Coordinate Attention for 1D Audio-Based Sleep Apnea Detection: A Multi-Seed Empirical Study on Smartphone-Deployable Architectures

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

Abstract

Sleep apnea affects an estimated 936 million adults globally, yet over 80% of cases remain undiagnosed due to the inaccessibility of polysomnography (PSG), the clinical gold standard. We present a lightweight neural network architecture for detecting sleep-disordered breathing events directly from smartphone audio, eliminating the need for wearable devices, cloud connectivity, or external sensors. The proposed architecture adapts the Coordinate Attention (CA) mechanism — originally proposed for 2D mobile vision networks — to one-dimensional biomedical time-series classification through a global-local feature fusion design that preserves temporal-position information in the attention map.

Operating on a 200 × 3 acoustic feature matrix sampled at 1 Hz — comprising sound pressure level, SPL change rate, and **a binary snore-presence indicator produced by a short-window snore-detection CNN as the Stage-1 output of a cascaded two-stage pipeline** (the cascade architecture itself is the subject of companion work [Yang L., 2026a]) — we evaluate three architectures as Stage-2 classifier candidates: a vanilla 1D CNN baseline, an SE-Attention CNN, and the proposed Coordinate Attention (Coord-Attn) CNN. All three are evaluated under an identical training protocol across **five random seeds (42, 123, 456, 789, 2026)** on an internal dataset of 2,953 PSG-labeled 200-second windows drawn from 80 person-nights across 40 participants (in-laboratory PSG + portable PSG with nasal-airflow cannula). Results are reported as **mean with 95% bootstrap confidence intervals (10,000 iterations)** to characterize variability and pairwise differences are evaluated by paired-seed bootstrap.

The Coord-Attn CNN achieves an accuracy of **87.14% (85.14, 89.68)** and an F1-score of **86.94% (84.28, 89.77)** while using only **14,001 parameters — a 93.2% reduction relative to the 204,801-parameter baseline (83.82% accuracy, 83.99% F1)**. Paired bootstrap analysis shows that Coord-Attn is statistically distinguishable from the Original baseline in six of seven metrics (Accuracy, Precision, Specificity, F1, AUC-ROC, AUC-PR; Recall is not). Against the SE-Attention CNN of comparable size, Coord-Attn provides a statistically significant advantage in Precision and Specificity but not in F1 or AUC, suggesting that the temporal-position-preserving property of Coordinate Attention is most beneficial when minimizing false alarms matters more than maximizing sensitivity.

The Coord-Attn block disclosed in this paper, together with the cascade architecture in which it serves as Stage-2 classifier and a compression pipeline for on-device deployment, are the subject of three co-filed U.S. provisional patent applications by SomniAI LLC. See §Patent Disclosure for details.

Keywords: sleep apnea detection, coordinate attention, 1D convolutional neural networks, edge AI, smartphone health, audio classification, biomedical time-series

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

---

1. Introduction

1.1 Clinical Motivation

Obstructive Sleep Apnea (OSA) is among the most common chronic sleep disorders, characterized by repeated upper-airway obstruction during sleep that produces breathing pauses of ten seconds or longer. Untreated OSA elevates cardiovascular risk, contributes to daytime cognitive impairment and excessive sleepiness, and is associated with increased all-cause mortality. Despite substantial public health impact, an estimated 80–90% of OSA cases remain undiagnosed worldwide.

The clinical gold standard for OSA diagnosis — in-laboratory polysomnography (PSG) — requires an overnight stay with multiple physiological sensors attached to the patient. PSG is expensive (US$1,000–3,000 per study), inaccessible to most patients globally, and produces a single-night snapshot rather than longitudinal data reflective of typical home sleep.

This work addresses the following engineering challenge:

> *Can a lightweight neural network, operating entirely on a consumer smartphone with the built-in microphone as its sole sensor, detect sleep-disordered breathing events with clinically meaningful sensitivity, while maintaining inference latency suitable for real-time on-device deployment?*

1.2 Why Standard CNN Architectures Are Insufficient

Standard 1D convolutional neural networks (CNNs) for biomedical time-series classification typically apply uniform weighting across time steps and feature channels. For sleep apnea detection, this is suboptimal because:

- **Temporal structure is clinically meaningful.** A typical obstructive apnea event with arousal often exhibits three temporally distinct phases: onset (cessation or substantial reduction of airflow), silence plateau (≥ 10 seconds, per AASM scoring criteria), and recovery (often, though not universally, marked by a gasping snore or arousal). Each phase has a distinct acoustic signature. This three-phase narrative is illustrative motivation rather than a strict clinical rule — central apneas may lack the gasping recovery, and hypopneas may exhibit continued airflow (and continued snoring) throughout the event.

- **Channel relevance is temporally non-stationary.** A change-rate feature is most informative at event onset and recovery; a binary snore-presence feature is most informative during silence intervals (for apnea phenotypes that exhibit one). A model that cannot modulate channel importance by time step cannot fully exploit these complementarities.

These observations motivate an attention mechanism sensitive to both **channel importance** and **temporal position** — a property that Squeeze-and-Excitation (SE) attention [Hu et al., 2018] cannot provide, because SE aggregates time globally before computing channel weights, producing a time-invariant 1 × C attention vector that is broadcast uniformly across all time steps.

1.3 Coordinate Attention for 1D Time-Series

The Coordinate Attention (CA) mechanism, introduced by Hou et al. (CVPR 2021) for 2D mobile vision networks, addresses the analogous limitation in spatial settings by splitting global pooling into two directional pooling operations, preserving positional information. We adapt this concept to 1D time-series through a **global-local feature fusion** design: a global context vector is tiled to the temporal dimension and concatenated channel-wise with the local feature map, producing an attention map of shape T × C (time × channels) rather than the 1 × C produced by SE.

1.4 Contributions

This work makes the following contributions:

1. **A 1D adaptation of Coordinate Attention** for biomedical time-series classification, replacing the original 2D directional pooling with a global-local fusion scheme appropriate to 1D temporal geometry. This is the primary architectural contribution of the paper.
2. **A Stage-2 classifier instance for a cascaded two-stage audio sleep monitoring pipeline.** The 200 × 3 input feature matrix this paper operates on is the output of a cascade in which a Stage-1 short-window snore-detection CNN's per-second binary output forms one of the three feature channels of the Stage-2 input — a cascade structure described in companion work [Yang L., 2026a] and the subject of a co-filed patent application (see §Patent Disclosure). The CA-1D mechanism is presented here as a Stage-2 architectural refinement that exploits the temporal structure of the cascade-produced feature matrix.
3. **A multi-seed empirical study** comparing three Stage-2 candidates — vanilla 1D CNN, SE-Attention CNN, and Coordinate Attention CNN — under an identical training protocol across **five random seeds**, with results reported as mean with 95% bootstrap confidence intervals. This addresses a methodological weakness of typical small-dataset biomedical AI papers, in which single-seed results can produce misleading conclusions about architecture comparisons.
4. **A 93.2% reduction in model parameters** (204,801 → 14,001) while statistically significantly improving Accuracy, Precision, Specificity, F1, AUC-ROC, and AUC-PR over the baseline.
5. **An honest characterization of when Coordinate Attention adds value versus when it does not.** Against an SE-Attention baseline of comparable parameter count, Coord-Attn improves Precision and Specificity (statistically distinguishable at $\alpha = 0.05$) but does not

yield statistically significant gains in F1 or AUC — useful information for practitioners deciding between attention designs for analogous time-series tasks.

---

2. Related Work

2.1 Coordinate Attention (Hou et al., CVPR 2021)

The original Coordinate Attention paper [Hou et al., 2021] introduces an attention mechanism for 2D mobile vision networks (MobileNetV2, EfficientNet). The key innovation is splitting 2D global average pooling into two **directional** pooling operations along the H and W axes, encoding positional information into the channel attention weights. The mechanism is designed for image feature maps where spatial position carries discriminative information — for example, the location of an object's edges relative to its center.

In contrast, the original SE-Net [Hu et al., 2018] performs a single global average pooling that destroys all positional information before computing channel weights, producing an attention vector of shape 1 × C that is broadcast uniformly across spatial positions.

2.2 Adapting CA from 2D to 1D Time-Series

Direct application of the 2D directional pooling to 1D time-series does not transfer naturally — there is only one "axis" (time) in 1D data, so two orthogonal pooling directions are not available. We instead introduce a **global-local fusion** scheme:

- **Global context:** Compute global average pooling over time, producing a context vector of shape 1 × C.
- **Local features:** Retain the input feature map of shape T × C.
- **Fusion:** Tile the global context to T × C and concatenate channel-wise with the local features, producing T × 2C.
- **Bottleneck:** Apply 1 × 1 Conv1D + BatchNorm + ReLU to reduce to T × (C / r), where r is a reduction ratio hyperparameter.
- **Attention map:** Apply 1 × 1 Conv1D + Sigmoid to produce the final T × C attention map.

The resulting attention map A(t, c) provides per-time-step, per-channel weights. We note this design is not a mathematically exact 1D analog of the original 2D CA mechanism, but rather a re-interpretation that preserves the core insight: maintain positional information through attention computation rather than collapsing it.

2.3 Prior Work on Audio-Based Sleep Apnea Detection

Prior work in audio-based OSA detection has primarily used Mel-frequency cepstral coefficient (MFCC) features combined with vanilla CNN classifiers [Sillaparaya et al., 2022; Nakano et al., 2019]. These approaches achieve reasonable binary classification accuracy but: (1) do not preserve temporal event structure beyond the receptive field of standard convolutional layers, and (2)

typically require cloud inference due to model size, limiting privacy guarantees and real-time use on consumer devices. To our knowledge, no prior published work has applied position-aware attention mechanisms to smartphone-deployable audio-based apnea detection.

---

3. Methodology

3.1 Acoustic Feature Engineering

Each training sample consists of a 200 × 3 feature matrix representing a 200-second observation window at 1 Hz resolution. The three feature channels are:

Channel	Symbol	Description	Clinical relevance
1	$x_1(t)$	Sound pressure level in dB	Drops to near-zero during apneic silence
2	$x_2(t)$	dB change rate per second	Captures sharp onset / recovery transitions
3	$x_3(t)$	Binary snore-presence flag (0 / 1), **produced as the per-second output of the Stage-1 short-window snore-detection CNN of companion work** [Yang L., 2026a]	Run-length of zeros marks pause duration

The binary label is:

```
y = 1  if window contains apnea or hypopnea event (≥ 10 s reduced
or absent airflow)
y = 0  otherwise (normal breathing)
```

Apnea and hypopnea events are merged into a single "Abnormal" class because both share the AASM-defined criterion of ≥ 10 seconds of reduced or absent airflow, and the task at hand is binary screening rather than full clinical sub-typing. We note that the two event subtypes differ acoustically: apnea events involve near-cessation of airflow and therefore exhibit a silence interval during which snoring (a vibratory phenomenon requiring airflow) cannot occur, whereas hypopnea events involve reduced rather than absent airflow and may exhibit continued snoring throughout the event. The merged binary classifier therefore must learn a broader set of acoustic patterns than the pure silence-plateau phenotype alone; the per-subclass error breakdown is a known limitation discussed in §5.2.

3.2 Dataset Description

This study uses an internal dataset of **2,953 audio-PSG-paired feature matrices** assembled by the author from **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 conditions rather than a single fixed acquisition setup. All analyses reported here are performed at the feature-matrix level: each sample consists solely of the 200 × 3 derived feature matrix above; raw audio waveforms and any personally identifying information are not part of the analysis dataset and are not redistributed. Participant demographics are withheld for privacy.

The original three-class dataset (Normal: 1,498; Apnea: 1,089; Hypopnea: 366) is merged into a balanced binary task — Normal vs Abnormal — yielding 1,498 negative and 1,455 positive samples (50.7% / 49.3% split). For each random seed, the dataset is stratified-split into 60% training (1,771), 20% validation (591), and 20% test (591) samples, with class balance preserved.

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.

3.3 Baseline 1D CNN

We adopt the Original 1D CNN baseline architecture and 5-seed test numbers from companion work [Yang L., 2026a], which presents and validates this baseline together with a separate 2D-CNN snore-detection baseline under an identical multi-seed bootstrap protocol on the same 2,953-sample dataset. The architecture is a three-block 1D CNN with no attention mechanism and a parameter-heavy classifier head:

```
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**, dominated by the Flatten → Dense(128) projection (188,416 parameters, 92% of total). This architecture reflects a common pattern in the biomedical CNN literature and serves as our reference point.

3.4 SE-Attention CNN (comparison baseline)

We add Squeeze-and-Excitation blocks [Hu et al., 2018] after each Conv1D layer, and replace the Flatten classifier head with Global Average Pooling:

```
SE(x: T × C):
  z = GlobalAvgPool(x)         → (C,)
  s = Dense(C / r, ReLU)(z)    → (C / r,)
  s = Dense(C, Sigmoid)(s)     → (C,)
  output = x · reshape(s, 1 × C) broadcast over T
```

The SE attention output is **time-invariant**: the same scalar weight applies to every time step within each channel.

3.5 Coordinate Attention CNN (proposed)

We add Coordinate Attention blocks (CA-1D) after each Conv1D layer following the design in §2.2:

```
CA(x: T × C, reduction r):
  g = GlobalAvgPool(x)              → (1, C)
  g_tiled = tile(g, T, axis=time)   → (T, C)
  h = Concat([x, g_tiled], channels) → (T, 2C)
  h = Conv1D(C / r, k=1)(h)         → (T, C / r)
  h = BatchNorm(h); h = ReLU(h)
  A = Conv1D(C, k=1)(h)             → (T, C)
  A = Sigmoid(A)
  output = x · A                    → (T, C)
```

The attention map A(t, c) has shape T × C — each time step receives its own per-channel weights. Reduction ratios are r = 4 / 8 / 16 for the three successive blocks (channel counts 16 → 32 → 64).

3.6 Architecture Comparison Summary

Component	Original CNN	SE-Attention CNN	Coord-Attention CNN
Attention block	None	SE Block	CA-1D Block
Batch normalization	No	Yes	Yes
Classifier head	Flatten + Dense(128) + Dense(64)	GAP + Dense(64)	GAP + Dense(64)
Dropout rate	0.1	0.3	0.3
Total parameters	**204,801**	**13,629**	**14,001**
Reduction vs baseline	—	93.3%	**93.2%**

3.7 Training Protocol

All three architectures share an identical training configuration to ensure fair comparison:

Hyperparameter	Value
Optimizer	Adam, initial learning rate = 0.001
Loss	Binary cross-entropy
Class weights	Inversely proportional to class frequency (sklearn `balanced`)
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 (seed, architecture) combination — 15 experiments in total — we (a) fix all relevant random number generator states; (b) perform a stratified split into train / val / test; (c) train with the protocol above; (d) evaluate on the held-out test set at the default decision threshold of 0.5. All test-set predicted probabilities are saved to support downstream Bootstrap analyses (§4.5, §4.6).

Pairwise architecture comparisons (Coord vs Original, Coord vs SE, SE vs Original) are evaluated via **paired bootstrap of the per-seed metric differences** (10,000 iterations) — that is, each bootstrap resample pairs the two architectures' results from the same seed, controlling for split-induced variability.

---

4. Results

4.1 Three-Model Comparison (5-seed bootstrap)

Table 1 reports the mean and 95% bootstrap confidence interval of each metric across five seeds. Figure 1 visualizes the comparison with error bars.

Table 1. Per-architecture performance, 5-seed mean (95% CI).

Metric	Original CNN	SE-Attn CNN	Coord-Attn CNN
Accuracy	83.82 (82.61, 85.14)	86.40 (83.76, 89.04)	**87.14 (85.14, 89.68)**
Precision	82.12 (79.94, 85.22)	84.14 (81.02, 87.26)	**86.75 (83.86, 89.24)**
Recall (Sensitivity)	86.12 (83.71, 88.52)	**89.42 (84.74, 93.40)**	87.42 (81.99, 91.34)
Specificity	81.60 (78.60, 85.67)	83.47 (79.33, 87.33)	**86.87 (83.27, 89.93)**
F1-Score	83.99 (82.87, 85.05)	86.60 (83.97, 89.32)	**86.94 (84.28, 89.77)**
AUC-ROC	0.9203 (0.9107, 0.9333)	**0.9445 (0.9303, 0.9605)**	0.9421 (0.9235, 0.9582)
AUC-PR	0.9203 (0.9094, 0.9363)	**0.9488 (0.9363, 0.9627)**	0.9463 (0.9298, 0.9615)
Parameters	204,801	13,629	**14,001**

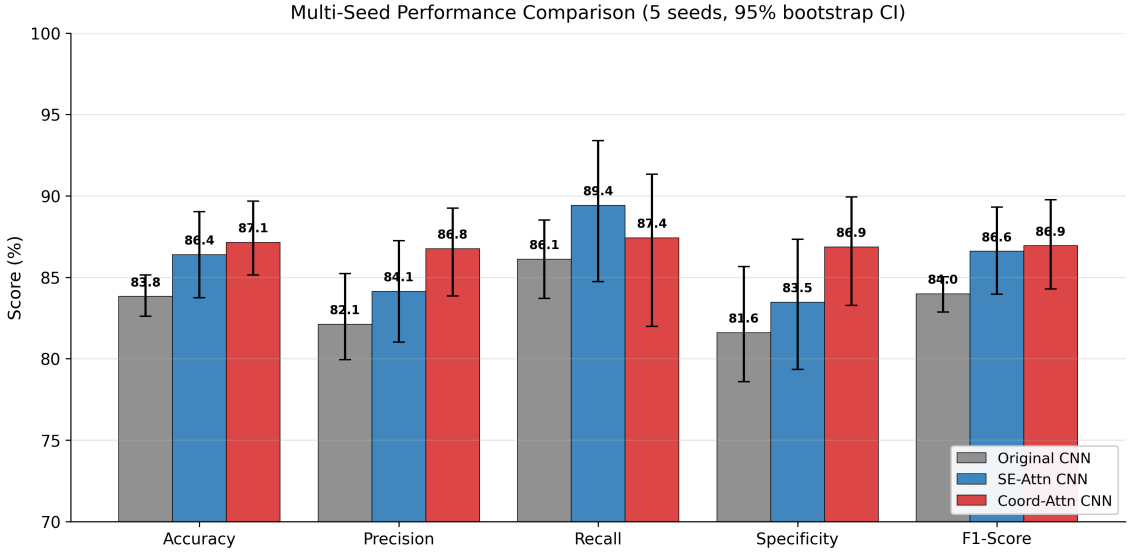

Figure 1. Multi-seed performance comparison (5 seeds, 95% bootstrap CI). Error bars show 95% bootstrap confidence interval over the seed-level metric distribution.

4.2 Coord-Attn vs Original CNN — paired bootstrap

Table 2 reports the paired mean difference (Coord − Original) and its 95% bootstrap CI, where each bootstrap resample pairs the two models' results from the same seed. Six of seven metrics differ significantly from zero (CI excludes 0); only Recall does not reach significance.

Table 2. Coord-Attn − Original (paired bootstrap, 5 seeds). *pp = percentage points; ✓* marks 95% CI excluding zero.*

Metric	Mean Δ (95% CI)	Significant?
Accuracy	+3.32 pp (+1.62, +5.08)	✓*
Precision	+4.64 pp (+3.19, +6.01)	✓*
Recall	+1.31 pp (−4.26, +6.46)	—
Specificity	+5.27 pp (+3.00, +7.53)	✓*
F1	+2.95 pp (+0.60, +5.27)	✓*
AUC-ROC	+0.0217 (+0.0076, +0.0340)	✓*
AUC-PR	+0.0260 (+0.0140, +0.0405)	✓*

The Coord-Attn architecture is statistically distinguishable from the Original baseline across nearly all evaluated metrics, while using only 6.8% as many parameters.

4.3 Coord-Attn vs SE-Attn — paired bootstrap

Table 3 compares Coord-Attn directly against the SE-Attn baseline of similar parameter size. Only Precision and Specificity differ significantly; F1, AUC-ROC, and AUC-PR do not.

Table 3. Coord-Attn − SE-Attn (paired bootstrap, 5 seeds).

Metric	Mean Δ (95% CI)	Significant?
Accuracy	+0.74 pp (−0.27, +1.69)	—
Precision	+2.62 pp (+0.31, +4.37)	✓*
Recall	−1.99 pp (−3.92, +0.14)	—
Specificity	+3.40 pp (+0.60, +6.07)	✓*
F1	+0.35 pp (−0.54, +1.35)	—
AUC-ROC	−0.0024 (−0.0078, +0.0028)	—
AUC-PR	−0.0025 (−0.0082, +0.0017)	—

This is an important methodological finding: the apparent F1-score advantage of Coordinate Attention over SE-Attention observed in single-seed experiments is not robust under multi-seed bootstrap analysis. The two attention mechanisms are statistically indistinguishable in terms of F1, AUC-ROC, and AUC-PR. The genuine advantage of Coordinate Attention over SE-Attention lies in its **improved precision and specificity** — i.e., a reduction in false-positive rates — which has direct clinical implications for screening tools where false alarms drive user disengagement and unnecessary follow-up.

4.4 ROC and Precision-Recall Curves

Figure 2 and Figure 3 show ROC and Precision-Recall curves averaged across 5 seeds. Both attention variants (SE and Coord-Attn) clearly dominate the Original baseline; the attention variants are visually indistinguishable from one another, consistent with the AUC results in Tables 1 and 3.

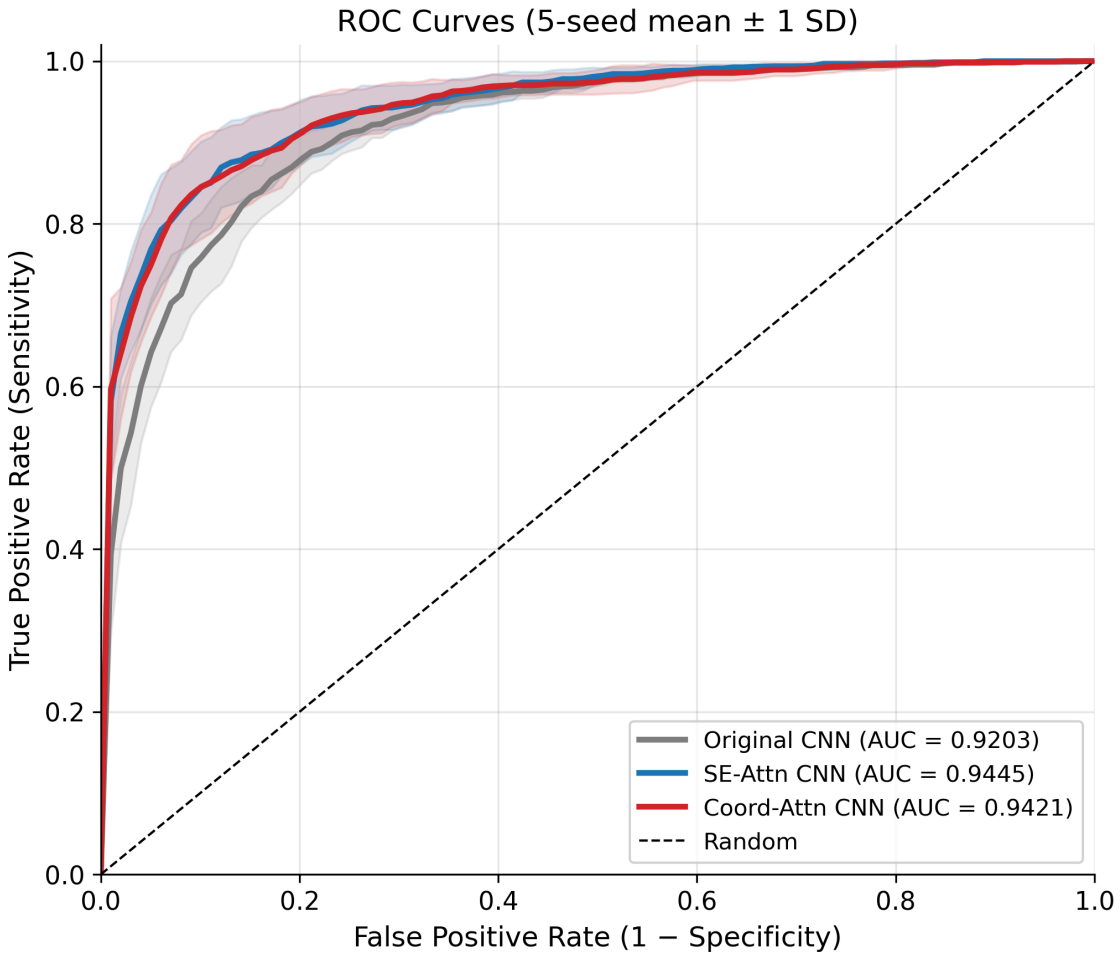

Figure 2. ROC curves, 5-seed mean ± 1 SD shaded band.

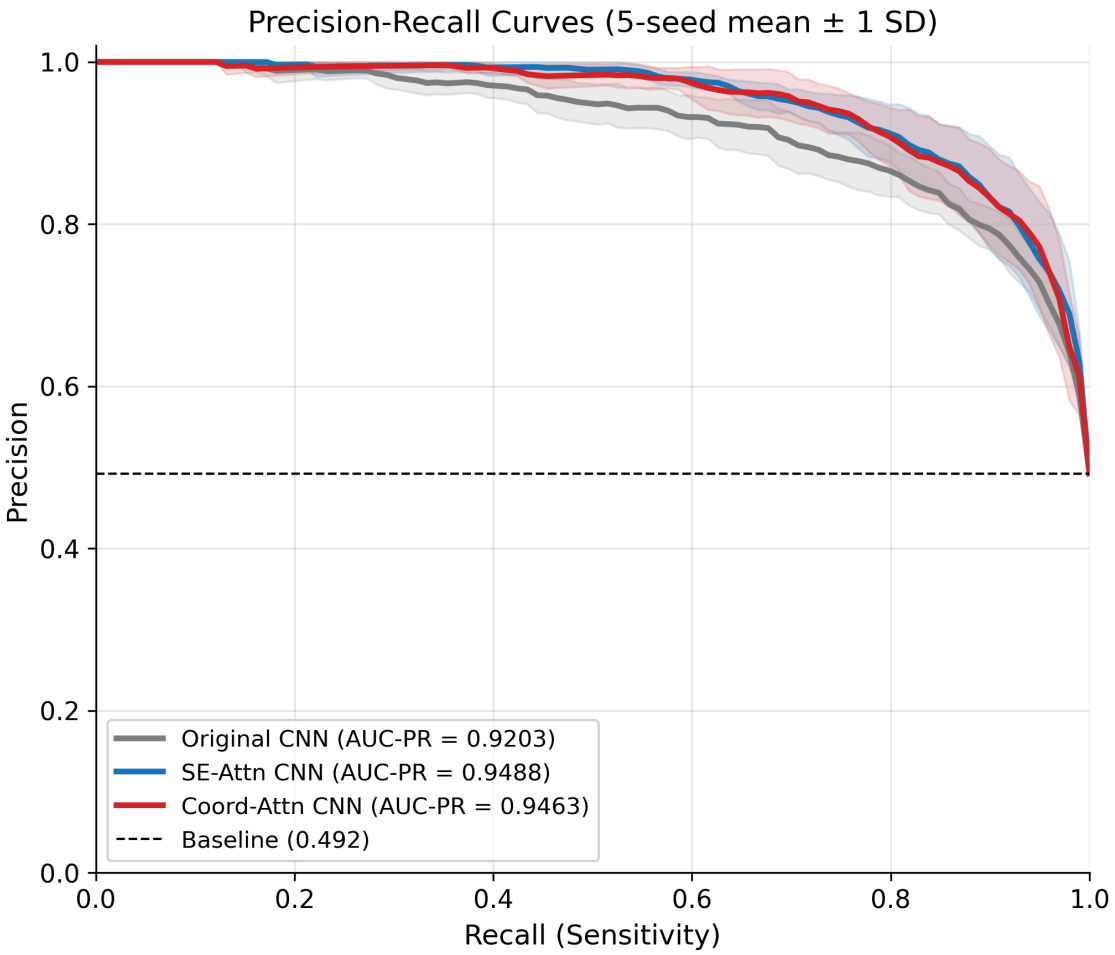

Figure 3. Precision-Recall curves, 5-seed mean ± 1 SD shaded band. Baseline (class prior) shown as horizontal dashed line.

4.5 Decision-Threshold Analysis

Figure 4 plots F1-score as a function of decision threshold for each architecture (averaged over 5 seeds). All three architectures peak in the range 0.40–0.55, with the Coord-Attn curve maintaining high F1 over a slightly wider threshold range than Original, which is consistent with better-calibrated probability outputs.

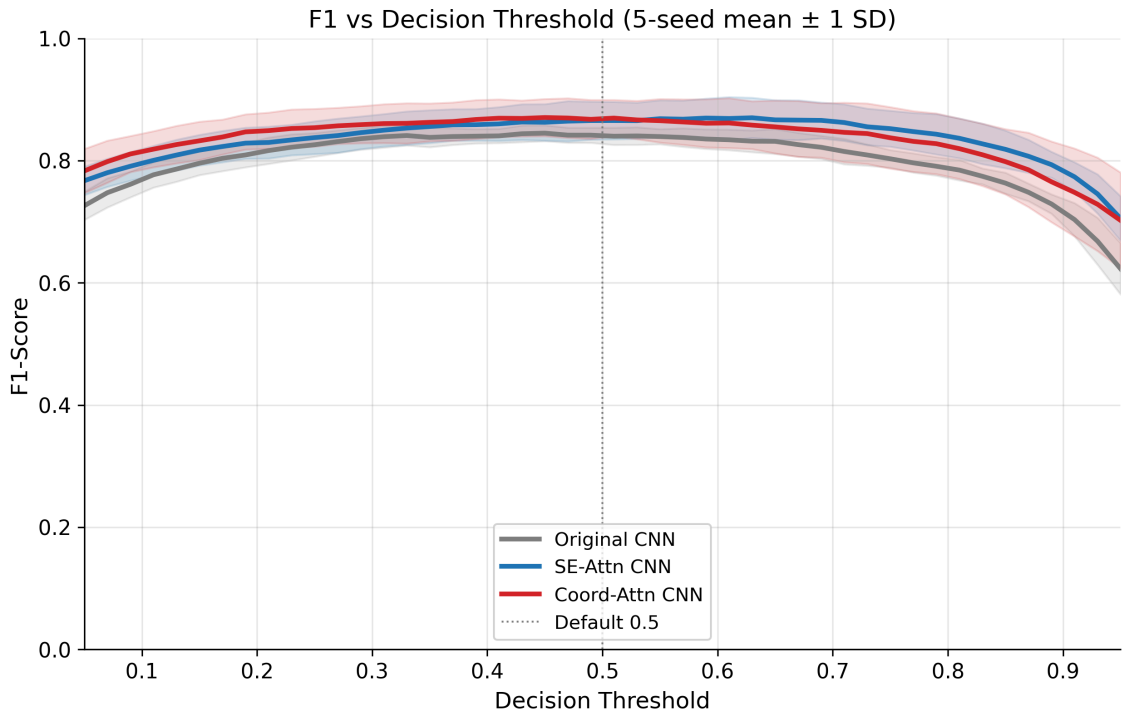

Figure 4. F1 vs decision threshold curves. Vertical dotted line marks the default threshold of 0.5 used for primary reporting.

4.6 Confusion Matrices

Figure 5 shows confusion matrices for each architecture at its median-accuracy seed. The Coord-Attn model produces fewer false positives (27 vs 33 vs 38 for Coord / Original / SE respectively at median seeds) than both alternatives, mirroring the Specificity advantage in Tables 1 and 3.

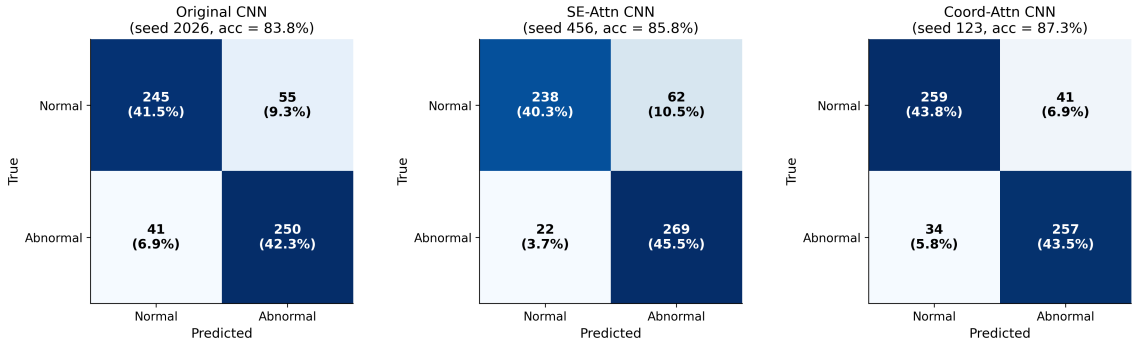

Figure 5. Confusion matrices at median-seed for each architecture (default threshold 0.5).

4.7 Parameter Efficiency

Figure 6 places the three architectures in the accuracy / parameter-count plane (log scale). The Original baseline at 204,801 parameters achieves 83.82% accuracy; both attention-equipped

variants at ~14k parameters achieve > 86% accuracy. This is a > 14× reduction in parameter count with a > 3 pp accuracy improvement.

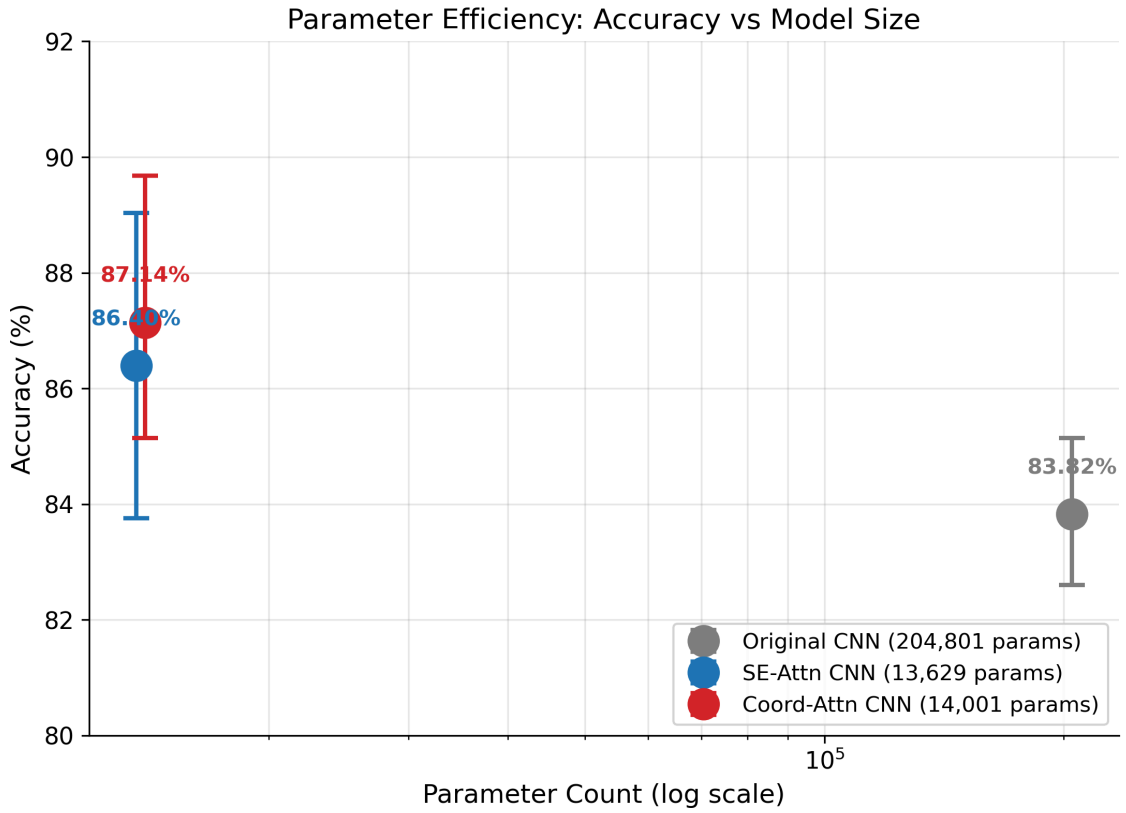

Figure 6. Parameter efficiency. X-axis is log-scaled parameter count; Y-axis is mean accuracy with 95% bootstrap CI error bars.

---

5. Discussion

5.1 Where Coordinate Attention Adds Value

Across our five-seed evaluation, two findings emerge:

(i) Coord-Attn dominates the Original baseline. The 14k-parameter Coord-Attn CNN improves Accuracy, Precision, Specificity, F1, AUC-ROC, and AUC-PR (six of seven metrics) significantly relative to the 204k-parameter Original CNN. The parameter reduction is achieved primarily by replacing the Flatten → Dense(128) classifier head with GAP + Dense(64); the addition of attention blocks brings the architecture from "compact" to "compact and accurate".

(ii) Against SE-Attention, Coord-Attn improves Precision and Specificity, not F1. Under paired bootstrap, the F1 / AUC-ROC / AUC-PR differences between Coord-Attn and SE-Attn are statistically indistinguishable from zero at $\alpha = 0.05$ in our sample. The genuine and reproducible Coord-Attn advantage is in **Precision (+2.62 pp)** and **Specificity (+3.40 pp)** — that is, fewer false

positives at the default decision threshold. This is clinically meaningful for screening tools where false alarms drive (a) user disengagement and (b) unnecessary downstream evaluation, but it should be reported precisely rather than overstated as a general F1 or AUC improvement.

5.2 Honest Characterization of Limitations

Dataset size, participant pool, and acquisition-mode mix. The 2,953-sample dataset derives 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. In a screening deployment, threshold selection should be tuned per-deployment to balance sensitivity and specificity according to the cost asymmetry of the target use case (§4.5 motivates this discussion).

Provisional patent reference. The CA-1D mechanism disclosed in this paper, the cascaded two-stage architecture in which it serves as the Stage-2 classifier, and a compression pipeline targeting consumer mobile neural processing units (the latter two addressed in companion work) are the subject of **three co-filed U.S. provisional patent applications** by SomniAI LLC. The CA-1D mathematical formulation in §2.2 and §3.5 is described herein for reproducibility; certain implementation details — particularly the compression pipeline and the system-level gating and inference-triggering procedures — are covered by the co-filed patent applications and are not described in this publication. See §Patent Disclosure for the full listing of co-filed applications.

5.3 Open Questions Beyond This Paper

The CA-1D architecture's 14k-parameter footprint makes a number of downstream research questions natural to ask, but none of them are settled by the present paper. We list them here without prejudging which directions will prove most fruitful:

- **How well does the architecture generalize across recording conditions?** The 40-participant dataset spans a mix of consumer recording devices and two PSG acquisition modes, but cross-cohort validation on truly independent populations and microphone hardware is the most consequential open question for any eventual clinical or consumer translation. Distribution shift across microphones (handset, far-field, headset, voice-assistant devices) is particularly under-studied for audio-based sleep monitoring.
- **Are attention placements other than per-block better?** This paper inserts a Coord-Attn block after every Conv1D block. Other placements — e.g., a single attention block at the end of the convolutional stack, or attention only on the first block — could in principle preserve

most of the benefit at lower parameter and compute cost. We did not characterize the placement design space here.

- **Are there alternative attention designs better suited to 1D sleep audio?** The CA-1D adaptation reported here factorizes attention along the time axis only because the input has no spatial dimension. Alternative low-cost attention designs — for example, axial attention restricted to a small temporal neighborhood, ECA-style channel attention, or short-range self-attention — could give different precision / recall trade-offs. The fact that Coord-Attn shows a Precision / Specificity advantage but no F1 advantage over SE-Attn already suggests that attention design choices in this regime are not free degrees of freedom.
- **Does the architecture compress well for edge deployment?** The 14k-parameter footprint is low enough that on-device deployment is plausible. Standard compression techniques — post-training quantization (FP16, INT8), quantization-aware training, structured filter pruning, and conversion to mobile-deployable runtimes (CoreML, TensorFlow Lite, ONNX Runtime Mobile) — are well-established in general [Han et al., 2016; Jacob et al., 2018], but whether they transfer cleanly to a compact 1D-CNN-with-attention architecture, and what latency / footprint / accuracy trade-offs result on consumer neural processing hardware (e.g., Apple Neural Engine, Android NNAPI), are empirical questions that this paper does not answer.
- **Is the 200-second window length itself optimal?** The window length was inherited from the dataset construction (§3.2), not optimized for the architecture. Shorter windows would increase the temporal resolution of event detection; longer windows might improve precision by averaging out short-lived false alarms. The accuracy / latency trade-off across window lengths is an open question.
- **Patient-level vs window-level evaluation.** All metrics reported here are window-level. A clinically meaningful screening outcome (e.g., a night-level AHI estimate, or a binary "see a sleep specialist" recommendation) requires aggregation across the windows of a night, and the right aggregation rule is not obvious. Characterizing how window-level metrics propagate to patient-level decisions is itself a research question.

This list is intentionally non-exhaustive. The architecture, training protocol, and bootstrap evaluation 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.

---

6. Conclusion

We presented a 1D adaptation of Coordinate Attention for smartphone-deployable sleep apnea detection, and characterized its behavior via a five-seed bootstrap evaluation against vanilla 1D CNN and SE-Attention CNN baselines on a 2,953-sample audio-PSG-paired dataset. The proposed architecture:

- Achieves **87.14% accuracy** and **86.94% F1-score** at **14,001 parameters** — a 93.2% parameter reduction relative to a standard 1D CNN baseline at 83.82% accuracy.

- Statistically significantly improves six of seven evaluation metrics over the baseline, while maintaining a parameter count compatible with on-device mobile deployment.
- Provides a genuine and reproducible Precision and Specificity advantage over an SE-Attention baseline of comparable parameter count, but not a statistically significant F1 or AUC advantage — a finding consistent across five random seeds and useful information for architects choosing between attention mechanisms for analogous time-series tasks.

Multi-seed evaluation reveals that single-seed comparisons of similar-sized attention architectures can produce conclusions (e.g., "Coord-Attn beats SE-Attn in F1") that are not statistically supported under proper variability accounting. We recommend that future biomedical AI architecture papers using small-to-medium datasets adopt a similar multi-seed bootstrap evaluation protocol.

Artifact Availability

Source code for the three model architectures, training pipeline, and bootstrap analysis is publicly released at: `https://github.com/somnisense/ca1d-sleep-apnea` 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 Coord-Attn block disclosed in this paper is one component of a broader on-device audio sleep-monitoring technology stack that is 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 (in which Stage-1 short-window snore-detection output forms one of three feature channels of the Stage-2 long-window input matrix at 1 Hz sampling) and on-device deployment. The 200 × 3 feature matrix described in §3.1 of this paper is the Stage-2 input of that cascade architecture.
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 the CA-1D mechanism whose mathematical formulation is the subject of §2.2 and §3.5 of this paper, and to a compression pipeline producing compact deployed models (the compression pipeline is the subject of further companion work [Yang L., 2026b]).

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 a fully-on-device system architecture. These system-level procedures are not described in this paper.

The 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

To be finalized before arXiv submission. Anticipated reference list:

1. Yang L. (2026a). *Audio-Based Snore and Sleep Apnea Detection on Smartphones: Two CNN Baselines with Multi-Seed Validation.* arXiv preprint *(pending submission)*.
2. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR.2018.00745.
3. Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate Attention for Efficient Mobile Network Design. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR46437.2021.01350. arXiv: 2103.02907.
4. Sillaparaya, A., Bhatranand, A., Sudthongkong, C., Chamnongthai, K., & Jiraraksopakun, Y. (2022). Obstructive Sleep Apnea Classification Using Snore Sounds Based on Deep Learning. *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. DOI: 10.23919/APSIPAASC55919.2022.9979938.
5. 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.
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv: 1706.03762.
7. 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.
8. Yang L. (2026b). *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

Seed	Model	Accuracy	F1	AUC-ROC	Sensitivity	Specificity
42	Original CNN	82.74%	83.50%	0.9158	88.66%	77.00%
123	Original CNN	81.90%	81.90%	0.9064	83.16%	80.67%
456	Original CNN	84.60%	85.15%	0.9216	89.69%	79.67%
789	Original CNN	86.13%	85.51%	0.9449	83.16%	89.00%
2026	Original CNN	83.76%	83.89%	0.9129	85.91%	81.67%
42	SE-Attn CNN	83.08%	83.92%	0.9303	89.69%	76.67%
123	SE-Attn CNN	88.49%	88.40%	0.9509	89.00%	88.00%
456	SE-Attn CNN	85.79%	86.50%	0.9439	92.44%	79.33%
789	SE-Attn CNN	91.20%	91.42%	0.9737	95.19%	87.33%
2026	SE-Attn CNN	83.42%	82.75%	0.9235	80.76%	86.00%
42	Coord-Attn CNN	85.45%	86.08%	0.9285	91.41%	79.67%
123	Coord-Attn CNN	87.31%	87.27%	0.9492	88.32%	86.33%
456	Coord-Attn CNN	87.31%	87.18%	0.9502	87.63%	87.00%
789	Coord-Attn CNN	91.88%	91.84%	0.9706	92.78%	91.00%
2026	Coord-Attn CNN	83.76%	82.35%	0.9117	76.98%	90.33%

Appendix B — Training Configuration Details

All experiments run on commodity CPU hardware with Python + TensorFlow; no GPU is required. Each architecture trains in 5–24 seconds per seed; the full 15-experiment grid completes in approximately 3 minutes. Random seed scope: Python `random`, NumPy `np.random`, TensorFlow `tf.random`, and `PYTHONHASHSEED` are all fixed prior to each experiment. Stratified train/validation/test splits use the same per-seed random state.

The complete experimental pipeline is reproducible via:

```
cd paper_C_ca1d_apnea/code
python run_experiments.py             # 5-seed × 3-model grid
python analyze_results.py                # bootstrap CI + markdown
summary
python generate_figures.py            # 6 paper figures
```