages helps identify whether an improvement arises from acoustic representation, learning and integration, or the construction of the patient result.

Detailed examples were selected for an explicit prediction target, an identifiable patient-validation design, and component or input comparisons linking model choices to a clinical output. They include Zhang et al.'s paired-posture speech model, Xie et al.'s snoring-feature model, Romero et al.'s sensor-fusion model, and two studies from our group on complementary respiratory features and clinically guided speech screening [4,14–17]. These examples allow representation, learning, and output construction to be examined within complete modeling pipelines.

### *Review Scope and Literature Selection*

This narrative review focuses on adult OSA assessment and studies published or available online from 1 January 2022 to 29 September 2026. Online-first articles were eligible,

including Wang et al.'s speech study assigned to a 2027 volume [17]. Earlier sources were retained for established diagnostic guidance, datasets and modeling methods. Iterative retrieval and citation checking were supplemented by PubMed and Europe PMC searches on 13 October 2026, retaining the original publication window. These database searches were a supplementary route to the literature rather than an exhaustive search of clinical and engineering databases. Selection prioritized complementary design comparisons and clinically informative validation. P.W. performed selection and extraction. Data S1 records the exact queries, eligibility criteria and retrieval history; Data S2 identifies the provenance of every appraised report.

Eligible primary reports used recorded speech, breathing, or snoring in an automated model for OSA screening, respiratory-event analysis, or severity estimation, with a defined reference and an evaluation result. Journal and conference reports were considered; preprints, retracted publications, non-acoustic-only models, self-reported snoring without acoustic input, and reports lacking a relevant target or evaluation were excluded. Pediatric OSA assessment was outside the scope; the mixed-age snore-recognition study [18] (12–81 years) was retained only as a supporting acoustic task. Age eligibility and reported ranges are recorded in Data S4; missing lower age bounds are not assumed to establish adult-only enrollment. Snore recognition and sleep staging supplied supporting evidence, while reviews, datasets, physiological studies, and methods papers provided context. The evidence tables compare 30 primary reports in 29 rows, with two analyses of a shared cohort combined in one row. Extraction covered population, inputs, reference, patient separation, validation setting, and performance level. Quantitative entries were checked against the available primary reports. The 19 additional reports not selected for detailed synthesis are summarized in Data S5. The synthesis was qualitative; pooled estimates are those of the cited meta-analyses.

A structured appraisal covered the same 30 primary reports and the prospective acoustic-device comparator. Six domains addressed participant selection and data completeness (P), predictor availability (I), reference and target definition (R), evaluation safeguards (V), precision (Q), and clinical applicability (A). Review-specific criteria informed by PROBAST + AI [19] were applied without an overall numerical score. C.-Y.T. independently assessed all six domains for the five detailed examples before consulting the original ratings; the initial agreement was 28/30 judgments (93.3%). The two differing validation judgments and their supporting reasons are reported in Data S2 and incorporated into the appraisal. Data S2 provides the supporting observations for the domain judgments. The predictor judgment for the speech study [17] concerns pre-sleep screening; its post-sleep recording condition is not included in the headline screening results.

## 2. Acoustic AI for OSA: Tasks and Analytical Framework

### *2.1. Prediction Tasks*

Patient screening combines acoustic observations into a risk group or referral decision. Local respiratory-event detection instead labels windows or locates event boundaries. Severity estimation predicts the apnea–hypopnea index (AHI), defined as scored apneas and hypopneas per hour of sleep, or an AHI category. A severity estimate may be obtained by counting predicted events or by direct regression from recording summaries. Supporting tasks, including snore recognition and sleep staging, supply intermediate information [18,20]. This distinction determines the evaluation unit: a sound, window, event, recording night, or patient. For example, speech clips inherit the patient's OSA label, while a localized respiratory event has an onset and an offset within a night. Evidence at one level does not automatically extend to another.

### *2.2. From Acoustic Signals to Patient-Level Outputs*

Figure 1 summarizes the pathway followed in this review: acquisition and representation, learning and integration, output construction, and validation. Acquisition determines which signal is recorded and under what conditions. Explicit descriptors, task-trained features, and pretrained encoders retain different information from that signal. Learning objectives and fusion then relate the information to a prediction target, using clinical measurements where available. Finally, clip probabilities may be aggregated for screening, window predictions merged and counted, or recording features regressed directly to severity [7,16,17,21]. Evaluation concerns this complete procedure, including patient separation, recording failures, and performance in new settings. Relating results to these stages helps identify whether a gain comes from representation, learning, or the construction of the clinical output. The following sections use this structure to compare reported findings; untested extensions are collected in Section 9.3.

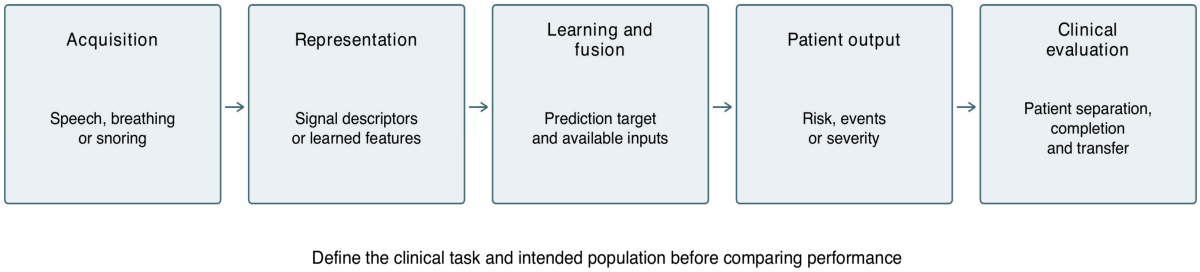



**Figure 1.** General pathway from an acoustic recording to clinical assessment. Each stage can change the information available or the meaning of the final result. Screening, event detection and severity estimation use different outputs within this common structure.

## 3. Clinical Applications

### *3.1. Screening and Referral Prioritization*

Acoustic screening is most useful when it improves referral decisions at an acceptable recording burden. Brief clinical risk instruments, including STOP-Bang and NoSAS, provide relevant comparators, while models using the same clinical variables without audio isolate the contribution of sound. A score for severe OSA does not exclude milder disease. The clinical question is therefore which additional patients reach diagnostic testing, how many unnecessary referrals are generated, and whether recording failures delay care. Table 1 brings together studies with prospective or separate-cohort evaluation and a direct home comparison with a clinical questionnaire.

Published meta-analyses provide a quantitative clinical context. Tan et al. combined 16 overnight-audio studies and estimated a sensitivity of 90.3% (95% credible interval, 86.9–93.1%) and a specificity of 86.7% (83.1–89.7%) [11]. At AHI thresholds of 5, 15, and 30 events/h, the sensitivity/specificity were 94.3/78.5%, 86.3/87.3%, and 86.3/89.5%, respectively. Gao et al. combined eight awake-speech studies, estimating a sensitivity of 82.9% (80.0–86.4%) and a specificity of 83.3% (80.7–86.1%) [12]. All eight used professional microphones in controlled hospital settings. These are estimates from separate reviews with different studies and targets, rather than a direct comparison of overnight audio with awake speech. In Tan et al.'s meta-regression, train–test split evaluation, higher OSA prevalence, higher sampling frequency and non-contact microphones were associated with higher sensitivity [11]. These study-level associations show why the pooled estimate cannot be transferred unchanged to an unselected screening population; they do not establish that any one design choice caused the difference.

**Table 1.** Clinical evaluation beyond model-development comparisons. Completion denominators refer to the reported recruitment or recording process, not a common failure definition.

| Study | Evaluation | Population and Completion | Comparator and Patient Result | Interpretation |
|---|---|---|---|---|
| Han et al. [9] | Prospective home; fixed model; separate cohort | 101/117 complete (86.3%); sleeping alone; 50.5% female | Level II PSG; severe-OSA: iOS sensitivity/specificity 93.3/94.4% | Developer-affiliated investigators; tested phone protocols |
| Wang et al., OSAnet [22] | Same-center temporal test after development | 59 subsequently recruited test patients | PSG; severe-OSA sensitivity/specificity 95.6/91.6% | External-center and home transfer not tested |
| Guan et al. [23] | Prospective fixed-model test; three hospitals | 121/152 usable snoring recordings (79.6%) | PSG and wearable; severe-OSA AUROC 0.964; AHI limits −41.25 to +23 events/h | 134 complete clinical records; modality-specific available cases |
| Xu et al. [24] | Separate external cohort after PSG-Audio development | 60 external participants; selected recording quality | PSG; AHI MAE 6.03 events/h | Laboratory evaluation and quality selection |
| Stephens et al. [25] † | Home app evaluation; hypertensive adults | 180 analyzed; OSA 91.7%; moderate-to-severe 65.5% | HSAT and STOP-Bang; AUROC 0.831 at AHI ≥ 15 | Cutoff selected in evaluation cohort |
| Sanchez Gomez et al. [26] | Prospective device comparison with PSG | 63/80 paired results (78.8%) | PSG; sensitivity/specificity 92.86/97.14% at AHI ≥ 15 | Device comparator; 5 non-diagnostic, 3 not started, 9 reference exclusions |

† Abstract or publisher-preview assessment. AHI thresholds are in events/h. A separate cohort does not necessarily imply investigators independent of the model developer. MAE, mean absolute error.

Stephens et al. compared a sleep-sound app with STOP-Bang in 180 hypertensive adults using home sleep apnea testing (HSAT) as a reference [25]. For AHI ≥15 events/h, the app's AUROC was 0.831 (95% confidence interval, 0.773–0.890), versus 0.728 and 0.754 for the original and modified questionnaire. The sensitivity/specificity were 72.0/85.5% at an app cutoff of 7.1 selected in the same cohort. OSA prevalence was 91.7%, with moderate-to-severe disease in 65.5%; these results describe a high-prevalence hypertensive population and do not establish the performance of a prespecified rule in general screening. The pooled comparison with questionnaires in Tan et al. instead used separate published evidence bases [11].

### *3.2. Overnight Assessment and Monitoring*

Overnight recording can be used to detect when respiratory events occur, estimate their frequency, or classify OSA severity. Its feasibility at home depends on capturing the correct patient, maintaining recording, and obtaining sufficient usable audio, as well as on the model's performance. Romero et al. evaluated smartphone recordings from 103 participants over 157 nights, using participant-separated cross-validation and simultaneous level III testing [21]. Han et al. evaluated a previously developed model prospectively against level II home PSG, with complete smartphone and PSG data in 101 participants [9]. Together, they demonstrate home acquisition under both internal evaluation and prospective testing of a

predeveloped model. Non-acoustic multinight research shows that severity classification can vary across nights [27], motivating repeated acoustic assessment. Such assessment must distinguish changes in breathing from changes in recording conditions. Monitoring treatment response additionally requires evidence that acoustic change tracks the relevant clinical outcome.

## 4. Acoustic Signal Types and Data Acquisition

### *4.1. Speech Signals*

Because awake speech depends on the vocal source, articulation, and vocal-tract acoustics, its association with OSA may reflect anatomy or function, shared anthropometry, or differences in the recording task. Phonation does not directly indicate recurrent airway collapse during sleep, and classification performance cannot identify the underlying mechanism [2,10]. Saha et al. measured upper-airway dimensions with ultrasound during vowel production in 31 adults [28]. Acoustic features were associated with measured dimensions, and dimensional changes differed between NoSAS risk groups. As the grouping used a risk score rather than PSG diagnosis, the study supports testing acoustic–anatomical relationships rather than diagnostic accuracy.

Sustained vowels restrict articulatory variation, whereas standardized reading samples transitions across sounds. Natural speech offers convenient acquisition but varies in words, language, and speaking style. Zhang et al. used sentences in 254 Chinese men with selected clinical variables [4]. Wang et al. recorded a standardized Mandarin passage before and after overnight PSG in the same 112 participants [17]. Standardized content reduces linguistic variation, and the paired design allows recording conditions to be compared within individuals. The pre-/post-sleep results describe this protocol; their physiological explanation and relevance to unscheduled daytime screening remain open.

Population coverage also differs: Han et al.'s home cohort was 50.5% female [9], whereas the two Wang cohorts were predominantly male [16,17]; Zhang et al. and Xue et al. enrolled only men [4,29]. Language and sex coverage therefore need to be considered alongside the choice of speech representation.

TaghiBeyglou et al. studied 151 participants from shelters, a sleep clinic, and a cohort including people after stroke; sex-specific analyses at AHI 30 reported selected word/model balanced accuracies of 0.74 in women and 0.78 in men [8]. The two Wang cohorts were predominantly male: 49/73 in the respiratory-sound study and 83/112 in the speech study [16,17]. Their published evaluations do not provide sex-stratified performance. Recruitment diversity and subgroup results therefore provide information beyond overall accuracy, particularly for transfer to populations unlike the development cohort.

The recording task can also be designed to emphasize particular articulatory movements. Xue et al. collected controlled Mandarin articulations from 301 men and combined explicit paralinguistic features with transfer-learned spectrogram features in PTF-Net [29]. Here, domain knowledge enters through the sounds requested from a patient as well as through the network, extending the design choices beyond standard reading and frozen speech representations. Their reported analysis against drug-induced sleep endoscopy (DISE) supplies an anatomical reference for that particular investigation. Because the selected features depend partly on phonetic content, transfer across languages involves both the encoder and the articulations sampled by the recording task.

### *4.2. Awake Breathing Sounds and Simulated Snoring*

For awake tracheal breathing, which requires a different recording procedure from speech, Alqudah and Moussavi collected supine contact recordings at the suprasternal notch during prescribed nasal and mouth breathing in 199 participants [30]. Their subse-

quent meta-model reused the cohort [31]. Sensor placement and breathing instructions provide control but impose different requirements from smartphone reading. Simulated snoring requires another recording protocol; Fang et al. used instructed sounds with controlled posture and endoscopic assessment [32]. The acoustic information and feasibility therefore depend on a supervised maneuver, rather than ordinary speech or spontaneous sleep snoring.

### *4.3. Nocturnal Breathing Sounds and Snoring*

Snore detection identifies an acoustic event, whereas detection of PSG-defined apneas and hypopneas may depend on pauses, reduced breathing sounds, recovery sounds, and their context. Continuous overnight audio additionally contains quiet breathing, movement, environmental noise, and wakefulness. Selecting individual snores therefore provides different information from analyzing the whole recording. A smartphone bench study used playback files derived from 18 participants to test snore detection [33]. Its reference was known snore occurrence in constructed audio, giving a technical measure of sound detection. Xie et al. used temporal variability and energy-related snoring features, with age and body mass index (BMI), to estimate severity in 172 participants [14]. This approach illustrates why the temporal organization of sounds can carry information beyond the spectrum of an isolated snore. Conversely, the hybrid respiratory-sound study analyzed overlapping 20 s windows for local respiratory-event classification and then merged positive windows for night-level assessment [16]. One approach summarizes patterns to predict a patient-level target; the other builds a patient output from local event-related decisions.

Li et al. combined waveform and visibility-graph image branches for snore recognition in home recordings from 88 participants, with subject-separated evaluation [18]. Their output identifies snoring rather than respiratory obstruction. In a separate 124-participant study, Li et al. used visibility-graph convolutional neural network (CNN) predictions from individual snores to construct a patient screening index, reserving 40 participants for testing [34]. The additional aggregation and clinical reference change the task. A snore detector can make analysis of long recordings more efficient, but excluding everything it calls non-snoring may also remove pauses, quiet breathing, and recovery context. The usefulness of snore selection therefore depends on the information needed for the final OSA prediction, especially in recordings with few detected snores.

### *4.4. Data Acquisition and Preprocessing*

Tan et al.'s meta-regression found no detected accuracy difference between home-smartphone and laboratory-microphone studies, while higher sampling frequency and non-contact microphones were associated with higher sensitivity [11]. These study-level comparisons support the feasibility of home audio; paired device studies are still needed to isolate a change in hardware within the same model. The two detailed studies use different recording and preprocessing protocols. The hybrid respiratory-sound study collected synchronized 16-kHz, 16-bit smartphone audio during PSG in 73 patients at a hospital sleep laboratory, without dedicated sound attenuation [16]. Eligibility required more than six hours of PSG recording and sleep efficiency above 40%. Handcrafted features were extracted from the original recordings before the preprocessing used for the learned branch. Gaussian noise and amplitude scaling between 0.9 and 1.1 were used for augmentation. The resulting model combines original-audio descriptors with features learned from a separately processed input. Acoustic assessment is not uniformly contactless: bedside recording, prescribed vocal tasks, and contact sensors impose different patient requirements. A difference in smartphone placement and inputs is illustrated by Frija et al.'s Apneal proof of concept, which attached a phone to the chest and recorded sound and motion [35]. Its

first automatic prototype used a heart-rate time series and an acceleration envelope, whereas audio was among the signals available to manual scorers. Its automatic results therefore concern heart-rate and motion inputs. The distinction matters when comparing automatic acoustic methods with multisensor acquisition systems.

Recording conditions shape the population and signal that a model can assess. Wang et al.'s speech study retained 112 of 115 recruited patients with paired recordings and complete PSG information [17]. Its controlled Mandarin reading task differs from a brief phone recording in routine care. Acquisition and processing details for this and other frameworks are collected in Data S3.

The effect of preprocessing on the final prediction was examined by Cho et al., who compared four combinations of denoising and feature selection in 423 laboratory smartphone recordings; the best AHI regression used selected features without denoising [36]. For AHI regression, denoising offered no advantage over selected features alone. A noisy-clinic tracheal-sound study illustrates the interaction of acquisition and selection: 247 people were recruited, 228 remained after recording-quality exclusions, and 196 entered the main analysis after excluding AHI 10–15 [37]. Nested patient-level cross-validation gave sensitivity/specificity values of 84.3/59.9% for AHI >15 versus <10. The practical trade-off is visible in both the false-positive rate and the patients left outside the main evaluation.

In Wang et al.'s respiratory study, descriptors from the original audio were combined with a learned representation of a smoothed log-RMS amplitude envelope [16]. The envelope represents slow changes in sound level, while the separately extracted descriptors retain acoustic spectral information. Removing the reported filtering and resampling step reduced the window AUROC from 0.919 to 0.909, supporting its use within the tested pipeline. The signal path is clarified in Data S3.

Noise augmentation and denoising address different aspects of acquisition. Le et al. added recorded home noise to laboratory audio [38], whereas Cho et al. compared denoising and feature selection [36]. These experiments assess the model's response to specific transformations; home studies additionally measure performance under naturally varying noise and incomplete recording.

### *4.5. Reference Labels and Acoustic Observations*

Reference labels specify the prediction target. Speech clips inherit a patient severity label, whereas respiratory windows are assigned local event labels; neither is equivalent to a precisely localized event boundary [16,17]. PSG scoring rules, included event types, window overlap, and audio synchronization therefore affect performance. Saha et al. examined central-versus-obstructive event classification in 50 selected participants [39], while Han, Xia, and Mascolo compared binary, three-class, and five-class event tasks in 194 PSG-Audio participants [40]. These studies address different label distinctions. Song et al.'s label-correction method further illustrates the importance of event–window alignment during development [41]. Evaluation against an unchanged physiological reference preserves the clinical meaning of the resulting scores.

## 5. Examples of Complete Modeling Frameworks

### *5.1. Feature-Based Severity Estimation and Sensor Fusion*

Xie et al. modeled severity using snoring timing and energy and then added age, BMI and further acoustic descriptors [14]. Adding clinical variables increased the correlation with AHI from 0.676 to 0.731 but reduced the four-category accuracy from 54.7% to 51.7%; extended features achieved a correlation of 0.786 and an accuracy of 59.3%. Continuous prediction and decisions at severity boundaries thus responded differently to feature additions. Romero et al. compared sound and respiratory-effort inputs in 103 participants

over 157 home nights [15,21]. For severe OSA, audio alone yielded a sensitivity/specificity of 0.78/0.93, versus 0.93/0.92 with decision fusion. Repeated nights remained in the same patient fold, but patient-clustered precision for these night-level estimates was not reported.

### *5.2. Paired Speech Learning and Clinical Guidance*

Zhang et al. used paired sitting and lying speech, a Siamese learning design, clinical characteristics, and auxiliary AHI regression in 254 men; the reported severe-OSA F1 was 0.725 on 75 test patients [4]. Wang et al.'s speech study from our group used standardized Mandarin reading in 112 patients, a frozen XLS-R speech encoder [42], alignment with clinical profiles, severity-supervised contrastive learning, and clinical modulation [17]. It examined alternative fusion rules and training objectives, then combined clip probabilities into a patient result. These studies use different supervision to make speech informative for patient assessment: paired posture and continuous severity in one, and patient–clinical correspondence and severity groups in the other. Section 6.4 compares their components alongside independent work.

### *5.3. Hybrid Respiratory Features*

Wang et al.'s respiratory study combined original-audio descriptors with features learned from a smoothed log-RMS envelope in 73 laboratory participants [16]. Auxiliary sleep-stage supervision and clinical information contributed to window classification; positive windows were merged into a recording-time index. This combines different transformations of one recording, whereas Romero et al. combined sound with a separately measured physiological signal [15]. Table 2 compares the five detailed examples using common criteria; the implementation details are in Data S3.

**Table 2.** Detailed examples selected for a defined clinical target, patient-validation design, and component or input comparisons. Results identify the evaluation unit and endpoint.

| Study and Inputs | Design Comparison | Output and Validation | Reported Result |
| --- | --- | --- | --- |
| Xie et al. [14]; snoring, age/BMI | Baseline and extended explicit features; addition of clinical variables | AHI regression and four-category patient severity; internal five-fold evaluation | Patient AHI $r$: 0.676 to 0.731 after clinical inputs; four-class accuracy: 54.7% to 51.7%. |
| Zhang et al. [4]; sitting/lying speech | Paired-posture learning, clinical characteristics, auxiliary AHI regression | Severe-OSA patient classification; 149/30/75 patient split | Patient F1 0.725 with neck circumference and auxiliary regression. |
| Romero et al. [15]; sound and effort | Feature, latent, and decision fusion | Severe OSA per night; 103 patients/157 home nights; patient-separated folds | Night sensitivity/specificity 0.93/0.92 with decision fusion; audio sensitivity 0.78. |
| Wang et al. [16]; authors' respiratory study | Original-audio descriptors and learned log-RMS envelope features; auxiliary supervision and clinical fusion | Window classification: 15 held-out patients; patient severity: separate ten-fold cross-validation | Window AUROC 0.919 (reported SD 0.008); 0.864 without latent features. |
| Wang et al. [17]; authors' speech study | Frozen encoders; clinical modulation; contrastive objectives; aggregation | Severe-OSA patient output; ten patient-separated folds; 112 patients | Pre-sleep segment AUROC 0.8345; patient accuracy 85.71% (95% CI 81.36–90.18%). |

## 6. Artificial Intelligence Methods

### *6.1. Acoustic Features and Representation Learning*

Handcrafted descriptors summarize named signal properties, task-trained networks learn from development recordings, and pretrained encoders transfer information from larger datasets. Skapova et al.'s supervised i-vector classifiers exceeded an exploratory frozen ECAPA-TDNN pipeline in 93 participants [43]. Xue et al. combined paralinguistic and spectrogram features from controlled articulation [29]. These approaches retain different information and have different data requirements. Pooled comparisons found no clear diagnostic advantage of deep over conventional machine learning but did not isolate architecture from acquisition, patient selection and evaluation [11,12].

Paired channels provide a more direct test of representation transfer. Han, Xia, and Mascolo compared filter-bank and HuBERT features in ambient and tracheal recordings from 194 participants [40]. With the same EBranchformer architecture for three-class event recognition, HuBERT increased ambient-sound F1 from 0.694 to 0.714 but decreased tracheal-sound F1 from 0.760 to 0.746. The opposite changes show that the usefulness of speech pretraining depends on the input channel. OPERA assembled approximately 136,000 respiratory recordings and evaluated 19 downstream tasks; contrastive pretraining more often favored classification, whereas generative pretraining favored regression [5]. ApneaWhisper provides another approach by adapting a pretrained Whisper encoder to sleep-audio segmentation [44]. These results make the pretraining domain and objective relevant alongside network size.

### *6.2. Clinical Variables and Fusion*

Clinical-only, acoustic-only and combined models answer different questions when evaluated on the same patients. Xie et al.'s improvement in AHI correlation after adding age and BMI, despite lower categorical accuracy, illustrates this endpoint dependence [14]. In Wang et al.'s internally cross-validated speech cohort, adding speech to seven clinical variables was associated with higher patient accuracy: 96/112 (85.71%) versus 81/112 (72.32%) [17]. Severe OSA affected 48/112 patients (42.86%), and classification used a fixed probability threshold of 0.5. The accuracy difference was not formally tested, and the clinical-only logistic regression was not a validated questionnaire. These results describe an internal comparison rather than established incremental screening utility.

Encoding the same clinical measurements in a different form changes their representation without adding a measurement. Wang et al. compared numerical and ClinicalBERT-based text encodings of seven variables [17,45], and Li et al.'s DFNet combined acoustic and medical-language representations [6]. Qiu et al. instead derived semantic representations from acoustic features and included event-record features [46]. Feature-wise linear modulation (FiLM) provides a general mechanism for conditioning learned features on another input [47]. For clinical use, the distinction is whether each input is available when the prediction is made. The encoding and fusion comparisons are summarized in Data S3.

Fusion at the decision stage offers an alternative to interactions within a shared network. TaghiBeyglou et al. combined predictions from self-supervised, handcrafted, and time–frequency speech representations [8]. Romero et al. compared feature, latent, and decision fusion of sound and abdominal effort during home recording; sum-based decision fusion increased severe-OSA sensitivity from 0.78 for audio alone to 0.93, with a specificity of 0.92 [15]. Fixed patient characteristics provide context, whereas a second physiological signal supplies time-varying information. Their costs and failure modes also differ: approximately 7% of Romero et al.'s recordings lost the effort signal.

### *6.3. Supervision and Learning Objectives*

Auxiliary targets and contrastive pairs encourage a model to retain different relationships. Zhang et al. combined paired-posture speech with continuous AHI regression [4]. Wang et al. aligned speech with clinical profiles and severity groups [17,48], while auxiliary REM/non-REM supervision shaped the envelope branch of the respiratory model [16]. The latter supported event classification within that pipeline and did not provide a separately validated sleep–wake output. Detailed losses and ablations are reported in Data S3.

Lillini et al.'s SiCRNN learned paired representations of tracheal-sound spectrograms using a Siamese CNN and bidirectional recurrent network, followed by clustering [49]. Its 37-patient dataset was evaluated with three patient-separated splits, each containing 23 training, six validation, and eight test patients. The mean six-second-window F1 was 0.90 (SD 0.008). A positive window required an apnea throughout its duration, so this result concerns window recognition rather than event-boundary recovery. The identity of paired observations and their reference labels is consequently more informative than the architecture name alone.

Imbalance also depends on how observations are sampled. Repeated or overlapping windows can increase exposure to a small number of events without increasing patient diversity. The reviewed models use focal weighting, label smoothing, or contrastive objectives to address aspects of learning, but these choices do not correct inaccurate event timing. AUROC and precision–recall performance describe complementary properties; the latter also depends on prevalence [50]. Their interpretation is strongest when the evaluation preserves the intended patient population and event distribution.

### *6.4. Component Comparisons and Computational Cost*

Table 3 summarizes selected within-study comparisons, with the evaluation unit stated for each result. Ablations estimate the effect of a component within a retrained pipeline; their differences are not additive shares of model performance. Xie et al.'s endpoint-dependent results and the opposite channel-specific changes reported by Han et al. show why a gain at one output need not extend to another [14,40]. Data S3 compares the implementation details.

**Table 3.** Selected within-study comparisons. Evaluation units are stated explicitly. These comparisons concern different datasets and do not constitute a cross-study ranking.

| Comparison | Reported Result and Unit | Interpretation |
| --- | --- | --- |
| Xie et al.: clinical inputs [14] | Patient AHI *r*: 0.676 to 0.731; four-class accuracy: 54.7% to 51.7%. | A gain in correlation did not improve the categorical endpoint. |
| Han et al.: HuBERT versus filter banks [40] | Three-class window F1: ambient 0.694 to 0.714; tracheal 0.760 to 0.746. | The same pretrained representation had opposite effects across channels. |
| Romero et al.: sound plus effort [15] | Night-level severe-OSA sensitivity: audio 0.78; decision fusion 0.93 (specificity 0.92). | A second physiological input improved sensitivity within this home cohort. |
| Xu et al.: energy and recurrent components [24] | Frame macro-F1: full 0.64; without energy 0.62; without BiLSTM 0.36. | Energy and temporal modeling contributed differently to local recognition. |
| Wang et al.: respiratory features [16] | Window AUROC: full 0.919; no latent features 0.864; latent only 0.855. | Original-audio descriptors and a learned log-RMS envelope representation contributed within the reported pipeline. |
| Wang et al.: speech aggregation [17] | Pre-sleep patient accuracy: voting 92/112 (82.14%); aggregation 96/112 (85.71%). | Four additional correct classifications; the paired difference was not formally tested. |

Respiratory window AUROC was measured in 15 held-out patients; SD 0.008 describes variability across runs, not a patient-level confidence interval [16]. The speech aggregation difference corresponds to four patients and has no reported paired test or confidence interval for the difference [17].

Practical computation includes feature extraction and preprocessing as well as model execution. A frozen encoder still incurs inference costs, and timing on laboratory hardware does not establish phone performance. Data S3 reports the available model sizes and hardware-specific timings; recording duration and repeated attempts remain part of the patient's total burden.

## 7. Patient-Level Prediction and Severity Estimation

### *7.1. Aggregation of Local Predictions*

Patient aggregation can change the final decision even when segment predictions are unchanged. In Wang et al.'s pre-sleep speech analysis, voting classified 92/112 patients correctly (82.14%), and statistical aggregation classified 96/112 (85.71%; reported 95% patient-clustered bootstrap CI, 81.36–90.18%) [17]. This four-patient difference was not accompanied by a paired significance test or an interval for the difference. Both the probability summaries and decision rule changed, with a fixed threshold of 0.5; the comparison was with voting rather than a trained mean-score model. The study author clarified that, within each patient-wise fold, the aggregator was fitted only on summary features from training patients and then applied unchanged to held-out test patients. Generating its training inputs from internal out-of-fold predictions was not specified.

For sleep sounds, output formation may merge positive windows, localize individual events, or regress severity directly. Wang et al. merged positive 20 s windows with 10 s overlap, and Romero et al. similarly grouped positive windows [16,21]. Merging determines whether repeated detections become one event and whether adjacent events remain distinct. Song et al.'s later work used an ensemble to revise automatically aligned window labels before event learning and severity regression [41]. This addresses training-label uncertainty; the evaluation reference remains a separate part of the design.

OSAnet directly localized events in spectrograms and reported an event precision of 0.81 and a sensitivity of 0.78 in 59 subsequently recruited test patients [22]. Xu et al.'s dual-stream CRNN evaluated frame classification, event boundaries, and patient AHI, with an external mean absolute error (MAE) of 6.03 events/h in 60 patients [24]. Their explicit timing outputs make event splitting, merging, and boundary errors observable. These errors can cancel in a final count, which explains why agreement at the patient level alone gives an incomplete account of event detection.

### *7.2. Event Counts and Time Normalization*

PSG AHI divides the scored event count by sleep time, whereas some acoustic systems normalize predicted events by recording duration:

$$\hat{I}_{\mathrm{rec}} = \frac{\hat{N}_{\mathrm{acoustic}}}{T_{\mathrm{recording}}}, \ \mathrm{AHI}_{\mathrm{PSG}} = \frac{N_{\mathrm{PSG}}}{T_{\mathrm{sleep}}}. \tag{1}$$

For an acoustic counter with positive durations $T$ and $T_{\mathrm{sleep}}$, the following identity separates a count term from a denominator term:

$$\frac{\hat{N}}{T} - \frac{N}{T_{\mathrm{sleep}}} = \frac{\hat{N} - N}{T_{\mathrm{sleep}}} + \hat{N}\left(\frac{1}{T} - \frac{1}{T_{\mathrm{sleep}}}\right). \tag{2}$$

Here, $N$ denotes the reference event count, $\hat{N}$ the acoustic count, and $T$ the operational duration. All quantities refer to the same observation interval, with durations in hours. Equation (2) first evaluates the count error using reference sleep time and then changes the denominator while holding the acoustic count fixed. Overcounting and a denominator longer than the sleep time can partly offset one another.

The recording-time acoustic index is analogous in its time normalization to the respiratory event index (REI) used in HSAT. It is denoted $\hat{I}_{\mathrm{rec}}$ here because its acoustic event definition and recording duration are not necessarily equivalent to the respiratory measurements and monitoring time used by a clinical HSAT. Wang et al. and Romero et al. evaluated recording-time indices against reference severity [16,21]. Pitkänen et al.'s analysis of 1561 PSG records showed that both the time denominator and the inclusion of arousal-associated hypopneas affect agreement with the AHI [51]. Substitution of reference sleep time can identify the denominator contribution in an acoustic counter. A direct regression against the PSG AHI instead estimates severity without an observable event count to decompose.

For Wang et al.'s recording-time index, a patient-level MAE or Bland–Altman comparison with the PSG AHI was not reported [16]. Its ten-fold severe-OSA classification result therefore does not establish agreement of the continuous index with the AHI.

### *7.3. Sleep-Time Estimation and Joint Outputs*

Acoustic sleep staging provides a possible sleep-time estimate, with its own errors. HomeSleepNet achieved 76.2% three-state epoch accuracy on 45 paired home PSG/audio recordings [20]. Xie et al. jointly modeled respiratory events and sleep–wake state using cardiac and audio signals from 161 participants [52]. Cross-validation yielded an event F1 of 0.588, AHI correlation of 0.825, and four-category severity accuracy of 57.8%. The model estimates both events and sleep time, coupling the numerator and denominator of the final index. Sleep-duration bias and the timing of misclassified wake periods can both affect which events enter the final index.

## 8. Model Validation and Clinical Evaluation

### *8.1. Internal and External Validation*

Patient separation is central because clips, overlapping windows, repeated conditions, and nights share patient identity and recording characteristics. Splitting these observations across training and testing can produce optimistic estimates for new patients. Allocation before segmentation and augmentation, with feature selection and aggregation training confined to development patients, preserves the clinical unit of evaluation [19,53]. Song et al. reported a segment accuracy of 88.32% with a subject-dependent split and 68.55% with a subject-independent split; the F1 was 86.28% and 60.29%, respectively [7]. The split is therefore essential context for the result. Li et al. reported a similar difference in 32 participants: the four-class snore-event accuracy fell from 83.52% with subject-dependent evaluation to 62.31% with subject-independent evaluation; the embedded implementation achieved 58.85% under leave-one-subject-out testing [54]. Evaluating the deployed model and unseen patients together reveals a different performance level from a development result alone.

Internal validation uses held-out patients or cross-validation within a development cohort. External validation evaluates a fixed model in a separately recruited cohort without using its outcomes for tuning. A later same-center cohort provides temporal validation; a new center examines additional transfer. Multicenter recruitment remains internal when patients from all centers contribute to development. Dataset-specific retraining on public and private data evaluates the method in two datasets, rather than external transfer of one fixed model. Li et al.'s adaptation using unlabeled new-patient recordings is a further setting in which performance depends on the target data available during adaptation [55].

Tan et al.'s meta-regression associated train–test splitting with higher sensitivity than other evaluation designs and higher disease prevalence with higher sensitivity [11]. These associations do not show that internal validation uniformly inflates accuracy, but they

reinforce the need to compare evaluation designs and disease spectra before interpreting numerical differences.

Wang et al.'s speech study used ten patient-separated folds, with eight for training, one for validation, and one for testing [17]. The respiratory study used two evaluation protocols: an 80%/20% patient split with 15 of 73 patients held out for window-level testing (AUROC 0.919), and a separate ten-fold evaluation of patient-level severity, training on nine folds and testing on the remaining fold [16]. The latter yielded a mean accuracy of 89.0% (SD 1.1 percentage points) at AHI 30. A separate evaluation using dataset-specific training on a 305-participant PSG-Audio subset yielded a window AUROC of 0.8065 and AUPRC of 0.7459. The 73-person respiratory-sound cohort and the 112-person speech cohort had no participants in common. Other reuse is explicit: the two awake-tracheal studies used the same 199 participants, and Romero et al.'s sound and sensor-fusion reports used the same 103-participant home cohort [15,21,30,31].

Tables 4 and 5 pair numerical results with their evaluation unit, reference threshold, and validation design. OSAnet tested 59 subsequently recruited patients at the same center [22]; Xu et al. used a separate external cohort of 60 [24]. Independent recruitment, prospective enrollment, and unattended home acquisition address different aspects of clinical applicability.

**Table 4.** Selected acoustic OSA studies: populations, inputs, validation, and patient-level or recording-night results. Performance values are presented with their reported endpoints and validation settings.

| Study | Population and Setting | Acoustic Input and Target | Validation and Separation | Performance: Unit and Threshold |
|---|---|---|---|---|
| Zhang et al., 2023 [4] | 254 Chinese men; laboratory | Sitting/lying sentence speech plus clinical characteristics; severe OSA | Internal patient split: 149 training, 30 validation, 75 testing | Patient F1 0.725 with neck circumference and auxiliary AHI regression. |
| Wang et al., online 2026; volume 2027 [17] | 112 patients (48 severe OSA; 42.86%); controlled Mandarin reading before sleep | Mandarin reading plus seven clinical variables; severe OSA | Internal patient-separated ten-fold evaluation; paired recordings remain one cohort | Severe-OSA patient accuracy 85.71% (96/112); clinical-only 72.32% (81/112); fixed threshold 0.5; paired difference not formally tested. |
| TaghiBeyglou et al., 2026 [8] | 151 participants across three settings | Prescribed speech; self-supervised and other representations; severity | Internal sex-specific cross-validation and dataset analyses | At AHI 30: sex-specific selected word/model balanced accuracy 0.74 in women and 0.78 in men. |
| Alqudah and Moussavi, 2025/2026 [30,31] | Same 199 participants; supine awake recordings | Contact tracheal breathing sounds plus anthropometry; severity | Internal holdout: 169 development/30 test participants; cohort reused | 2026 four-class meta-model test accuracy 76.7%; earlier pairwise results concern different tasks. |
| Cho et al., 2022 [36] | 423 laboratory participants; smartphone | Sleep breathing sounds; severity classification and AHI regression | Internal holdout: 256 training/167 test patients | At AHI 30: patient accuracy 81.7%, AUROC 0.90; best AHI regression $r = 0.78$. |

**Table 4.** *Cont.*

| Study | Population and Setting | Acoustic Input and Target | Validation and Separation | Performance: Unit and Threshold |
|---|---|---|---|---|
| Xie et al., 2023 [14] | 172 sleep-clinic participants | Nocturnal snoring timing/energy features, age and BMI; severity | Internal patient-level five-fold evaluation with inner tuning | Extended features: AHI MAE 7.26 events/h; four-category accuracy 59.3%. Reference excludes central apneas. |
| Wang et al., 2026 [16] | 73 laboratory participants (49 men); smartphone; separate PSG-Audio subset of 305 participants | Respiratory audio plus five clinical variables; local events and night index | Window test: 15 held-out patients; severity: separate ten-fold cross-validation; dataset-specific retraining on PSG-Audio | Window test: AUROC 0.919 (run SD 0.008), 15 held-out patients. Separate ten-fold patient protocol: severe-OSA accuracy 89.0% (fold SD 1.1 percentage points). Continuous AHI agreement not reported. |
| Song et al., 2025 [7] | Laboratory PSG-Audio; 50 held-out patients for AHI | 9 s sleep-audio windows; ConvNeXt–LSTM and AHI regression | Internal patient split: 205/50/50; separate subject-dependent segment experiment | Patient AHI MAE 10.29 events/h, $r = 0.804$ in the full-text analysis. |
| Romero et al., 2022 [21] | 103 participants; 157 home nights; level III reference | Smartphone breathing sounds; recording-time-normalized event index | Internal ten-fold patient separation; repeated nights stay in one fold | Night-level severe-OSA sensitivity/specificity 0.78/0.93 across repeated nights. |
| Le et al., 2023 [38] | 1018 PSG-Audio and 297 smartphone recordings; laboratory | Breathing sounds with home-noise augmentation; event/severity prediction | Internal 1165/150 split; simultaneous dual-microphone recordings excluded for subject-independent evaluation | Patient severity sensitivity/specificity 0.85/0.84 at AHI ≥15; epoch accuracy 86% is a separate outcome. |
| Han et al., 2024 [9] | 117 home attempts; 101 complete cases; level II PSG | iOS/Android smartphone breathing sounds; AHI/severity | Prospective home test of fixed model in a separate cohort; developer-affiliated investigators; sleeping alone | At AHI ≥30: sensitivity/specificity 93.3%/94.4% (iOS), 92.9%/94.3% (Android); 12 phone and four PSG failures. |
| Guan et al., 2026 [23] | Three hospitals; 152 recruited, 134 complete clinical records; 144 PSG, 121 snoring, 107 wearable records | Snoring model; PSG-referenced severity and wearable comparison | Prospective independent test of fixed model; PSG comparisons: snoring $n = 121$, wearable $n = 107$ | Severe-OSA patient AUROC 0.964 (95% CI 0.922–1.000); AHI bias −9.12 events/h, limits −41.25 to +23; modality AUROC differences not significant. |
| Qiu et al., 2024 [46] | 250 hospital participants | Snoring features, derived semantics, and event-record features; events/severity | Internal 200 training/50 test patients; event-record availability matters at inference | Patient four-class severity accuracy 58%; event-classification metrics describe a different target. |

**Table 4.** *Cont.*

| Study | Population and Setting | Acoustic Input and Target | Validation and Separation | Performance: Unit and Threshold |
|---|---|---|---|---|
| Stephens et al., 2026 [25] † | 180 hypertensive adults at home; OSA 91.7%, moderate-to-severe 65.5%; HSAT reference | Smartphone sleep sounds; moderate-to-severe OSA; STOP-Bang comparator | Clinical app evaluation; cutoff selected in evaluation cohort | At AHI ≥15: patient AUROC 0.831 (95% CI 0.773–0.890), sensitivity/specificity 72.0/85.5% at app cutoff 7.1. |

AHI, apnea–hypopnea index; AUROC, area under the receiver operating characteristic curve; BMI, body mass index; MAE, mean absolute error; PSG, polysomnography. AHI thresholds are in events/h. Internal denotes a split or cross-validation within the development cohort; separate-cohort testing uses separately recruited evaluation patients and does not imply investigator independence. Acquisition setting and validation design are reported separately. Recording-based splits are identified when patient separation could not be verified. Night-level results can include repeated observations from one patient. Reference [25] was assessed from its abstract and publisher preview. CI, confidence interval; SD, standard deviation. Values are selected operating points from the cited reports and depend on case mix, reference scoring, thresholds, and validation design. † identifies the relevant study rows assessed from abstracts or publisher previews.

**Table 5.** Additional acoustic studies: task, modeling approach, validation, and evaluation level. Output levels distinguish supporting acoustic tasks from patient OSA screening.

| Study | Population, Acoustic Input and Task | Modeling Approach | Validation and Separation | Performance: Unit and Threshold |
|---|---|---|---|---|
| Xue et al., 2025 [29] † | 301 men; controlled Mandarin articulation; OSA severity. | Explicit paralinguistic and transfer-learned spectrogram streams in PTF-Net. | Development study; external validation not established in the examined abstract. | Patient severity: reported AUC/UAR improvements 5.1/5.8%; absolute values and specific threshold unavailable in abstract. |
| Skapova et al., 2026 [43] | 93 participants; daytime speech; OSA screening. | MFCC and i-vector pipelines; exploratory frozen ECAPA-TDNN scoring. | Internal subject-independent evaluation; 93-person clinical dataset. | Patient OSA classification: i-vector SVM AUROC 0.987 (95% CI 0.974–0.997), accuracy 96.3%. |
| Wang et al., 2022, OSAnet [22] | 135 development participants; subsequently enrolled test cohort of 59. | Spectrogram event detection; duration and patient-level assessment. | 135 development patients; 59 subsequently recruited at the same center (temporal test). | Event precision/sensitivity 0.81/0.78; patient severe-OSA sensitivity/specificity 95.6/91.6%. |
| Li et al., 2023 [34] | 124 participants at home; 40 reserved for testing; OSA screening. | Visibility-graph images, CNN, and snore-based patient index. | Internal holdout: 40 of 124 participants reserved for testing; actual home acquisition. | Patient OSAHS classification: accuracy 92.5%, sensitivity/specificity 93.9/91.2%; index cutoff 0.5. |
| Li et al., 2023 [18] | 88 participants, ages 12–81, at home; snore/non-snore recognition. | Waveform and visibility-graph CNN branches; subject-separated folds. | Subject-separated folds; home recordings; supporting sound-recognition task. | Sound-level snore/non-snore AUROC 0.947; accuracy 89.3%; not a patient OSA endpoint. |

**Table 5.** *Cont.*

| Study | Population, Acoustic Input and Task | Modeling Approach | Validation and Separation | Performance: Unit and Threshold |
|---|---|---|---|---|
| Romero et al., 2023 [15] | 103 participants, 157 home nights; sound and respiratory effort. | Feature, latent, and decision fusion; participant-separated evaluation. | Internal participant-separated evaluation; same 103-person/157-night home cohort as 2022 study. | Severe-OSA sensitivity/specificity 0.93/0.92 with decision fusion; recording-night endpoint. |
| Lillini et al., 2024 [49] | 37 patients; tracheal audio; apnea-window identification. | Siamese CNN–GRU representation followed by clustering. | Three patient-separated splits: 23 train/six validation/eight test in each. | Six-second-window F1 0.90 (SD 0.008); apnea must span the full positive window. |
| Saha et al., 2025 [39] † | 50 participants with central and obstructive events. | Transfer-learned CNNs on tracheal spectrograms. | Event-subtype evaluation in 50 selected participants. | Central/obstructive event accuracy 83.66%; sensitivity and specificity each above 83%; patient separation not established in abstract. |
| Han, Xia, and Mascolo, 2026 [40] | 194 PSG-Audio participants; ambient and tracheal channels. | Multiple representations and networks; binary and finer event tasks. | Participant-independent public-data splits; paired channels from the same participants. | Three-class window F1 with EBranchformer: ambient 0.694 (filter banks), 0.714 (HuBERT); tracheal 0.760/0.746. |
| Xu et al., 2026 [24] | PSG-Audio development; 60-participant external cohort. | Spectrogram and energy streams with recurrent temporal modeling. | PSG-Audio development plus a separate 60-person external cohort. | External patient AHI MAE 6.03 events/h; at AHI ≥30: AUROC 0.99, sensitivity/specificity 89.29/100%. |
| Aygün Çakıroğlu et al., 2026 [56] † | 32 adults; six-person subject-level test set. | Compact CNN, attention, and transformer models. | Subject-level six-person test set, including one positive case. | Six test patients, one positive; too imprecise for comparative performance interpretation (Section 8.4). |
| Tran et al., 2023 [20] | 45 paired home PSG/audio recordings for testing; sleep staging. | Hospital supervision, home-domain learning, and noise consistency. | Home-domain learning followed by testing on 45 paired home PSG/audio recordings. | Three-state sleep-staging accuracy 76.2% at the epoch level. |
| Xie et al., 2025 [52] | 161 participants; concurrent cardiac and audio signals; sleep-disordered breathing. | Multi-input, multitask event and sleep–wake modeling. | Internal patient-separated four-fold evaluation with inner validation; thresholds selected on validation data. | Event F1 0.588; patient AHI $r = 0.825$; four-class severity accuracy 57.8%. |

**Table 5.** *Cont.*

| Study | Population, Acoustic Input and Task | Modeling Approach | Validation and Separation | Performance: Unit and Threshold |
|---|---|---|---|---|
| Bastani Najafabadi et al., 2026 [37] | 247 recruited; 228 after quality exclusions; 196 in main analysis; awake tracheal sound and anthropometry. | Two-level ensemble with subgroup-specific feature selection and bagged trees. | Nested patient-level CV; four outer folds; excludes AHI 10–15 from main comparison. | Patient AHI >15 versus <10: AUROC 0.77 (SD 0.01); sensitivity/specificity 84.3/59.9%. |
| Li et al., 2023 [54] † | 32 participants; snore audio; four respiratory-event classes. | Mel filter-bank features, bidirectional LSTM with attention; pruning and quantization for embedded hardware. | Subject-dependent and subject-independent comparisons; deployed model tested with leave-one-subject-out validation. | Event accuracy 83.52/62.31% before compression; deployed hardware accuracy 58.85% under leave-one-subject-out testing. |

UAR, unweighted average recall; SVM, support vector machine; OSAHS, obstructive sleep apnea–hypopnea syndrome; CNN, convolutional neural network; GRU, gated recurrent unit; MFCC, mel-frequency cepstral coefficient; PSG, polysomnography; AHI, apnea–hypopnea index; MAE, mean absolute error. References [29,39,54,56] were assessed from abstracts or publisher previews. The mixed-age snore-recognition study is included only as a supporting acoustic task. Validation refers to the reported separation and testing procedure; details that could not be verified from available material are identified. Results are reported at the sound, event, epoch, or patient level appropriate to each task. † identifies the relevant study rows assessed from abstracts or publisher previews.

Table 6 summarizes the review-specific appraisal, with explanations in Data S2. Participant selection was unclear for 27 of 31 reports; U denotes insufficient information, not demonstrated high risk of bias. Tan et al. applied QUADAS-2 and graded their selected overnight-audio evidence as high quality, with no included study judged at high risk [11]. Their study set, clinical diagnostic targets and assessment rules differ from this review's broader mixture of speech, event, supporting-task and device studies. Our selection domain also considers recording completeness and consequences of exclusions. These domain judgments therefore cannot be equated with Tan et al.'s overall quality grading. Concrete concerns here include evaluation-cohort cutoff selection, restricted disease spectra and small independent-patient test samples.

**Table 6.** Domain-level appraisal of the reported results. P, participant selection; I, predictor availability; R, reference; V, evaluation safeguards; Q, precision; A, applicability.

| Study | P | I | R | V | Q | A |
|---|---|---|---|---|---|---|
| Zhang et al. 2023 [4] | U | L | L | C | U | C |
| Wang et al. speech study [17] (pre-sleep) | U | L | L | U | L | C |
| TaghiBeyglou et al. 2026 [8] | U | L | L | U | U | C |
| Alqudah and Moussavi 2025 [30] | U | L | L | L | C | C |
| Alqudah and Moussavi 2026 [31] | U | L | L | L | C | C |
| Cho et al. 2022 [36] | U | L | L | L | U | C |
| Xie et al. 2023 [14] | U | L | L | L | L | C |
| Wang et al. respiratory-sound study [16] | U | L | L | L | C | C |
| Song et al. 2025 [7] | U | L | L | L | U | C |
| Romero et al. 2022 [21] | U | L | L | L | U | C |

**Table 6.** *Cont.*

| Study | P | I | R | V | Q | A |
|---|---|---|---|---|---|---|
| Le et al. 2023 [38] | C | L | L | L | U | C |
| Han et al. 2024 [9] | U | L | L | L | L | C |
| Guan et al. 2026 [23] | U | L | L | L | L | C |
| Qiu et al. 2024 [46] | U | U | L | L | U | C |
| Stephens et al. 2026 [25] † | U | U | L | C | U | C |
| Xue et al. 2025 [29] † | U | U | U | U | U | C |
| Skapova et al. 2026 [43] | U | L | L | L | L | C |
| Wang et al. OSAnet 2022 [22] | U | L | L | L | U | C |
| Li et al. OSA screening 2023 [34] | U | L | L | L | U | C |
| Li et al. sound recognition 2023 [18] | U | L | L | L | U | N |
| Romero et al. sensor fusion 2023 [15] | U | L | L | L | U | C |
| Lillini et al. 2024 [49] | C | L | L | L | C | C |
| Saha et al. 2025 [39] † | U | U | L | U | U | C |
| Han, Xia and Mascolo 2026 [40] | U | L | L | L | U | C |
| Xu et al. 2026 [24] | C | L | L | L | L | C |
| Aygun Cakiroglu et al. 2026 [56] † | U | U | U | U | C | C |
| Tran et al. 2023 [20] | U | L | L | L | U | N |
| Xie et al. 2025 [52] | U | L | L | L | U | C |
| Bastani Najafabadi et al. 2026 [37] | C | L | L | L | C | C |
| Li et al. embedded model 2023 [54] † | U | U | L | U | C | C |
| Sanchez Gomez et al. acoustic device [26] | U | L | L | L | L | C |

L, low concern; C, concern identified; U, insufficient information; N, not applicable. U is not high risk of bias. † Abstract or publisher-preview assessment. Ratings address the result discussed, not every experiment in a report; supporting observations are in Data S2.

### *8.2. Home Recording and Independent Populations*

Romero et al.'s internal patient-separated evaluation of 157 home nights from 103 participants yielded severe-OSA sensitivity/specificity of 0.78/0.93 for audio alone and 0.93/0.92 for sound–effort fusion [15,21]. Patient-clustered precision was not reported for these night-level estimates. Han et al. prospectively tested a previously developed model against level II home PSG in a separately recruited cohort [9]. At AHI ≥30 events/h, the sensitivity/specificity were 93.3/94.4% for iOS and 92.9/94.3% for Android. Complete data were available in 101/117 attempts (86.3%); 12 smartphone failures and four PSG attachment errors accounted for the remainder. Several investigators were affiliated with the model developer, Asleep Research Institute: the cohort was separate from development, but the investigators were not independent of the developer. Performance applies to the tested phones and sleeping-alone protocol.

Guan et al. prospectively tested a fixed snoring model at three hospital centers [23]. Of 152 recruits, 134 had complete clinical and demographic information, explaining the abstract denominator. Valid recordings comprised 144 PSG, 121 snoring and 107 wearable records; comparisons with PSG used the available records for each modality. Thus, 31/152 recruits (20.4%) lacked a usable snoring recording. Against the PSG AHI, the snoring estimate had a correlation r = 0.79 and ICC = 0.753; bias was −9.12 events/h, with agreement

limits of −41.25 to +23. The severe-OSA AUROC was 0.964 (95% CI, 0.922–1.000). For the source's paired mild-disease comparison, the snoring and wearable AUROCs were 0.83 and 0.72 ($p = 0.07$). Good discrimination and correlation therefore coexisted with wide individual AHI differences, without a statistically demonstrated advantage over the wearable comparator.

Recording feasibility includes obtaining a result from the intended patient. Shared rooms, variable microphone placement, background sounds, and interrupted recording differ from controlled single-sleeper protocols. Quality rejection, repeat attempts, and missing reference data affect the fraction of patients who receive a usable assessment. A low-error result among complete recordings can still be difficult to use if many attempts fail. Local processing and controlled audio retention also matter because bedside microphones capture household speech as well as breathing. For identifiable health-related audio, GDPR or equivalent requirements include a lawful basis for processing, protection of access and retention limited to a justified purpose and period [57].

### *8.3. Clinical Performance and Comparators*

The clinical value of an acoustic result lies in additional cases identified at a given referral burden, with missed severe cases and repeat recordings included. Stephens et al. provide a direct comparison with STOP-Bang, but their high-prevalence hypertensive cohort and locally selected cutoff restrict transfer to general screening [25]. Wang et al.'s speech comparison was internal, used an unvalidated clinical-only logistic regression and did not formally test the accuracy difference [17]. Neither study establishes how an acoustic strategy changes completed testing or patient outcomes in routine practice.

Overnight oximetry measures oxygen desaturation, which sound does not measure directly. Technically adequate HSAT records respiratory physiology and remains an established diagnostic option in appropriate adults [3]. Acoustic tools offer different acquisition requirements, but their value relative to oximetry or HSAT depends on diagnostic yield, recording completion, and downstream testing. Guan et al.'s wearable comparison illustrates the need to evaluate both discrimination and agreement: no significant AUROC difference demonstrated superiority of either modality [23]. The indirect comparisons in the overnight-audio meta-analysis likewise summarize separate evidence bases rather than clinical interchangeability [11]. Automated acoustic medical devices offer a useful clinical comparison. Sanchez Gomez et al. prospectively evaluated AcuPebble against PSG in Spain: 63 of 80 recruited patients provided paired diagnostic results, with sensitivity/specificity of 92.86/97.14% at AHI 15 [26]. Five device outputs were non-diagnostic, three recordings were not started, and nine exclusions concerned the reference equipment or recording. The proprietary system is considered here as a clinical comparator, rather than evidence for a particular AI architecture. Its results illustrate how diagnostic yield includes the availability of a usable output as well as accuracy among completed studies.

Among the devices considered here, AcuPebble SA100 has US FDA 510(k) clearance (K210480) for adult home OSA prescreening to assist a healthcare professional [58]. This product-specific indication does not confer clearance on other research models or replace full PSG when its additional measurements are needed.

Calibration, sensitivity, specificity, predictive values, and patient-level uncertainty describe how scores translate into decisions [59]. For continuous severity, absolute error and agreement supplement correlation. Population prevalence and disease spectrum affect predictive values and referral burden, while threshold optimization on evaluation patients overstates the performance of a fixed deployment rule. Early clinical evaluation also concerns time to completed testing and clinicians' use of predictions, as emphasized by DECIDE-AI [60]. These outcomes connect model performance with access to care.

### *8.4. Population Coverage and Reproducibility*

The number of independent patients determines precision more directly than the number of clips. Aygün Çakıroğlu et al. reported a subject-level AUROC of 1.00 with a 95% CI of 0.00–1.00 in six test patients, only one positive [56]. Small held-out samples also limit the interpretation of window-based models such as Wang et al.'s respiratory study (15 test patients) and SiCRNN (eight per split) [16,49]. A reported SD across runs or splits is not a patient-level confidence interval. Tables 4 and 5 distinguish these quantities where available.

Language, sex, body habitus, comorbidities, and referral pattern influence the population represented. Mandarin reading and Chinese male cohorts provide evidence under specific linguistic conditions [4,17,29]. A multilingual pretrained encoder does not itself validate an OSA model across languages. Han et al.'s home study and Guan et al.'s multicenter study broaden separate-cohort evaluation under specified protocols [9,23]. Reproducible public-data comparisons complement this evidence when channel selection, event labels, exclusions, and patient splits are documented. PSG-Audio is valuable for such comparisons, but repeated analyses of its participants are not additional independent clinical replications [40,61].

## 9. Discussion and Future Directions

### *9.1. Synthesis of the Evidence*

The pooled overnight-audio and awake-speech estimates summarize different clinical settings [11,12]. Tan et al.'s associations between sensitivity, train–test splitting and OSA prevalence reinforce the need to interpret performance through evaluation design and patient spectrum. Within-study comparisons explain a different source of variation: clinical-feature additions improved Xie et al.'s AHI correlation while reducing categorical accuracy; pretrained features helped ambient but not tracheal audio in Han et al.; and energy and recurrent components improved Xu et al.'s frame recognition [14,24,40]. A representation is useful in relation to its signal, prediction target and output.

Fusion and output construction can also change the final assessment. Romero et al.'s additional physiological channel increased severe-OSA sensitivity, while Wang et al.'s aggregation comparison changed the classification of 4 of 112 patients [15,17]. The respiratory study combined original-audio descriptors with a learned log-RMS envelope representation [16]. These comparisons identify useful modeling hypotheses within the tested pipelines. Demonstrating added clinical value requires paired patient-level comparisons, appropriate uncertainty and evaluation against the tests used in the intended pathway.

Separate-cohort evaluations provide a stronger test of transfer than internal model comparisons. Han et al.'s home results and Guan et al.'s multicenter study broaden this evidence while also showing the importance of recording completion and individual AHI agreement [9,23]. A model may support referral prioritization while estimating continuous severity imprecisely. Clinical usefulness depends on producing an interpretable result for the intended patients, including those whose first recording fails.

### *9.2. Clinical Implications*

A plausible near-term role is referral prioritization where the acoustic result adds to clinical risk assessment and leads to timely PSG or HSAT. Awake speech requires a brief task, whereas overnight sound adds recording burden but may capture event-related information. Their usefulness depends on which patients are additionally identified, how many extra referrals are generated, and whether incomplete recordings delay care. Low acoustic scores do not override symptoms or other indications for diagnostic testing [3]. Repeated acoustic

recording may support accessible follow-up, but monitoring disease or treatment response requires evidence beyond cross-sectional severity discrimination.

### *9.3. Research Priorities*

Three priorities emerge: matched model-design comparisons, evaluation of complete clinical outputs, and generalization with practical deployment. Shared patient partitions, clinical inputs and tuning opportunities help identify whether gains arise from representation, fusion, supervision or output formation. Testing components across more than one pipeline would establish how consistently their benefits transfer.

Complete-output evaluation connects local predictions to patient decisions. Relevant questions include whether richer aggregation improves on a trained mean-score comparator, whether temporal context improves event timing rather than only smoothing scores, and whether count or duration errors dominate severity estimates. For models using patient profiles, within-patient event discrimination helps separate local recognition from differences in baseline patient risk.

Generalization involves new patients, languages, devices, and home conditions, with clinical-only comparators and recording failures retained in the evaluation. Prospective testing of a fixed model gives different evidence from adaptation using target-population data. Computation, acquisition time, repeat attempts, and subgroup performance determine whether any statistical gain is practical. A longer-term direction is physiological interpretation: DISE or imaging can test anatomical associations, while synchronized airflow, effort, oxygenation, and arousal clarify which respiratory changes a model recognizes [62]. Associations with hypoxic burden or outcomes beyond AHI also require paired physiological evidence [63]. Feature attributions can help select acoustic patterns for these physiological investigations.

### *9.4. Limitations*

Purposive selection and extraction may limit coverage and introduce selection bias. The supplementary database searches did not comprehensively cover engineering literature. The review-specific appraisal provides domain judgments rather than a formal certainty grade; incomplete reporting also limited verification for some studies. Authorship of two detailed examples is disclosed. Differences in populations, reference scoring, inputs, targets and validation prevent reliable cross-study ranking. The synthesis supports interpretation of model-design choices and research priorities; comparative clinical effectiveness requires direct evaluation.

## 10. Conclusions

Acoustic AI offers promising approaches to screening and referral, with substantial differences in the evidence supporting individual tools. The published pooled sensitivity/specificity are approximately 90/87% for overnight audio and 83/83% for awake speech (rounded values); the speech studies used professional microphones in hospitals, and overnight-audio sensitivity varied with evaluation design and disease prevalence. Within-study comparisons link performance to representations, fusion and patient-output construction but do not by themselves establish incremental clinical utility. Separate-cohort studies add evidence on transfer, recording completion and AHI agreement. The next clinical priority is direct comparison with risk instruments and existing tests under intended-use conditions. Narrative selection and heterogeneous evaluations support design insights rather than a ranking of effectiveness. Acoustic AI does not currently replace PSG or technically adequate HSAT.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/medicina62112149/s1: Data S1: Literature search and study selection; Data S2: Study quality and clinical applicability; Data S3: Model implementation and component comparisons; Data S4: Study-level extraction record; Data S5: Additional reports in the retrieval inventory; Table S1: Study assessments and supporting observations; Table S2: Retrieval provenance for all appraised reports; Table S3: Implementation choices across acoustic OSA frameworks. All items are provided in one supplementary document.

**Author Contributions:** Investigation, P.W.; writing—original draft preparation, P.W.; writing—review and editing, P.W.; supervision, A.M. and C.-Y.T.; resources, A.M. and C.-Y.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The sources supporting this review are cited in the article.

**Conflicts of Interest:** The authors coauthored the respiratory-sound and speech studies discussed in this review [16,17], constituting a non-financial interest in their interpretation. The authors declare no financial conflicts of interest.

## References

1. Cao, S.; Rosenzweig, I.; Bilotta, F.; Jiang, H.; Xia, M. Automatic detection of obstructive sleep apnea based on speech or snoring sounds: A narrative review. *J. Thorac. Dis.* **2024**, *16*, 2654–2667. [CrossRef] [PubMed]
2. Alqudah, A.M.; Elwali, A.; Kupiak, B.; Hajipour, F.; Jacobson, N.; Moussavi, Z. Obstructive sleep apnea detection during wakefulness: A comprehensive methodological review. *Med. Biol. Eng. Comput.* **2024**, *62*, 1277–1311. [CrossRef] [PubMed]
3. Kapur, V.K.; Auckley, D.H.; Chowdhuri, S.; Kuhlmann, D.C.; Mehra, R.; Ramar, K.; Harrod, C.G. Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline. *J. Clin. Sleep Med.* **2017**, *13*, 479–504. [CrossRef] [PubMed]
4. Zhang, K.; Cao, L.; Ding, Y.; Li, Y.; Zhang, C.; Wu, J.; Han, D. Obstructive Sleep Apnea Detection using Pre-trained Speech Representations. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023; pp. 1139–1143. [CrossRef]
5. Zhang, Y.; Xia, T.; Han, J.; Wu, Y.Y.; Rizos, G.; Liu, Y.; Mosuily, M.; Chauhan, J.; Mascolo, C. Towards Open Respiratory Acoustic Foundation Models: Pretraining and Benchmarking. In Proceedings of the Advances in Neural Information Processing Systems 37, Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 27024–27055. [CrossRef]
6. Li, B.; Qiu, X.; Tan, X.; Yang, L.; Tao, J.; Fang, Z.; Huang, J. An end-to-end audio classification framework with diverse features for obstructive sleep apnea-hypopnea syndrome diagnosis. *Appl. Intell.* **2025**, *55*, 427. [CrossRef]
7. Song, Y.; Ding, L.; Peng, J.; Song, L.; Zhang, X. Screening for obstructive sleep apnea hypopnea using sleep breathing sounds based on the PSG-audio dataset. *Biomed. Signal Process. Control* **2025**, *103*, 107472. [CrossRef]
8. TaghiBeyglou, B.; Geng, J.; McDaulid, D.; Gnaneswaran, P.; Yasokaran, O.; Chow, A.; Ng, R.; Chervin, R.D.; Brown, D.L.; Yadollahi, A. Self-Supervised Speech Representations for Sleep Apnea Severity Prediction. *IEEE Trans. Biomed. Eng.* **2026**, *73*, 3601–3611. [CrossRef] [PubMed]
9. Han, S.C.; Kim, D.; Rhee, C.S.; Cho, S.W.; Le, V.L.; Cho, E.S.; Kim, H.; Yoon, I.Y.; Jang, H.; Hong, J.; et al. In-Home Smartphone-Based Prediction of Obstructive Sleep Apnea in Conjunction With Level 2 Home Polysomnography. *JAMA Otolaryngol.–Head Neck Surg.* **2024**, *150*, 22–29. [CrossRef] [PubMed]
10. TaghiBeyglou, B.; Čuljak, I.; Bagheri, F.; Suntharalingam, H.; Yadollahi, A. Estimating the severity of obstructive sleep apnea during wakefulness using speech: A review. *Comput. Biol. Med.* **2024**, *181*, 109020. [CrossRef] [PubMed]
11. Tan, B.K.J.; Gao, E.Y.; Tan, N.K.W.; Yeo, B.S.Y.; Tan, C.J.W.; Ng, A.C.W.; Leong, Z.H.; Phua, C.Q.; Uataya, M.; Goh, L.C.; et al. Machine Listening for OSA Diagnosis: A Bayesian Meta-Analysis. *CHEST* **2025**, *168*, 520–530. [CrossRef] [PubMed]
12. Gao, E.Y.; Hao, Y.; Tan, N.K.W.; Ng, A.C.W.; Leong, Z.H.; Phua, C.Q.; Ong, T.H.; Leow, L.C.; Huang, G.B.; Tan, B.K.J.; et al. Awake speech recordings for machine learning diagnosis of obstructive sleep apnea: A Bayesian meta-analysis. *J. Clin. Sleep Med.* **2026**, *22*, 5. [CrossRef] [PubMed]
13. Pillion, M.; Kim, P.; Larsson, M.; Jorge, E.; Orwa, J.; Gradisar, M. Accuracy of artificial intelligence-based prediction of moderate-to-severe obstructive sleep apnea using audio signals: A systematic review. *Sleep Med. Rev.* **2026**, *90*, 102357. [CrossRef] [PubMed]

14. Xie, J.; Fonseca, P.; van Dijk, J.; Overeem, S.; Long, X. Assessment of obstructive sleep apnea severity using audio-based snoring features. *Biomed. Signal Process. Control* **2023**, *86*, 104942. [CrossRef]
15. Romero, H.E.; Ma, N.; Brown, G.J.; Johnson, S. Obstructive sleep apnea screening with breathing sounds and respiratory effort: A multimodal deep learning approach. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023; pp. 5451–5455. [CrossRef]
16. Wang, P.; Lin, Y.C.; Liu, W.T.; Ma, E.; Majumdar, A.; Kang, J.H.; Gudnason, J.; Kuan, Y.C.; Lee, K.Y.; Feng, P.H.; et al. Deep representation learning with cross attention-based multi-feature fusion for sleep apnea detection using sleep respiratory sound. *Biomed. Signal Process. Control* **2026**, *113*, 109093. [CrossRef]
17. Wang, P.; Majumdar, A.; Liu, W.T.; Kang, J.H.; Gudnason, J.; Chen, Y.Y.; Liu, I.J.; Lee, K.Y.; Chen, T.T.; Lee, H.C.; et al. Dual contrastive learning with clinically guided multimodal fusion for speech-based severe sleep apnea risk screening. *Expert Syst. Appl.* **2027**, *333*, 133852. [CrossRef]
18. Li, R.; Li, W.; Yue, K.; Zhang, R.; Li, Y. Automatic snoring detection using a hybrid 1D–2D convolutional neural network. *Sci. Rep.* **2023**, *13*, 14009. [CrossRef] [PubMed]
19. Moons, K.G.M.; Damen, J.A.A.; Kaul, T.; Hooft, L.; Andaur Navarro, C.; Dhiman, P.; Beam, A.L.; Van Calster, B.; Celi, L.A.; Denaxas, S.; et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **2025**, *388*, e082505. [CrossRef] [PubMed]
20. Tran, H.H.; Hong, J.K.; Jang, H.; Jung, J.; Kim, J.; Hong, J.; Lee, M.; Kim, J.W.; Kushida, C.A.; Lee, D.; et al. Prediction of Sleep Stages Via Deep Learning Using Smartphone Audio Recordings in Home Environments: Model Development and Validation. *J. Med. Internet Res.* **2023**, *25*, e46216. [CrossRef] [PubMed]
21. Romero, H.E.; Ma, N.; Brown, G.J.; Hill, E.A. Acoustic Screening for Obstructive Sleep Apnea in Home Environments Based on Deep Neural Networks. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 2941–2950. [CrossRef] [PubMed]
22. Wang, B.; Tang, X.; Ai, H.; Li, Y.; Xu, W.; Wang, X.; Han, D. Obstructive Sleep Apnea Detection Based on Sleep Sounds via Deep Learning. *Nat. Sci. Sleep* **2022**, *14*, 2033–2045. [CrossRef] [PubMed]
23. Guan, H.; Huang, J.; Liu, D.; Yang, L.; Zeng, H.; Huang, J.; Miao, B. Multicenter Validation of an AI-Based Snoring Model vs. Wearables for Diagnosing Sleep Apnea. *Nat. Sci. Sleep* **2026**, *18*, 582714. [CrossRef] [PubMed]
24. Xu, M.; Li, Y.; Han, D. Fine-Grained and Lightweight OSA Detection: A CRNN-Based Model for Precise Temporal Localization of Respiratory Events in Sleep Audio. *Diagnostics* **2026**, *16*, 577. [CrossRef] [PubMed]
25. Stephens, T.N.; Wang, Y.; Guo, J.; Wang, J. Evaluation of a sleep sound analysis smartphone app for home obstructive sleep apnea screening among community-dwelling Chinese adults with hypertension. *Sleep Breath.* **2026**, *30*, 209. [CrossRef] [PubMed]
26. Sanchez Gomez, J.; Pramono, R.X.A.; Imtiaz, S.A.; Rodriguez-Villegas, E.; Valido Morales, A. Validation of a Wearable Medical Device for Automatic Diagnosis of OSA against Standard PSG. *J. Clin. Med.* **2024**, *13*, 571. [CrossRef] [PubMed]
27. Lechat, B.; Naik, G.; Reynolds, A.; Aishah, A.; Scott, H.; Loffler, K.A.; Vakulin, A.; Escourrou, P.; McEvoy, R.D.; Adams, R.J.; et al. Multinight Prevalence, Variability, and Diagnostic Misclassification of Obstructive Sleep Apnea. *Am. J. Respir. Crit. Care Med.* **2022**, *205*, 563–569. [CrossRef] [PubMed]
28. Saha, S.; Rattansingh, A.; Martino, R.; Viswanathan, K.; Saha, A.; Montazeri Ghahjaverestan, N.; Yadollahi, A. A pilot observation using ultrasonography and vowel articulation to investigate the influence of suspected obstructive sleep apnea on upper airway. *Sci. Rep.* **2024**, *14*, 6144. [CrossRef] [PubMed]
29. Xue, B.; Wang, Z.; Shao, Y.; Zhu, X.; Zhao, H.; Fu, C.H.; Xu, J.; Ding, N.; Qian, X.; Hong, H. Knowledge Guided Articulatory and Spectrum Information Fusion for Obstructive Sleep Apnea Severity Estimation. *IEEE J. Biomed. Health Inform.* **2025**. advance online publication. [CrossRef] [PubMed]
30. Alqudah, A.; Moussavi, Z. Assessing Obstructive Sleep Apnea Severity During Wakefulness via Tracheal Breathing Sound Analysis. *Sensors* **2025**, *25*, 6280. [CrossRef] [PubMed]
31. Alqudah, A.M.; Moussavi, Z. Awake Insights for Obstructive Sleep Apnea: Severity Detection Using Tracheal Breathing Sounds and Meta-Model Analysis. *Diagnostics* **2026**, *16*, 448. [CrossRef] [PubMed]
32. Fang, L.; Cai, J.; Huang, Z.; Tuohuti, A.; Chen, X. Assessment of simulated snoring sounds with artificial intelligence for the diagnosis of obstructive sleep apnea. *Sleep Med.* **2025**, *125*, 100–107. [CrossRef] [PubMed]
33. Brown, J.; Mitchell, Z.; Jiang, Y.A.; Archdeacon, R. Accuracy of Smartphone-Mediated Snore Detection in a Simulated Real-World Setting: Algorithm Development and Validation. *JMIR Form. Res.* **2025**, *9*, e67861. [CrossRef] [PubMed]
34. Li, R.; Li, W.; Yue, K.; Li, Y. Convolutional neural network for screening of obstructive sleep apnea using snoring sounds. *Biomed. Signal Process. Control* **2023**, *86*, 104966. [CrossRef]
35. Frija, J.; Millet, J.; Béquignon, E.; Covali, A.; Cathelain, G.; Houenou, J.; Benzaquen, H.; Geoffroy, P.A.; Bacry, E.; Grajoszex, M.; et al. Proposition of a new, minimally-invasive, software smartphone device to predict sleep apnea and its severity. *Sleep Breath.* **2025**, *29*, 282. [CrossRef] [PubMed]
36. Cho, S.W.; Jung, S.J.; Shin, J.H.; Won, T.B.; Rhee, C.S.; Kim, J.W. Evaluating Prediction Models of Sleep Apnea From Smartphone-Recorded Sleep Breathing Sounds. *JAMA Otolaryngol.–Head Neck Surg.* **2022**, *148*, 515–521. [CrossRef] [PubMed]

37. Bastani Najafabadi, V.; Ashraf, W.; Elwali, A.; Moussavi, Z. A Two-Level Ensemble Machine Learning Framework for OSA Classification Whilst Awake from Noisy Tracheal Breathing Sounds. *Sensors* **2026**, *26*, 1349. [CrossRef] [PubMed]
38. Le, V.L.; Kim, D.; Cho, E.; Jang, H.; Reyes, R.D.; Kim, H.; Lee, D.; Yoon, I.Y.; Hong, J.; Kim, J.W. Real-Time Detection of Sleep Apnea Based on Breathing Sounds and Prediction Reinforcement Using Home Noises: Algorithm Development and Validation. *J. Med. Internet Res.* **2023**, *25*, e44818. [CrossRef] [PubMed]
39. Saha, S.; Ghahjaverestan, N.M.; Yadollahi, A. Separating obstructive and central respiratory events during sleep using breathing sounds: Utilizing transfer learning on deep convolutional networks. *Sleep Med.* **2025**, *131*, 106485. [CrossRef] [PubMed]
40. Han, J.; Xia, T.; Mascolo, C. Continuous Mobile Audio Monitoring for Sleep Apnea Detection. *IEEE J. Biomed. Health Inform.* **2026**. advance online publication.. [CrossRef] [PubMed]
41. Song, Y.; Peng, J.; Ding, L.; Song, L.; Zhang, X. A noisy label correction framework for apnea-hypopnea index estimation from sleep breathing sounds. *Biocybern. Biomed. Eng.* **2026**, *46*, 162–175. [CrossRef]
42. Babu, A.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q.; Goyal, N.; Singh, K.; von Platen, P.; Saraf, Y.; Pino, J.; et al. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 2278–2282. [CrossRef]
43. Skapova, M.; Partila, P.; Tovarek, J.; Lubojacka, T.; Slonkova, J.; Hruby, O.; Genzor, S.; Mizera, J.; Mooz, P. Non-invasive detection of obstructive sleep apnea from speech in a limited clinical dataset. *Sci. Rep.* **2026**. advance online publication. [CrossRef]
44. Kim, Y.; Kim, M.; Shin, J.; Ko, M. ApneaWhisper: Transformer-Based Audio Segmentation for Fine-Grained Non-Invasive Sleep Apnea Detection. *Nat. Sci. Sleep* **2025**, *17*, 2455–2468. [CrossRef] [PubMed]
45. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.H.; Jin, D.; Naumann, T.; McDermott, M.B.A. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 72–78. [CrossRef]
46. Qiu, X.; Wang, C.; Li, B.; Tong, H.; Tan, X.; Yang, L.; Tao, J.; Huang, J. An audio-semantic multimodal model for automatic obstructive sleep Apnea-Hypopnea Syndrome classification via multi-feature analysis of snoring sounds. *Front. Neurosci.* **2024**, *18*, 1336307. [CrossRef] [PubMed]
47. Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; Courville, A. FiLM: Visual Reasoning with a General Conditioning Layer. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32, pp. 3942–3951. [CrossRef]
48. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. In *Proceedings of the Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 18661–18673. Available online: https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html (accessed on 29 September 2026).
49. Lillini, D.; Aironi, C.; Migliorelli, L.; Gabrielli, L.; Squartini, S. SiCRNN: A Siamese Approach for Sleep Apnea Identification via Tracheal Microphone Signals. *Sensors* **2024**, *24*, 7782. [CrossRef] [PubMed]
50. McDermott, M.B.; Zhang, H.; Hansen, L.H.; Angelotti, G.; Gallifant, J. A Closer Look at AUROC and AUPRC under Class Imbalance. In Proceedings of the Advances in Neural Information Processing Systems 37, Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 44102–44163. [CrossRef]
51. Pitkänen, M.; Nath, R.K.; Korkalainen, H.; Nikkonen, S.; Mahamid, A.; Oksenberg, A.; Duce, B.; Töyräs, J.; Kainulainen, S.; Leppänen, T. Respiratory event index underestimates severity of sleep apnea compared to apnea-hypopnea index. *Sleep Adv.* **2024**, *5*, zpad054. [CrossRef] [PubMed]
52. Xie, J.; Fonseca, P.; van Dijk, J.P.; Overeem, S.; Long, X. Multi-modal multi-task deep neural networks for sleep disordered breathing assessment using cardiac and audio signals. *Int. J. Med. Inform.* **2025**, *201*, 105932. [CrossRef] [PubMed]
53. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [CrossRef] [PubMed]
54. Li, H.; Lin, X.; Lu, Y.; Wang, M.; Cheng, H. Pilot study of contactless sleep apnea detection based on snore signals with hardware implementation. *Physiol. Meas.* **2023**, *44*, 085003. [CrossRef] [PubMed]
55. Li, H.; Lu, Y.; Qian, Y.; Zhou, L.; Wang, M.; Cheng, H. Learning generalizable representations with adversarial domain adaptation for snoring-based sleep apnea detection. *Biomed. Signal Process. Control* **2026**, *113*, 108937. [CrossRef]
56. Aygün Çakıroğlu, M.; Kızılkaya Aydoğan, E.; Bolattürk, Ö.F.; Aydoğan, S.; İsmailoğulları, S.; Delice, Y. Non-contact acoustic screening for sleep apnea: A subject-aware deep learning approach. *Sleep Breath.* **2026**, *30*, 36. [CrossRef] [PubMed]
57. Information Commissioner's Office. A Guide to the UK GDPR: Data Protection Principles, Lawful Basis and Special Category Data. Available online: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/ (accessed on 20 October 2026).
58. US Food and Drug Administration. AcuPebble SA100: 510(k) Summary and Clearance Letter, K210480. 6 July 2021. Available online: https://www.accessdata.fda.gov/cdrh_docs/pdf21/K210480.pdf (accessed on 20 October 2026).

59. Van Calster, B.; Collins, G.S.; Vickers, A.J.; Wynants, L.; Kerr, K.F.; Barreñada, L.; Varoquaux, G.; Singh, K.; Moons, K.G.M.; Hernandez-Boussard, T.; et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance. *Lancet Digit. Health* **2025**, *7*, 100916. [CrossRef] [PubMed]
60. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **2022**, *377*, e070904. [CrossRef] [PubMed]
61. Korompili, G.; Amfilochiou, A.; Kokkalas, L.; Mitilineos, S.A.; Tatlas, N.A.; Kouvaras, M.; Kastanakis, E.; Maniou, C.; Potirakis, S.M. PSG-Audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. *Sci. Data* **2021**, *8*, 197. [CrossRef] [PubMed]
62. Huang, Z.; Bosschieter, P.F.N.; Aarab, G.; van Selms, M.K.A.; Vanhommerig, J.W.; Hilgevoord, A.A.J.; Lobbezoo, F.; de Vries, N. Predicting upper airway collapse sites found in drug-induced sleep endoscopy from clinical data and snoring sounds in patients with obstructive sleep apnea: A prospective clinical study. *J. Clin. Sleep Med.* **2022**, *18*, 2119–2131. [CrossRef] [PubMed]
63. Labarca, G.; Vena, D.; Hu, W.H.; Esmaeili, N.; Gell, L.; Yang, H.C.; Wang, T.Y.; Messineo, L.; Taranto-Montemurro, L.; Sofer, T.; et al. Sleep Apnea Physiological Burdens and Cardiovascular Morbidity and Mortality. *Am. J. Respir. Crit. Care Med.* **2023**, *208*, 802–813. [CrossRef] [PubMed]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.