A Cascaded Two-Stage CNN Pipeline for Audio-Based Snore and Sleep Apnea Detection on Smartphones: Reference Baselines with Multi-Seed Validation

Yang L. (SomniAI LLC)
sun@somnisense.top

Abstract

Smartphone-based audio analysis offers a non-contact, low-cost path to sleep health screening at population scale. We present a **cascaded two-stage CNN pipeline** for audio-based sleep monitoring on consumer mobile devices: a short-window 2D CNN processes 1-second audio segments represented as 15 × 13 Mel-frequency cepstral coefficient (MFCC) matrices and produces a per-segment binary snore-presence indicator; that per-second indicator, together with two scalar acoustic features (sound pressure level and SPL change rate), is assembled into a 200 × 3 time-series feature matrix at 1 Hz sampling that forms the input to a long-window 1D CNN classifying the 200-second observation window as containing a sleep-disordered breathing event or not. The cascade structure — Stage-1 detector output forming a feature channel of the Stage-2 classifier's input — provides an ultra-compact intermediate representation (200 × 3 ≈ 600 floats per analysis window) suitable for on-device inference on consumer mobile hardware.

We evaluate both stages under a multi-seed protocol with 95% bootstrap confidence intervals (10,000 iterations). On a dataset of 13,538 labeled 1-second snore segments and 2,953 labeled 200-second apnea-detection windows — derived from **80 person-nights across 40 participants**, with PSG event annotations from a mix of in-laboratory PSG (10 nights) and portable / ambulatory PSG with nasal-airflow cannula (70 nights) — the two stages achieve, respectively, **94.29% (93.60, 95.02) accuracy** with **F1 90.28% (89.29, 91.38)** for Stage-1 (snore detection), and **83.82% (82.61, 85.14) accuracy** with **F1 83.99% (82.87, 85.05)** for Stage-2 (sleep apnea detection). The two-stage cascade and its compact intermediate representation are presented as reference baselines, with multi-seed bootstrap metrics that can serve as comparison points for downstream methodological work.

Keywords: snore detection, sleep apnea detection, MFCC, convolutional neural networks, multi-seed bootstrap, smartphone audio, biomedical time-series

arXiv Categories: cs.LG (primary), eess.AS (cross-listed)

1. Introduction

1.1 Motivation

Sleep-disordered breathing — particularly chronic snoring and obstructive sleep apnea (OSA) — is highly prevalent and substantially under-diagnosed worldwide. The clinical gold standard for diagnosis, in-laboratory polysomnography (PSG), is accurate but inaccessible to most patients due to cost, capacity, and the inherent inconvenience of an instrumented overnight stay. A growing body of work has therefore investigated lightweight, non-contact alternatives based on audio captured by consumer smartphones.

We approach this problem as a **cascaded two-stage pipeline** rather than as two parallel classification tasks. The pipeline consists of:

- **Stage 1 — Short-window snore detection.** A 2D CNN operates on a 15 × 13 MFCC representation of each successive 1-second audio segment and produces a per-second binary snore-presence indicator.
- **Stage 2 — Long-window sleep apnea detection.** A 1D CNN operates on a sliding 200 × 3 time-series feature matrix sampled at 1 Hz, where the three feature channels are sound pressure level, SPL change rate, and **the binary snore-presence indicator produced by Stage 1**. The output is a binary classification of whether the 200-second window contains a sleep-disordered breathing event.

The Stage-1 output forming an input channel to Stage 2 is the central architectural element of this pipeline: a 1-second classifier's binary output, sampled at 1 Hz, becomes part of the long-window classifier's input rather than being aggregated heuristically post-hoc. To the authors' knowledge, this cascaded architecture combined with the ultra-compact 200 × 3 intermediate representation has not been previously reported in smartphone-deployable audio-based sleep-disordered breathing detection.

On small-to-moderate clinical datasets, evaluations of audio sleep-monitoring CNNs have typically relied on single-seed train/validation/test splits. Single-seed comparisons can produce metric estimates whose variability across seed-induced split changes is comparable to the accuracy differences between architectures, making reported gains difficult to interpret.

1.2 Contributions

This paper makes the following contributions:

1. **A cascaded two-stage CNN pipeline for audio-based sleep-disordered breathing detection.** We present a pipeline in which a short-window (1-second) snore-detection CNN's per-second binary output forms one of three feature channels of an ultra-compact 200 × 3 time-series feature matrix sampled at 1 Hz, which serves as the input to a long-window (200-second) sleep apnea detection CNN. To the authors' knowledge, this combination of cascaded short-window-to-long-window output piping with a ≤ 5-channel 1 Hz intermediate

representation has not been previously reported in smartphone-deployable audio sleep monitoring.

2. **Two compact CNN architectures for the two stages.** We present small-footprint CNN designs — a 2D CNN over MFCC for Stage 1, and a three-block 1D CNN over the 200 × 3 feature matrix for Stage 2 — suitable for on-device inference on consumer mobile hardware. Per-stage CNN topologies follow established small-CNN design patterns; the per-stage topology choices are not themselves claimed as novel.
3. **Multi-seed bootstrap evaluation.** Each model is trained across five random seeds (42, 123, 456, 789, 2026) with stratified train/validation/test splits per seed. Results are reported as means with 95% bootstrap confidence intervals (10,000 iterations) over the seed-level metric distribution.
4. **Open evaluation code.** A complete Python implementation of the training pipeline, the multi-seed protocol, and the bootstrap analysis is released alongside this paper, enabling other researchers to compare under identical conditions.
5. **Reproducible numbers for downstream research.** All training/evaluation code and per-seed metric results are released, enabling fair comparison by future work — whether exploring alternative architectures, evaluation methodologies, or cross-site validation.

---

2. Related Work

2.1 Snore Detection from Audio

Snore detection has been studied extensively as a precondition for downstream sleep health analyses. Karunajeewa et al. (2008) introduced an early silence-breathing-snore classification system; Pevernagie et al. (2010) characterized the acoustic physiology of snoring sounds; Nakano et al. (2019) applied a deep neural network to tracheal sound analysis for sleep apnea detection. More recent deep-learning work has applied CNN architectures to MFCC-based snore vs non-snore classification with reported accuracies in the 85–95% range.

2.2 Audio-Based Sleep Apnea Detection

Audio-only sleep apnea detection from short or long windows has been investigated using a variety of features — MFCCs, sub-band energies, hand-crafted SPL trajectories — combined with neural network classifiers. To our knowledge, no public benchmark dataset exists for cross-paper comparison; methodological comparisons therefore rely on each paper's internal baselines.

---

3. Methodology

3.1 Datasets

Both tasks use derived acoustic features computed offline from overnight audio recordings paired with polysomnography (PSG)-based event annotations. The dataset, assembled by the author,

comprises **40 participants** and **80 person-nights**, distributed as:

- **5 participants × 10 person-nights** recorded at a hospital sleep center under full **in-laboratory PSG** (the clinical gold standard), and
- **35 participants × 70 person-nights** recorded in each participant's own home sleep environment with a **portable / ambulatory PSG unit** that included a nasal-airflow cannula.

Audio was captured by consumer recording devices in both settings, spanning a heterogeneous mix of smartphone and tablet microphones, room geometries, and ambient-noise profiles rather than a single fixed acquisition setup. Only the computed feature matrices and binary task labels are used for the analyses reported here; raw audio waveforms and any personally identifying information are not part of the analysis dataset and are not redistributed.

Snore dataset. 13,538 labeled 1-second audio segments, of which 9,636 are non-snore and 3,902 are snore. Each segment is represented as a 15 × 13 matrix of Mel-frequency cepstral coefficients (15 time frames × 13 cepstral coefficients per frame), computed using standard MFCC parameters consistent with prior literature.

Apnea dataset. 2,953 labeled 200-second observation windows, of which 1,498 are normal and 1,455 contain apnea or hypopnea events (1,089 apnea + 366 hypopnea, merged to binary "Abnormal" for the task). Each window is represented as a 200 × 3 matrix of per-second features sampled at 1 Hz: sound pressure level (dB), per-second SPL change rate, and **a binary snore-presence indicator produced as the output of the Stage-1 short-window CNN of §3.2 applied to the corresponding 1-second audio segments**. This per-second-Stage-1-output-as-Stage-2-feature-channel structure is the cascade link between the two stages.

For both datasets, we adopt a per-seed stratified 60% train / 20% validation / 20% test split, preserving class balance within each partition.

Ethics and data scope. The feature matrices analyzed in this paper were derived from audio recordings collected prior to this study in a sleep-medicine research context, spanning both in-laboratory and home recording settings. The original collection was conducted under participant consent; however, no formal IRB or equivalent ethics-review record is available to the authors, and the original consent did not explicitly address downstream publication of derived research findings. All analyses reported here operate exclusively on derived, non-identifying feature matrices; no raw audio, waveforms, labels, or personal identifiers are released or redistributed by the authors. Because of these data-provenance limitations, the results should be interpreted as exploratory and retrospective rather than as a formally ethics-cleared clinical study. Future work by the authors will use prospectively collected data under formal ethics review and consent procedures that explicitly permit downstream research publication.

3.2 Snore CNN Architecture (Task 1)

The snore CNN operates directly on the 15 × 13 MFCC feature image and is a standard small-footprint 2D CNN:

```
Input (15, 13, 1)
  → Conv2D(16, kernel=3×3, ReLU)
  → Flatten
  → Dense(128, ReLU)
  → Dense(64, ReLU)
  → Dropout(0.1)
  → Dense(1, Sigmoid)
```

Total parameters: **301,473**, dominated by the Flatten → Dense(128) projection. This is a compact design appropriate to the modest input size (15 × 13 = 195 input features) and suitable for on-device inference.

3.3 Apnea CNN Architecture (Task 2)

The apnea CNN operates on the 200 × 3 feature time series and is a three-block 1D CNN:

```
Input (200, 3)
  → Conv1D(16, k=3, ReLU) → MaxPool(2)
  → Conv1D(32, k=3, ReLU) → MaxPool(2)
  → Conv1D(64, k=3, ReLU) → MaxPool(2)
  → Flatten
  → Dense(128, ReLU)
  → Dense(64, ReLU)
  → Dropout(0.1)
  → Dense(1, Sigmoid)
```

Total parameters: **204,801**, again dominated by the Flatten → Dense(128) projection. This is a compact three-block 1D CNN design appropriate to the time-series input geometry, with the time dimension shrunk by max-pooling from 200 → 100 → 50 → 25 before flattening.

3.4 Training Protocol

Both models share an identical training configuration:

Hyperparameter	Value
Optimizer	Adam, initial learning rate = 0.001
Loss	Binary cross-entropy
Class weights	Inversely proportional to class frequency
Batch size	32
Maximum epochs	100
Early stopping	patience = 15 on validation loss
LR schedule	ReduceLROnPlateau (factor = 0.5, patience = 8, min_lr = 1e-6)
Random seeds	**5 seeds: 42, 123, 456, 789, 2026**
Train / val / test	60% / 20% / 20% stratified per seed

For each (task, seed) combination — 10 experiments in total — we fix all relevant random number generator states, perform a stratified split, train under the protocol above, and evaluate on the held-out test set at the default decision threshold of 0.5. Test-set predicted probabilities are saved to support reproducible downstream analyses.

3.5 Bootstrap Confidence Intervals

For each task and each metric we compute a percentile bootstrap confidence interval over the five seed-level values. With 10,000 bootstrap iterations, the 95% CI is given by the 2.5 and 97.5 percentiles of the bootstrap distribution of the sample mean. With only five seed-level observations, the bootstrap CI primarily characterizes seed-to-seed variability in our experimental setup; for larger-sample uncertainty on the test set itself, additional procedures (e.g., per-sample bootstrap on the held-out test predictions) would be required. We treat the seed-level CI here as a lower bound on result variability.

---

4. Results

4.1 Snore Detection (Task 1)

Table 1 reports the snore CNN's 5-seed mean and 95% bootstrap CI for each metric; Figure 1 visualizes the comparison with error bars.

Table 1. Snore CNN performance (5-seed mean, 95% CI).

Metric	Mean	95% CI Lower	95% CI Upper
Accuracy	94.29%	93.60%	95.02%
Precision	89.01%	86.30%	91.73%
Recall (Sensitivity)	91.67%	90.68%	92.72%
Specificity	95.35%	94.03%	96.67%
F1-Score	90.28%	89.29%	91.38%
AUC-ROC	0.9830	0.9806	0.9854
AUC-PR	0.9620	0.9573	0.9667

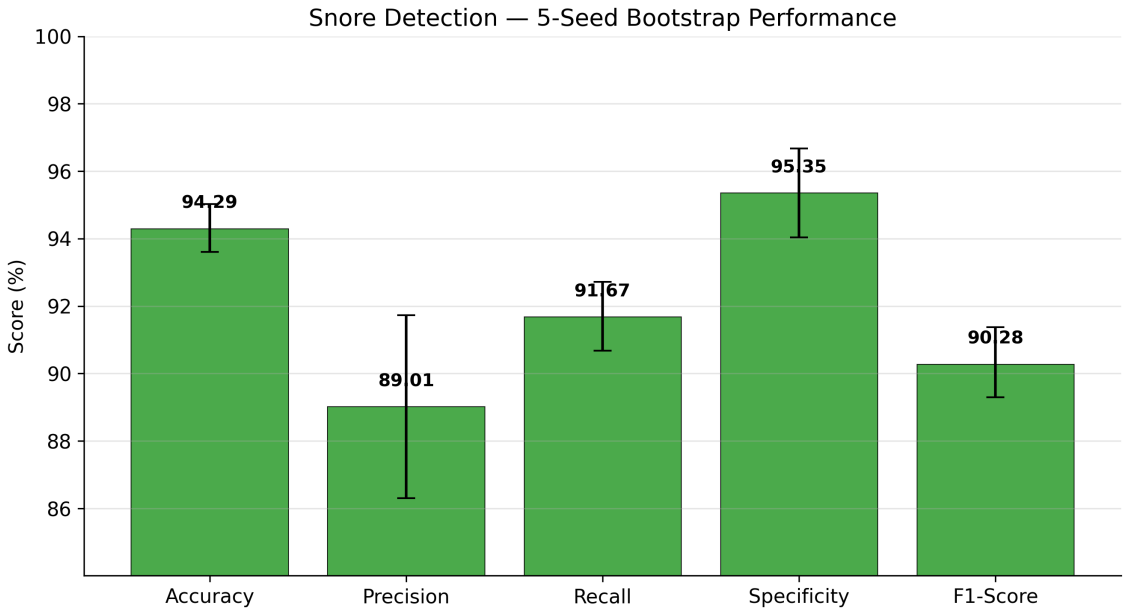

Figure 1. Snore CNN: 5-seed mean ± 95% bootstrap CI for the five primary metrics.

4.2 Sleep Apnea Detection (Task 2)

Table 2 reports the apnea CNN's 5-seed mean and 95% bootstrap CI; Figure 2 visualizes the comparison.

Table 2. Apnea CNN performance (5-seed mean, 95% CI).

Metric	Mean	95% CI Lower	95% CI Upper
Accuracy	83.82%	82.61%	85.14%
Precision	82.12%	79.94%	85.22%
Recall (Sensitivity)	86.12%	83.71%	88.52%
Specificity	81.60%	78.47%	85.67%
F1-Score	83.99%	82.87%	85.05%
AUC-ROC	0.9203	0.9107	0.9327
AUC-PR	0.9203	0.9094	0.9359

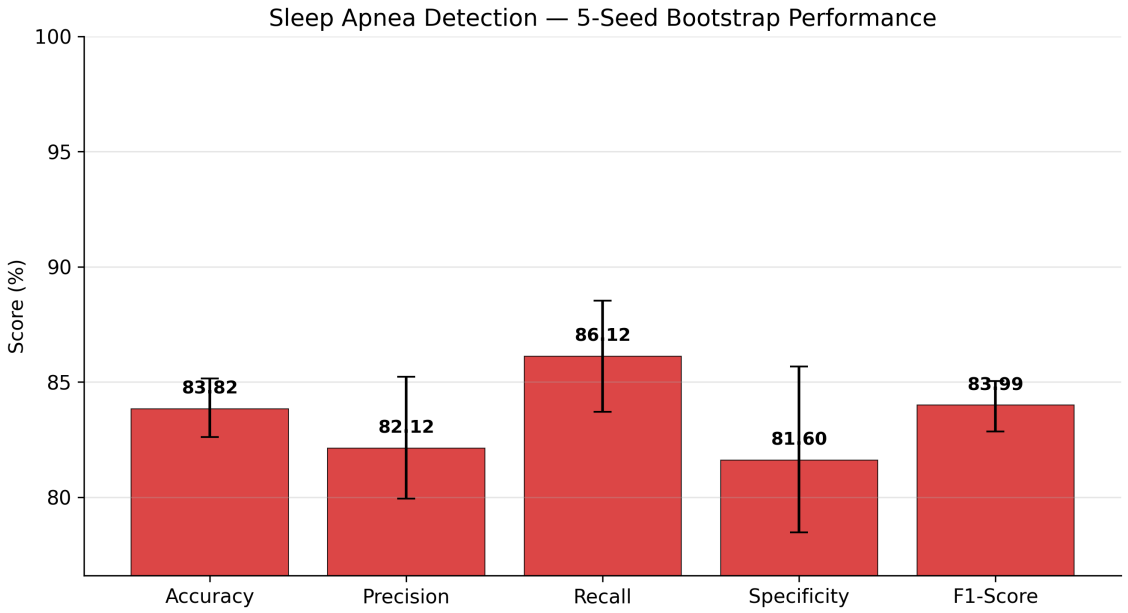

Figure 2. Apnea CNN: 5-seed mean ± 95% bootstrap CI for the five primary metrics.

4.3 ROC and Precision-Recall Curves

Figure 3 plots ROC curves for both models, averaged across 5 seeds, with ±1 SD shaded bands. Figure 4 plots Precision-Recall curves with the same averaging.

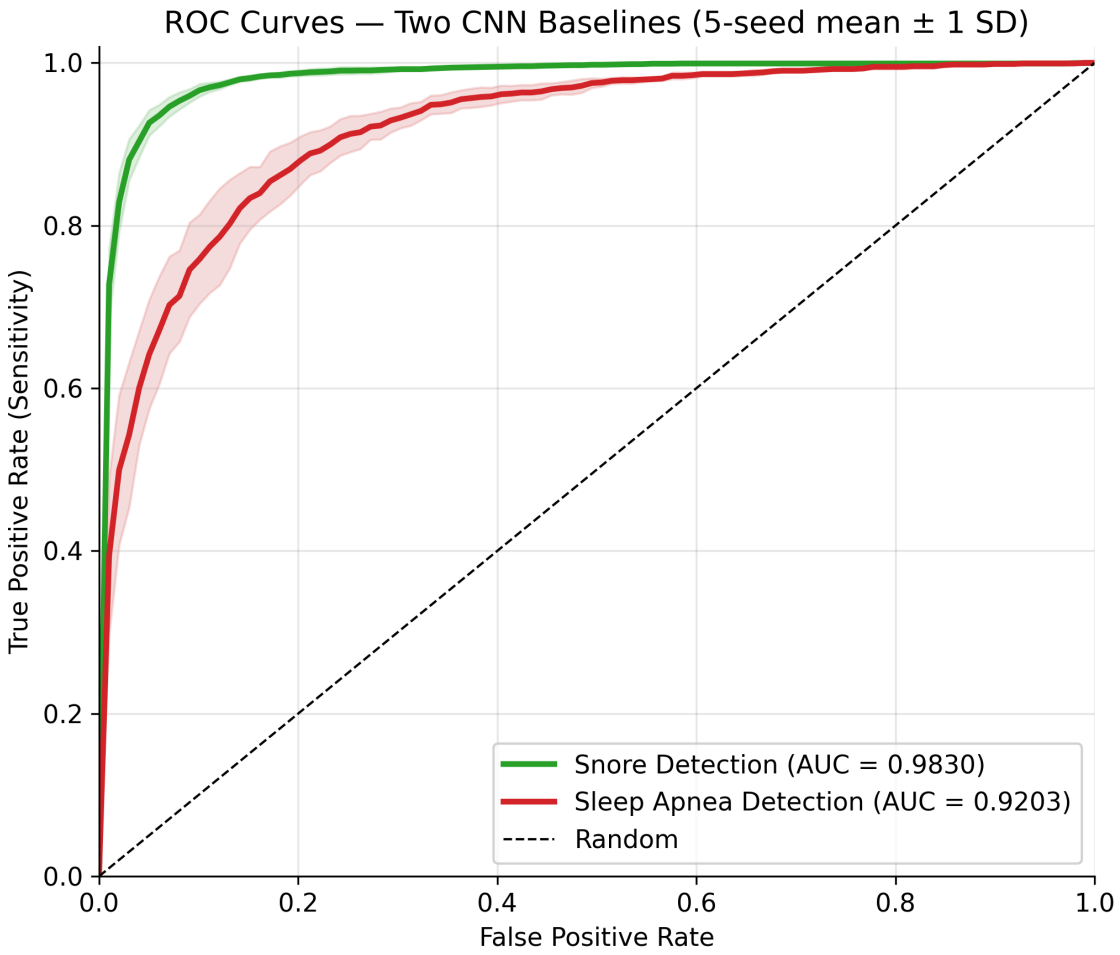

Figure 3. ROC curves, 5-seed mean ± 1 SD shaded band, for both baseline CNN models.

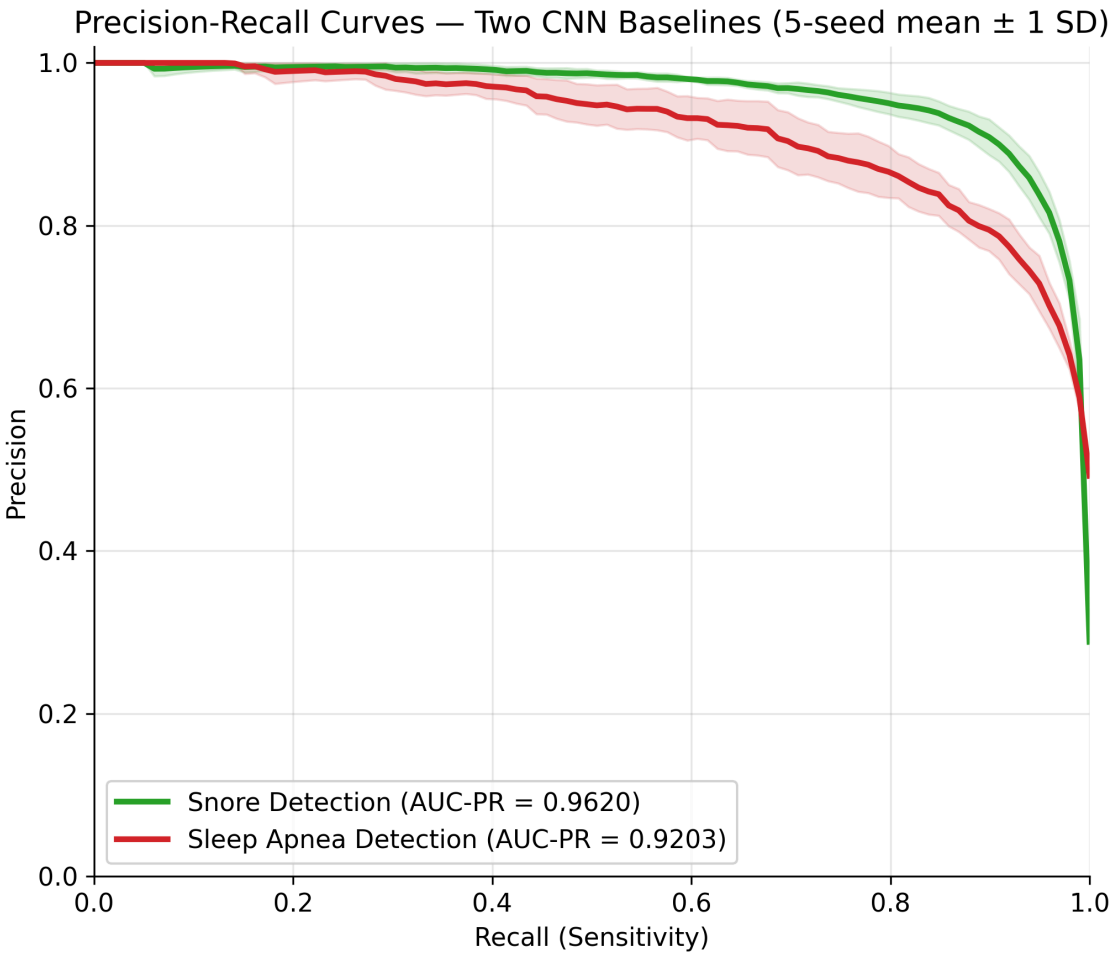

Figure 4. Precision-Recall curves, 5-seed mean ± 1 SD shaded band, for both baseline CNN models.

4.4 Confusion Matrices at Median-Accuracy Seeds

Figure 5 shows confusion matrices at the median-accuracy seed for each task, computed at the default decision threshold of 0.5.

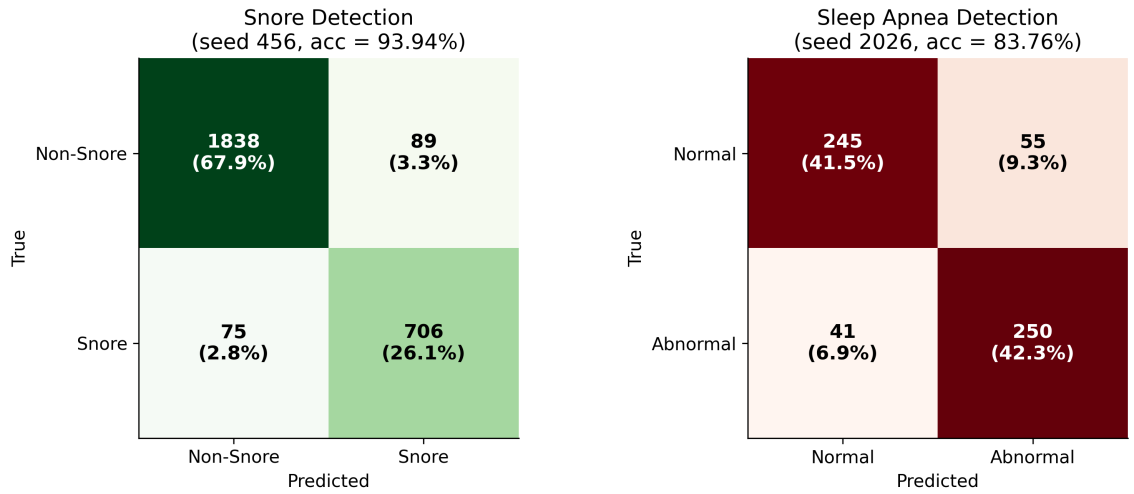

Figure 5. Confusion matrices for each task at the median-accuracy seed (default threshold 0.5).

---

5. Discussion

5.1 What These Results Establish

The two CNN baselines provide reproducible reference numbers on a 2,953-sample apnea dataset and a 13,538-sample snore dataset. The snore CNN reaches **94.29% mean accuracy** with **F1 90.28%** and **AUC 0.9830**; the apnea CNN reaches **83.82% mean accuracy** with **F1 83.99%** and **AUC 0.9203**. Both models are small-to-moderate in parameter count (~300k and ~205k respectively) and trained in seconds-to-minutes per seed on commodity CPU hardware without specialized acceleration.

The roughly 10-point accuracy gap between the snore task (94%) and the apnea task (84%) reflects the relative discriminability of the two tasks under audio-only observation: short-window snore-vs-non-snore boundaries are sharper acoustic phenomena than 200-second-window apnea-vs-normal boundaries, which require capturing temporally-extended patterns of silence and breathing-resumption rather than localized loud sounds.

5.2 What This Paper Does and Does Not Claim

This paper claims architectural novelty in the **cascaded two-stage pipeline structure** — specifically, that Stage 1's per-second binary output forms one of three feature channels in Stage 2's 200 × 3 input matrix at 1 Hz sampling — and in the **ultra-compact intermediate representation** that this structure enables (≈ 600 floats per 200-second analysis window, in contrast to the multi-MB spectrogram inputs typical of end-to-end audio classifiers).

This paper does **not** claim novelty in the per-stage CNN topology. The 2D CNN for Stage 1 and the three-block 1D CNN for Stage 2 follow well-established small-footprint CNN design patterns. Architectural refinements at the per-stage level — including a one-dimensional coordinate

attention mechanism for the Stage-2 classifier and a compression pipeline targeting mobile neural processing units — are addressed in companion work and are not the subject of this paper. The contribution here is the cascade structure and its evaluation rigor, not the per-stage network topology.

5.3 Open Questions Beyond This Paper

Several open questions are worth investigating. We list them here without prejudging which directions will prove most fruitful:

- **How robust are these numbers across recording conditions?** Validation on independent datasets recorded at different sites, with different microphone hardware, room acoustics, and patient populations, is the most important open question for clinical translation. Distribution shift across microphones (handset, far-field, headset) is particularly under-studied for this class of audio task.
- **Are there better features than MFCC + SPL?** Both tasks use hand-engineered front-ends established in prior literature. Whether learned front-ends (e.g., raw-waveform 1D convolutions, wavelet-based representations, or self-supervised audio embeddings such as YAMNet or CLAP) would change the achievable accuracy ceiling is an empirical question this paper does not answer.
- **Are CNNs even the right family?** For the 200 × 3 apnea input in particular, sequence models (small Transformers, state-space models such as Mamba, temporal convolutional networks) could in principle capture the long-range "silence-then-resumption" pattern more efficiently than a fixed-kernel CNN with a heavy Flatten + Dense head. Whether they actually outperform on $n \approx 3{,}000$ samples is non-obvious.
- **What is the impact of label noise and annotator agreement?** PSG event scoring has known inter-rater variability, especially for hypopneas. Quantifying how much of the residual 16-point gap on the apnea task is irreducible label noise versus model capacity is itself a research question.
- **Multi-class and patient-level extensions.** The current binary tasks could be extended — to multi-class apnea/hypopnea/normal, to snoring subtype classification, or to night-level AHI estimation aggregated across windows. Each extension changes the evaluation metric set and is not a simple drop-in replacement.
- **Deployment characterization.** For any eventual on-device use, model footprint, inference latency, and power draw on consumer mobile neural processing units (e.g., Apple Neural Engine, Android NNAPI) are practical questions distinct from accuracy on a held-out test set.

This list is intentionally non-exhaustive. The baselines and evaluation pipeline released with this paper are intended to make any of these directions cheaper to investigate by providing a fair comparison point, not to commit to a specific follow-up agenda.

5.4 Limitations

Dataset size and participant pool. The apnea dataset (2,953 windows) and the snore dataset (13,538 segments) derive from 40 distinct participants and 80 person-nights. The audio side is naturally diverse — multiple consumer recording devices, and participants' own home sleep environments for 70 of the 80 nights — which approximates a real deployment distribution rather than a controlled-laboratory one. The principal caveats are (i) that 40 participants is still a modest pool that is not formally stratified by geography or demographic group, so systematic per-device or per-population evaluation is not provided here, and (ii) that the dataset spans two distinct PSG acquisition modes — in-laboratory PSG (10 nights) and portable / ambulatory PSG with a nasal-airflow cannula (70 nights) — which is helpful for ecological validity but means a subset-level comparison of model performance between the two modes has not been characterized in this paper.

Default-threshold evaluation. All primary results are reported at the default decision threshold of 0.5. Deployment threshold tuning to balance sensitivity and specificity per use case is beyond the scope of this paper.

No paired-architecture comparison. Because this paper studies a single architecture per task, we report per-task per-metric CIs but do not perform paired bootstrap comparison between architectures. Such comparisons are outside the scope of this baseline-establishing paper.

---

6. Conclusion

We presented a cascaded two-stage CNN pipeline for audio-based snore and sleep apnea detection on consumer smartphones, in which a 2D-CNN snore detector's per-second binary output forms one of three feature channels of an ultra-compact 200 × 3 time-series feature matrix sampled at 1 Hz, which serves as the input to a 1D-CNN sleep apnea detector. Both stages were evaluated under a multi-seed bootstrap protocol on a combined dataset of 16,491 labeled samples derived from audio-polysomnography paired recordings, achieving 94% accuracy on Stage-1 snore detection and 84% accuracy on Stage-2 sleep apnea detection at the default decision threshold of 0.5.

The contributions of this paper are (i) the cascaded two-stage pipeline structure with ultra-compact intermediate representation, (ii) methodological rigor (multi-seed bootstrap CI on small-dataset audio classification), and (iii) reproducibility (public evaluation code). The reported numbers and open evaluation pipeline are intended to support comparative architectural and methodological work in audio-based sleep monitoring.

Artifact Availability

Python source code for the two model architectures, the training pipeline, and the bootstrap analysis is publicly released at: `https://github.com/somnisense/audio-sleep-cnn-baselines` under the MIT License. By design, **no audio recordings, no labels, and no derived feature matrices are distributed with the repository** — the original recordings were

collected under participant consent that does not cover public release of either the waveforms or the derived features. The training and evaluation scripts run against any dataset that conforms to the I/O contract documented in the repository README.

Patent Disclosure

The cascaded two-stage architecture and ultra-compact 200 × 3 intermediate representation described in this paper are the subject of three co-filed U.S. Provisional Patent Applications by SomniAI LLC:

1. **Cascaded Two-Stage Audio Architecture for Sleep-Disordered Breathing Detection with Ultra-Compact On-Device Feature Representation** — directed to the cascade pipeline structure, the at-1-Hz N-by-K ($K \leq 5$) feature matrix construction, and on-device deployment.
2. **Coordinate Attention Block for One-Dimensional Time-Series Classification and Compression Pipeline Comprising Architectural Redesign, Quantization-Aware Training, and Structured Filter Pruning** — directed to a 1D coordinate attention mechanism applicable as a Stage-2 architectural refinement and a compression pipeline producing compact deployed models, both addressed in companion work.
3. **Sound-Pressure-Level Multi-Stage Gating, Event-Driven Inference Triggering, and Privacy-Preserving On-Device System Architecture for Audio-Based Sleep Monitoring** — directed to multi-stage SPL gating, event-driven invocation of the Stage-2 classifier, and the fully-on-device system architecture.

Mathematical and architectural details described in this paper are presented for reproducibility; certain implementation specifics — particularly the compression pipeline, the multi-stage gating procedures, and the event-driven triggering logic — are covered by the co-filed patent applications and are not described in this manuscript.

References

1. Karunajeewa, A. S., Abeyratne, U. R., & Hukins, C. (2008). Silence-breathing-snore classification from snore-related sounds. *Physiological Measurement*, 29(2), 227–243. DOI: 10.1088/0967-3334/29/2/006.
2. Pevernagie, D., Aarts, R. M., & De Meyer, M. (2010). The acoustics of snoring. *Sleep Medicine Reviews*, 14(2), 131–144. DOI: 10.1016/j.smrv.2009.06.002.
3. Nakano, H., Furukawa, T., & Tanigawa, T. (2019). Tracheal Sound Analysis Using a Deep Neural Network to Detect Sleep Apnea. *Journal of Clinical Sleep Medicine*, 15(8), 1125–1133. DOI: 10.5664/jcsm.7804.
4. Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on*

Acoustics, Speech, and Signal Processing, 28(4), 357–366. DOI: 10.1109/TASSP.1980.1163420.

5. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*. URL: https://arxiv.org/abs/1704.04861.
6. Berry, R. B., et al. (2023). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 3.0*. Darien, IL: American Academy of Sleep Medicine.
7. Yang L. (2026). *A 1D Coordinate Attention CNN for Audio-Based Sleep Apnea Detection: Architecture and Multi-Seed Evaluation* (companion work).
8. Yang L. (2026). *Compression Pipeline for Sub-Millisecond Sleep Apnea Detection on Mobile Neural Engines* (companion work).
9. SomniAI LLC. (2026). *Cascaded Two-Stage Audio Architecture for Sleep-Disordered Breathing Detection with Ultra-Compact On-Device Feature Representation*. U.S. Provisional Patent Application.
10. SomniAI LLC. (2026). *Coordinate Attention Block for One-Dimensional Time-Series Classification and Compression Pipeline Comprising Architectural Redesign, Quantization-Aware Training, and Structured Filter Pruning*. U.S. Provisional Patent Application.
11. SomniAI LLC. (2026). *Sound-Pressure-Level Multi-Stage Gating, Event-Driven Inference Triggering, and Privacy-Preserving On-Device System Architecture for Audio-Based Sleep Monitoring*. U.S. Provisional Patent Application.

Appendix A — Raw Per-Seed Results

Task	Seed	Accuracy	Precision	Recall	Specificity	F1	AUC-ROC	AUC-PR
snore	42	93.72%	89.91%	92.82%	94.09%	89.49%	0.9822	0.9603
snore	123	93.32%	84.96%	93.33%	93.31%	88.94%	0.9795	0.9583
snore	456	93.94%	88.78%	90.40%	95.38%	89.59%	0.9808	0.9609
snore	789	95.01%	91.62%	91.04%	96.63%	91.33%	0.9854	0.9646
snore	2026	95.46%	93.69%	90.78%	97.35%	92.02%	0.9870	0.9659
apnea	42	82.74%	78.90%	88.66%	77.00%	83.50%	0.9158	0.9160
apnea	123	81.90%	80.67%	83.16%	80.67%	81.90%	0.9064	0.9043
apnea	456	84.60%	81.06%	89.69%	79.67%	85.15%	0.9216	0.9180
apnea	789	86.13%	88.00%	83.16%	89.00%	85.51%	0.9449	0.9500
apnea	2026	83.76%	81.97%	85.91%	81.67%	83.89%	0.9129	0.9133

Appendix B — Reproducibility

All experiments run on commodity CPU hardware with Python + TensorFlow; no GPU is required. Per-seed training time: snore CNN 10-15 seconds; apnea CNN 5-7 seconds. The complete 10-experiment grid completes in approximately 2 minutes.

Source code is hosted at `https://github.com/somnisense/audio-sleep-cnn-baselines` (see also the *Artifact Availability* section above). The full reproduction pipeline:

```
git clone https://github.com/somnisense/audio-sleep-cnn-baselines.git
cd audio-sleep-cnn-baselines/code
pip install -r requirements.txt        # tensorflow, numpy, scipy, matplotlib

python run_experiments.py              # 5-seed × 2-task grid (~2 min on CPU)
python analyze_results.py              # bootstrap CI + markdown summary
python generate_figures.py             # 5 paper figures
```

Random seed scope: Python `random`, NumPy `np.random`, TensorFlow `tf.random`, and `PYTHONHASHSEED` are all fixed prior to each experiment.