electronics

MDPI

Article

BCough: Design and Evaluation of a Bone-Conduction-Embedded AI Platform for On-Device Cough Detection

Mayur Sanap [1,*], Joseph de la Viesca [1], Aadesh Shah [1], Sameer Dalal [1], Jack Twiddy [2], Michael Daniele [2] and Edgar J. Lobaton [1]

[1] Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695, USA; jcdelavi@ncsu.edu (J.d.l.V.); atshah@ncsu.edu (A.S.); stdalal@ncsu.edu (S.D.); ejlobato@ncsu.edu (E.J.L.)
[2] Department of Biomedical Engineering, North Carolina State University, Raleigh, NC 27695, USA; jstwiddy@ncsu.edu (J.T.); mdaniel6@ncsu.edu (M.D.)
* Correspondence: msanap@ncsu.edu

Abstract

Continuous cough monitoring provides valuable insights into respiratory health; however, conventional air-conduction microphones are highly susceptible to ambient noise and raise privacy concerns. This work presents BCough, a bone-conduction-based embedded AI platform for on-device cough detection, designed and evaluated on the MAX78000 neural accelerator. The system employs a skin-contact bone-conduction sensor worn on the chest to capture body vibrations transmitted through bone and tissue, detecting cough events while minimizing environmental interference. The custom prototype integrates a bone-conduction microphone, a synchronized 6-axis IMU, power management circuitry, and an embedded neural accelerator to support real-time inference and future multimodal extensions. A compact 8-bit quantized convolutional neural network was optimized for deployment on the MAX78000 and evaluated using leave-one-subject-out cross-validation on one-second cough and non-cough segments derived from a corpus of 19,955 labeled events collected from five participants under controlled conditions. The deployed model achieved 0.80 Macro-F1, 0.81 balanced accuracy, 0.74 cough F1, and 0.89 AUC, with 15–16 ms inference latency and approximately 20 μJ energy per inference on chip. These results demonstrate the feasibility of low-power, privacy-preserving, bone-conduction cough detection on embedded AI hardware within an initial five-participant study. The current design is a benchtop prototype; the findings should therefore be interpreted as an initial feasibility assessment rather than evidence of robust performance across diverse users and real-world conditions. Future work will extend this platform toward miniaturized wearable implementations combining bone-conduction and inertial sensing for continuous multimodal respiratory monitoring.

Keywords: bone conduction; embedded AI; cough detection; embedded neural accelerator; MAX78000; privacy-preserving sensing; edge inference

Academic Editor: Ping-Feng Pai

Received: 1 April 2026
Revised: 22 April 2026
Accepted: 27 April 2026
Published: 1 May 2026

1. Introduction

Respiratory diseases, such as asthma and chronic obstructive pulmonary disease (COPD), remain among the leading causes of morbidity and mortality worldwide [1]. Coughing is one of the most frequent and clinically significant symptoms associated with these conditions, often serving as an indicator of disease progression and treatment effectiveness [2]. Accurate quantification of cough frequency provides valuable diagnostic

and prognostic insights, but traditional assessments rely primarily on patient self-reporting or short-term acoustic recordings [3]. These approaches are subjective and often fail to reflect true cough frequency in daily life. Moreover, continuous clinical monitoring setups are expensive, require bulky equipment, and are impractical for long-term home-based use. As a result, there is an increasing need for automated, objective, and privacy-preserving cough detection systems capable of continuous operation in real-world environments.

Earlier research in automated cough detection primarily utilized acoustic features and traditional classifiers, such as Hidden Markov Models and Support Vector Machines [4–6]. Several commercial devices have been developed for cough monitoring, including the Leicester Cough Monitor [3], VitaloJAK™ [7], and LEOSound® [8]. These systems have demonstrated reliable cough event detection in controlled environments but depend on long-term audio recordings processed offline, raising privacy and security concerns that may limit user comfort and adoption. More recently, wearable systems such as CoughWatch [9], CoughBuddy [10], and HearCough [11] have emerged, integrating microphones and inertial measurement units into smartwatches or earbuds. While these devices support ambulatory monitoring, they frequently rely on cloud-based computation or smartphone tethering, which increases energy consumption and latency, limiting their suitability for continuous daily monitoring.

To address these limitations, embedded AI platforms have gained traction as a means of performing on-device inference on constrained hardware [12–14]. Processing signals locally reduces the need for continuous data transmission and thereby minimizes latency, network dependency, and privacy risks. Recent studies have demonstrated lightweight models for on-device inference. For instance, HearCough [11] introduced Tiny-COUNET, a compact convolutional network that achieved 90% detection accuracy with a power draw of only 5.2 mW when deployed on commercial earbuds. More recently, Cough-E [15] proposed a multimodal, privacy-preserving cough detection algorithm combining acoustic and kinematic features, deployed on edge hardware with hardware-aware model design, and Orlandic et al. [16] contributed the first publicly available multimodal edge-AI cough dataset covering 15 subjects with synchronized audio and kinematic modalities. Despite these advances, most cough detection systems still depend on deep neural networks with large parameter counts, which often exceed the Flash and SRAM limits of ultra-low-power microcontrollers, such as the MAX78000 (Analog Devices, Inc., Wilmington, MA, USA) and STM32G4 (STMicroelectronics, Geneva, Switzerland) [17,18]. Furthermore, neither HearCough nor Cough-E targets bone-conduction sensing or the MAX78000 neural accelerator, leaving the co-design of bone-conduction-based sensing and hardware-aware embedded AI inference as an open challenge.

Bone-conduction (BC) microphones detect internal body vibrations transmitted through bone and tissue rather than airborne compression waves detected by conventional air-conduction (AC) microphones. Unlike AC microphones, BC sensors are inherently robust to environmental noise and preserve privacy by excluding bystander speech and background sounds. Moreover, BC microphones capture fewer spectral features of speech due to their limited bandwidth, further reducing the amount of personally identifiable acoustic information captured during monitoring [19]. Recent work has demonstrated the practical value of BC sensing in wearable applications beyond speech: He et al. [20] showed that bone-conducted vibrations captured by head-mounted wearables can support robust speech enhancement in real-world noisy environments, and Liu et al. [21] demonstrated that PMUT-based BC sensors achieve signal-to-noise ratios over five times greater than conventional AC microphones in high-noise settings. These advances confirm the growing maturity of BC sensing as a practical modality for privacy-preserving wearable health monitoring.

When combined with inertial measurement units (IMUs), BC microphones provide complementary motion cues corresponding to chest and neck movements during a cough [22,23], which can improve detection reliability in noisy or mobile conditions typical of daily life. Prior work has shown that IMU signals alone can serve as a privacy-preserving alternative to audio for cough detection: Diab and Rodriguez-Villegas [24] demonstrated that tri-axial accelerometer features achieve F1-scores above 88% under LOSO evaluation, confirming the discriminative value of inertial sensing for respiratory event classification. Prior work from our group has further demonstrated that combining air-conduction audio with IMU signals yields robust cough detection across diverse real-world conditions. Chen et al. [23] showed that an audio–IMU multimodal system enhanced with out-of-distribution (OOD) detection achieves over 90% accuracy at 16 kHz and maintains strong performance even at 750 Hz, where audio privacy is preserved. A follow-up study [25] further confirmed that inertial signals capture cough-induced body motion with sufficient discriminability to complement acoustic features across window sizes from 1 to 5 s. These results establish the IMU as a reliable sensing modality for wearable cough detection.

However, prior work has largely focused on offline multimodal analysis or smartphone-based computation using air-conduction microphones, without exploring hardware–software co-design for embedded, low-power execution. The present work extends this line of research by replacing air-conduction audio with bone-conduction sensing and targeting deployment on a microcontroller-class neural accelerator, a co-design challenge not previously addressed in this context. Thus, the integration of BC sensing and hardware-aware AI inference remains an open challenge for practical, on-device respiratory health monitoring.

In this paper, we present the design, implementation, and evaluation of BCough, a bone-conduction-based embedded AI system for cough detection on the MAX78000 microcontroller. The primary contribution of this work is an end-to-end co-design pipeline that integrates bone-conduction sensing, a hardware-constrained CNN, post-training quantization, and deployment on the MAX78000 neural accelerator within a unified framework. Supporting components of this system include a custom hardware prototype, a multimodal dataset, and a systematic design-space exploration. An overview of the full design and evaluation framework is illustrated in Figure 1. A comparison of closely related edge-deployed cough detection systems is provided in Section 7. Specifically, this work makes the following contributions, where the first item constitutes the primary contribution and the remaining items are supporting components:

- BCough: a bone-conduction-based embedded AI platform for privacy-preserving cough detection on the MAX78000 neural accelerator, incorporating a lightweight CNN achieving 0.80 Macro-F1, 0.81 balanced accuracy, 0.74 cough F1, and 0.89 AUC after 8-bit post-training quantization, with 15–16 ms inference latency and approximately 20 µJ energy per inference, evaluated in an initial five-participant feasibility study.
- A post-training quantization pipeline with scaling-based clipping that balances model accuracy, memory footprint, and energy efficiency for microcontroller-class deployment.
- A systematic design-space exploration across input resolution, sampling rate, feature representation, and quantization strategy to characterize the accuracy-efficiency trade-off for embedded cough detection.
- A custom hardware prototype integrating a bone-conduction microphone, audio codec, IMU, and SD-card interface for synchronized multimodal data acquisition and on-device validation.
- A multimodal dataset of 19,955 labeled events collected from five participants under controlled stationary and ambulatory conditions.

The rest of this paper is organized as follows. Section 2 describes the data collection procedure and the datasets used for training and evaluation. Section 3 presents the char-

acteristics of bone-conduction sensing, including its privacy-preserving properties and comparison with air-conduction audio. Section 4 outlines the embedded AI model design, including the CoughNet architecture, input configurations, and quantization pipeline. Section 5 describes the training and evaluation workflow, including the multi-stage pipeline, design-space exploration, and final model configuration. Section 6 discusses the embedded deployment on the MAX78000, including model selection, synthesis, and on-device performance. Section 7 presents the discussion, including a comparison with prior edge-deployed systems, BC vs. AC modality trade-offs, and the limitations of the present study. Finally, Section 8 concludes the article with a summary of key findings and directions for future work.

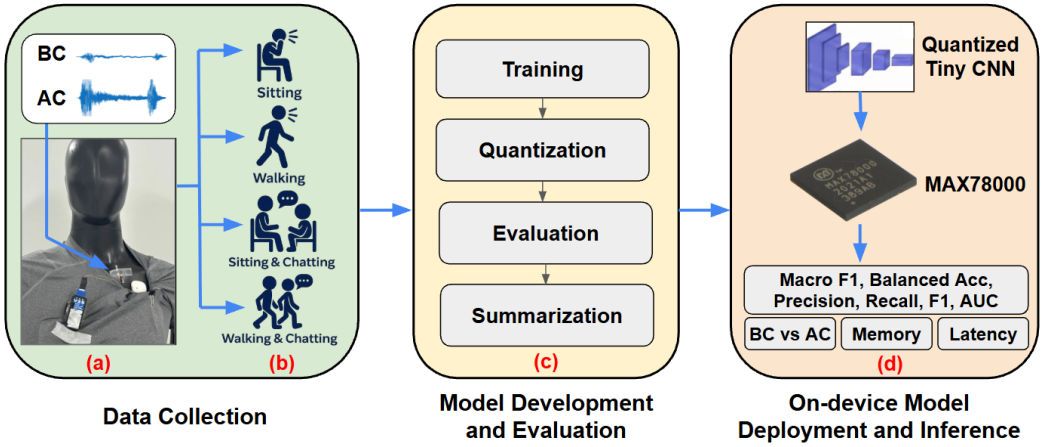


Figure 1. Overview of the BCough design and evaluation framework. (**a**) Wearable sensing platform with BC microphone and IMU. (**b**) Data collection protocol across four session conditions. (**c**) Model development and evaluation pipeline, including training, quantization, evaluation, and summarization stages. (**d**) On-device deployment on the MAX78000 neural accelerator with evaluation metrics. Blue arrows indicate the sequential flow from data collection through model development to on-device deployment and inference.

2. Data Collection

A custom bone-conduction cough detection dataset was collected using an on-body sensing platform integrating a bone-conduction vibration sensor, an air-conduction MEMS microphone, and a 6-axis inertial measurement unit (IMU). The data collection was performed under the approved NC State University Institutional Review Board (IRB) protocol 28453.

The objective of this human study was to capture multiple sensing modalities for the detection of speech and non-verbal vocalizations from individuals in both sedentary (sitting) and active (walking) states. The wearable data acquisition setup, shown in Figure 2, positions the bone-conduction sensor and IMU on the chest to capture vibration and motion signals during natural coughing. Table 1 summarizes the recorded signals and their corresponding sampling rates. Raw audio was captured at 48 kHz and downsampled to 16 kHz using a standard anti-aliasing filter prior to feature extraction. No dedicated noise reduction or denoising was applied to the bone-conduction channel, as the BC sensor is inherently robust to ambient acoustic noise by design. The apparent background activity visible in the BC waveforms reflects low-amplitude contact vibrations captured by the sensor, which are substantially lower in energy than cough transients. IMU data were recorded at 100 Hz with synchronized timestamps for future multimodal extensions.

Each 10 min session followed a structured protocol to capture diverse cough and non-cough activities under controlled and semi-natural conditions. The session included baseline speech followed by alternating one-minute segments of coughing, sneezing, throat clearing, arm or shoulder movement, and clothing friction with and without speech. This

design ensured a balanced mix of acoustic and motion interferences relevant to real-world cough detection.

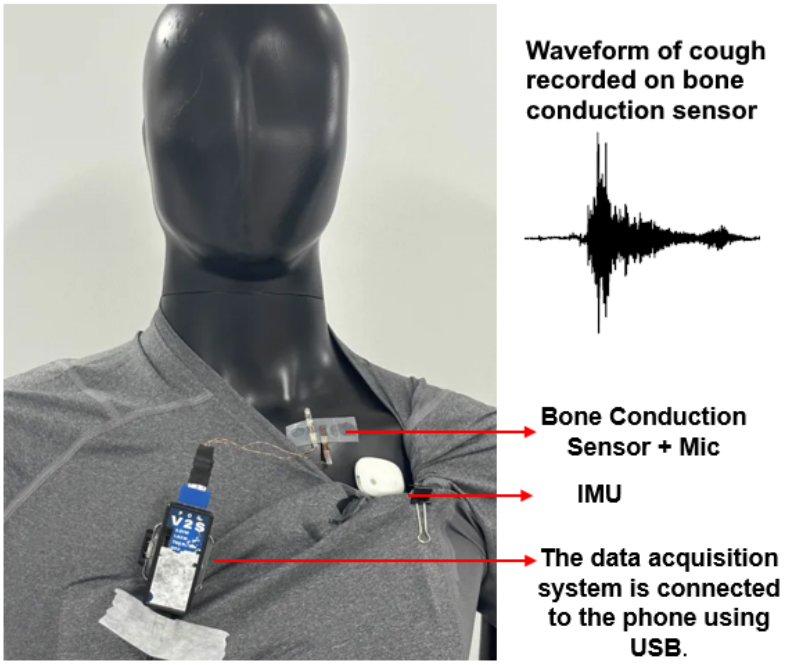


Figure 2. Wearable data acquisition setup used for synchronized bone-conduction and inertial sensing. The bone-conduction sensor and IMU are mounted on the chest, and the system connects to a smartphone via USB for synchronized data logging.

The audio front-end employed a V2S200D (Syntiant Corp., Irvine, CA, USA) bone-conduction sensor [26] and a SPH0141LM4H-1 (Knowles Corporation, Itasca, IL, USA) MEMS microphone [27] connected via a PDM-to-USB interface streaming stereo audio (16-bit, 48 kHz). The left channel recorded vibration data, and the right channel captured airborne audio. In parallel, a MetaMotionS (MbientLab Inc., San Francisco, CA, USA) sensor [28] recorded 3-axis accelerometer and 3-axis gyroscope data at 100 Hz, providing synchronized motion information for potential multimodal fusion.

Table 1. Recorded signals and their sampling rates.

Sensing Modality	Recorded Signals	Sampling Rate
Acoustic	Bone-conduction microphone (V2S200D), air-conduction MEMS microphone	16 kHz
Kinematic	Accelerometer (X, Y, Z), gyroscope (yaw, pitch, roll)	100 Hz

Data were collected from five participants (P1–P5), each recorded under four experimental conditions (S1–S4) designed to capture variations in body motion, environmental context, and conversational interference. As summarized in Table 2, the protocol includes stationary and walking conditions, as well as speaking-alone and conversation settings. In each session, participants performed a common set of engagement tasks consisting of intermittent coughs, coughs mixed with sneezing, coughs mixed with throat clearing, coughs with clothing rubbing, and coughs with arm/shoulder movement. Across all participants and sessions, the full corpus contains 19,955 labeled events. The data collection setup and session conditions are summarized in Figure 1a,b.

The present study was scoped as an initial hardware-deployment feasibility investigation. The cohort of five participants was selected to enable systematic data collection across all four session conditions under the approved IRB protocol (28453) within the constraints

of a benchtop prototype validation stage. This sample size is consistent with early-stage feasibility studies in closely related embedded and wearable health sensing work; for example, Orlandic et al. [16] collected a structured edge-AI cough dataset from 15 subjects in a similarly controlled laboratory setting, and the Cough-E study [15] used 20 healthy subjects for initial model development and validation. The present study represents an earlier stage of this trajectory, and data collection is ongoing under the same IRB protocol with a target expansion to 12–15 participants to improve generalization across diverse users and real-world conditions. The structured session protocol was designed to maximize the diversity of acoustically relevant events, including speech, throat clearing, sneezing, clothing friction, and motion artifacts, within a controlled acquisition setting, ensuring that the dataset captures the primary non-cough confounders relevant to real-world wearable deployment while remaining reproducible across participants. This design prioritizes signal diversity and controlled labeling accuracy over naturalistic cough occurrence rates, which is appropriate for a feasibility study focused on hardware–software co-design validation rather than population-level generalization.

Table 2. Recording session protocol used for multimodal cough data collection.

Session	Condition	Participant Engagement Tasks
S1	Stationary, speaking alone	(1) Random intermittent cough (∼10 s total); (2) cough for 10 s + sneezing for 10 s; (3) cough for 10 s + throat clearing for 10 s; (4) cough with clothing rubbing; (5) cough with arm/shoulder movement
S2	Stationary, conversation	Same engagement tasks as S1
S3	Walking, speaking alone	Same engagement tasks as S1, performed under outdoor walking conditions
S4	Walking, conversation	Same engagement tasks as S1, performed during outdoor conversation

For binary cough detection, the dataset used for model development and evaluation is derived from the larger corpus using leave-one-subject-out (LOSO) cross-validation. In each fold, the held-out participant forms the test set and the remaining four participants form the training set. All cough segments are retained in both splits. Non-cough segments are drawn from acoustically relevant events researcher speech, participant speech, throat clearing, sneezing, arm-motion artifacts, and clothing rubbing and selected independently for each split using the following three-stage procedure.

Non-cough samples with a bone-conduction RMS below 0.001 are removed in a first filtering step. This threshold acts as a noise-floor gate, discarding near-silent frames that carry no discriminative signal; it does not selectively remove acoustically challenging non-cough events. The threshold value of 0.001 was determined empirically through a systematic evaluation of multiple candidate thresholds, including 0.1, 0.005, and 0.001. For each candidate, the retained samples were manually verified through auditory inspection to confirm that segments above the threshold contained perceptible acoustic events, such as coughs, speech, throat clearing, and rubbing rather than sensor noise or near-silent contact artifacts. A threshold of 0.001 was selected as it maximized the number of retained samples while reliably excluding near-silent frames that carried no discriminative acoustic content.

The remaining non-cough samples within each label category are then ranked by BC RMS in descending order, retaining the most acoustically prominent segments. Samples are subsequently selected across label categories using a round-robin procedure that prioritizes underrepresented labels at each step: at each iteration, the label with the fewest selected samples so far is chosen, and among equally underrepresented labels, the sample with the highest BC RMS is selected. This continues until the total non-cough count reaches a target

2:1 non-cough-to-cough ratio. This procedure ensures both that the selected non-cough pool is acoustically diverse across event types and that the most discriminatively challenging samples those with the strongest BC contact signal are retained.

The 2:1 balancing is applied separately and independently to the training and test splits within each LOSO fold so that the test set for each held-out participant also maintains a 2:1 non-cough-to-cough ratio rather than reflecting the participant's natural event distribution. This ratio was chosen to reflect a tractable class imbalance for an initial feasibility study while still requiring the classifier to distinguish cough from a diverse set of non-cough events. It should be noted that this ratio simplifies the classification task relative to true continuous monitoring scenarios, where non-cough events occur far more frequently than cough events. This constitutes a known limitation of the present study and will be addressed in future work through sliding-window evaluation under more realistic class distributions.

Table 3 summarizes the participant-wise composition of the sampled dataset used by the classifier.

Table 3. Participant-wise sampled dataset used by the binary cough detection pipeline after exclusions, BC-RMS filtering, and 2:1 non-cough/cough balancing. Bold values in the last row indicate the total count across all participants.

Participant	Cough	Arm Move	Researcher	Rubbing	Sneeze	Speech	Throat Clear	Total NC
P1	254	113	112	113	20	113	37	508
P2	220	118	17	119	30	118	38	440
P3	137	51	51	52	17	52	51	274
P4	230	101	101	102	24	101	31	460
P5	204	83	82	83	36	83	41	408
Total	**1045**	**466**	**363**	**469**	**127**	**467**	**198**	**2090**

Overall, the dataset provides a controlled yet realistic benchmark for bone-conduction-based cough detection, capturing variability in vocalization, body motion, and environmental conditions while remaining compatible with low-power embedded AI systems. However, given the limited number of participants and the structured acquisition protocol, the results derived from this dataset should be interpreted as an initial feasibility assessment rather than a comprehensive evaluation of real-world generalization.

3. Bone-Conduction Sensing Characteristics

Figure 3 illustrates representative samples of air-conduction (AC), bone-conduction (BC), and inertial measurements recorded during cough (left) and speech (right). The signals were captured using off-the-shelf sensors placed on the sternum, a location chosen for its ability to capture both acoustic and physiological vibrations, though the proposed framework is generalizable to other on-body placements [29].

The AC signal magnitude alone is insufficient to reliably distinguish between the speaker and nearby bystanders, whereas the inertial signals exhibit distinct peaks corresponding to cough events. Speech, by contrast, produces smoother, lower-amplitude motion patterns. The BC microphone demonstrates strong robustness to environmental noise, activating clearly only during user-relevant speech or cough events, which confirms its suitability for body-worn acoustic detection [21]. This selectivity is the key practical advantage of BC sensing for privacy-preserving respiratory monitoring: because the sensor couples mechanically to the body surface, it attenuates bystander speech and ambient sounds that would otherwise contaminate an air-conduction recording.

Figure 4 compares the spectral and temporal characteristics of AC and BC signals recorded during (a) cough and (b) speech. The top row presents time–domain waveforms normalized to the full dynamic range of their respective sensors. During speech, both AC

and BC signals show amplitude variations corresponding to spoken syllables, while during coughs, short, high-energy transients dominate the waveform. The BC waveform displays less ambient background activity and clearer temporal isolation of individual cough events compared with the AC waveform, reflecting the mechanical coupling of the sensor to the body surface. The waveforms also illustrate the characteristic amplitude attenuation of BC sensing relative to AC sensing. This attenuation is a direct consequence of transmitting vibrations through tissue rather than air, and represents the fundamental trade-off of the BC modality: improved noise resilience at the cost of reduced signal amplitude and bandwidth.

The bottom row presents spectrogram representations of the same signals. Speech spectrograms show structured harmonic bands and temporal continuity, while cough spectrograms contain wideband, short-duration energy bursts. The BC spectrogram maintains high signal contrast even under noisy acquisition conditions, further confirming the modality's robustness to external interference.

The highlighted regions in Figure 4 reveal significant attenuation of BC audio in the 2–6 kHz band relative to AC. This effect is most pronounced for speech (Figure 4b), where spectral features in this band are clearly captured by the AC microphone but are nearly absent from the BC spectrogram. The BC spectrogram also shows elevated spectral energy near 7 kHz, attributed to the resonant frequency of the V2S200D sensor (Figure 5). Audio recorded via bone conduction generally exhibits significant high-frequency attenuation, with the extent of attenuation depending on the frequency response of the specific BC sensor [30]. As a result, BC recordings contain substantially less frequency information than AC recordings of the same event [19]. Despite this reduced spectral bandwidth, cough detection remains feasible from BC signals alone, as illustrated in Figure 3. Bone conduction therefore provides a minimal-information, privacy-preserving sensing modality for respiratory event detection: it captures sufficient discriminative information for cough classification while naturally suppressing the rich acoustic detail that would raise privacy concerns in conventional air-conduction recordings.

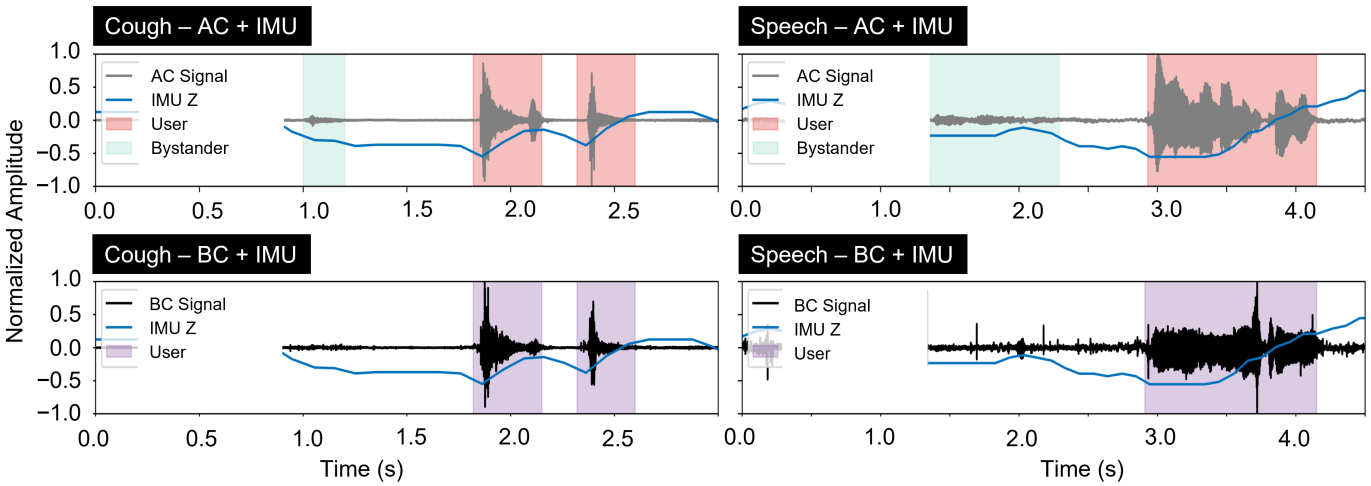
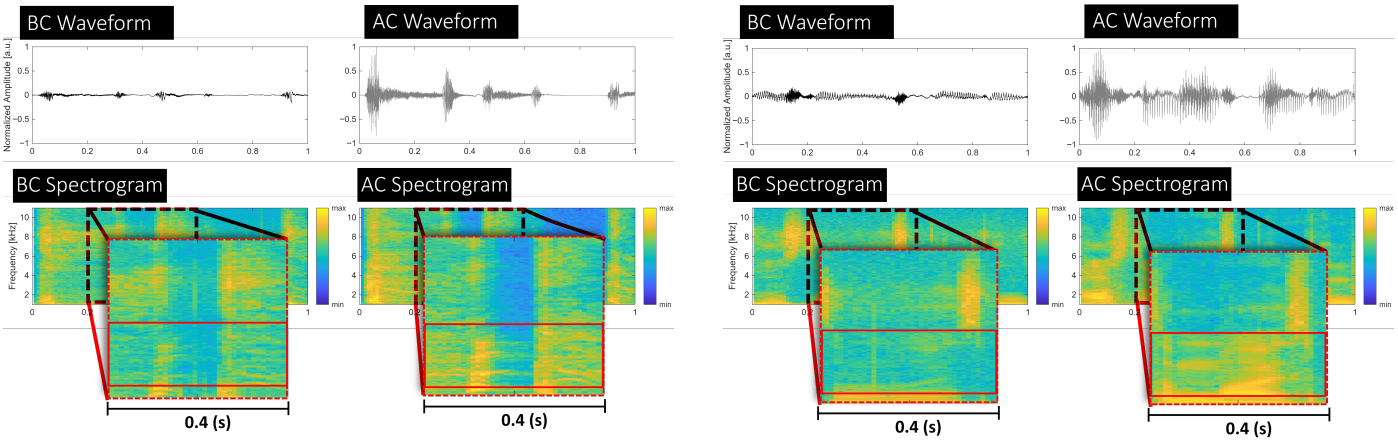


Figure 3. Comparison of air-conduction (AC), bone-conduction (BC), and IMU signals during cough (**left**) and speech (**right**), recorded with sensors placed on the sternum. Cough events appear as short, high-energy transients in both AC and BC channels and as sharp peaks in the IMU Z-axis, whereas speech produces sustained, lower-amplitude patterns. The BC signal activates only during body-coupled events, suppressing bystander speech visible in the AC channel (shaded regions).



(**a**) Cough (**b**) Speech

Figure 4. Comparison of bone-conduction (BC) and air-conduction (AC) signals during (**a**) cough and (**b**) speech, showing time–domain waveforms (**top**) and spectrograms (**bottom**). BC signals exhibit attenuated high-frequency content relative to AC, particularly in the 2–6 kHz band, indicated by the red dashed rectangle in the spectrograms. Elevated BC energy near 7 kHz, corresponds to the resonant frequency of the V2S200D sensor (Figure 5). Despite this reduced bandwidth, cough events remain clearly distinguishable in the BC channel as wideband, short-duration bursts, whereas speech harmonic structure is largely suppressed, illustrating the privacy-preserving property of the BC modality.

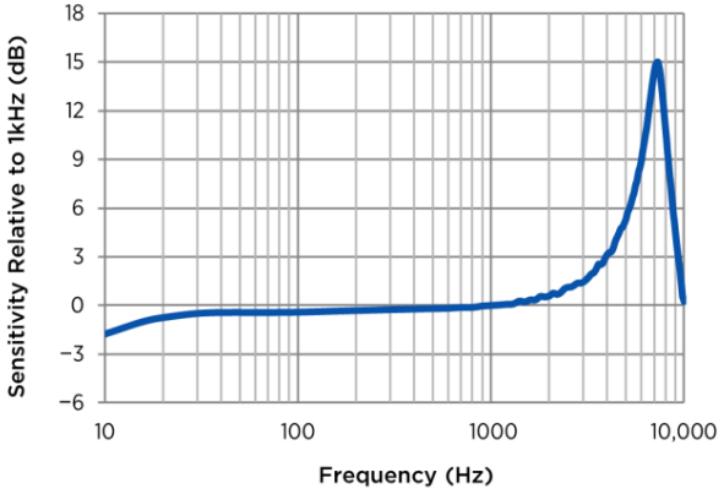


Figure 5. Frequency response of the V2S200D bone-conduction microphone (Syntiant Corp., Sunnyvale, CA, USA), showing a peak sensitivity near 7 kHz. This resonance peak is visible in BC spectrograms as elevated energy near 7 kHz and informed the design of the Selective SpecAugment strategy described in Section 4.

4. Embedded AI Model Design

4.1. Model Architecture

The proposed model is a lightweight convolutional neural network (CNN) designed for on-device cough detection on the MAX78000 neural accelerator. The network is implemented using the ai8x-training framework and is designed to satisfy the deployment constraints of the MAX78000 AI85 accelerator (Analog Devices, Inc., Wilmington, MA, USA), including 8-bit integer computation, limited on-chip memory, and fused convolutional operations for efficient execution.

The model takes as input a Mel-spectrogram representation derived from bone-conduction audio signals. In the primary configuration, the input is a single-channel 128×128 Mel-spectrogram. To study the effect of feature dimensionality on classification performance and deployability, additional ablations were performed using two-channel (Mel + Δ) and three-channel (Mel + Δ + Δ^2) input variants. These experiments were used to evaluate the contribution of temporal-derivative features under embedded hardware constraints, as described in Section Feature Extraction: Spectrogram, Temporal Derivatives, and MFCC Comparison.

The architecture comprises six convolutional layers followed by a fully connected classification layer. Each convolutional stage employs fused convolution and ReLU activation, with max-pooling applied in most intermediate layers to progressively expand the receptive field while reducing the spatial resolution from 128×128 to 8×8. The number of feature channels increases from 16 to 64 in the earlier layers to enable hierarchical feature extraction and is subsequently reduced to 32 and 16 in the later stages to balance representational capacity with on-chip memory constraints. This channel expansion-then-reduction strategy is motivated by the need to capture multi-scale spectro-temporal features in the early layers while keeping the final feature map small enough to fit within the MAX78000 fully connected layer memory limit (fc_inputs $\times$ dim$^2 \leq 1024$). A summary of the network architecture is provided in Table 4.

The final fully connected layer maps the flattened 16-channel feature representation to two output neurons corresponding to the cough and non-cough classes. During deployment, batch normalization and activation operations are folded into the convolutional layers to reduce runtime memory usage and inference latency. The architecture supports different input resolutions (e.g., 64×64 and 128×128) while preserving the same overall downsampling behavior and compatibility with the MAX78000 hardware pipeline.

Overall, the network contains approximately 57k trainable parameters and requires fewer than 350k multiply–accumulate (MAC) operations per inference. These counts were deliberately kept compact to satisfy the Flash and SRAM constraints of the MAX78000, which preclude the use of larger architectures common in smartphone- or server-based cough detection systems. The model supports both floating-point and 8-bit quantized inference with minimal performance loss, making it suitable for always-on, low-power respiratory event detection on embedded AI hardware.

Table 4. Layer-wise architecture of the proposed CoughNet model.

Layer	Output Dimension	Channels	Operation
Input	128×128	1/2/3	Spectrogram input
Conv1 + ReLU	128×128	16	3×3, padding = 1
Conv2 + Pool + ReLU	64×64	32	3×3, stride = 2
Conv3 + Pool + ReLU	32×32	64	3×3, stride = 2
Conv4 + Pool + ReLU	16×16	32	3×3, stride = 2
Conv5 + Pool + ReLU	8×8	32	3×3, stride = 2
Conv6 + ReLU	8×8	16	3×3, padding = 1
Fully Connected	1×1	2	Binary classification

4.2. Automated MAX78000 Training, Quantization, and Evaluation Pipeline

To enable reproducible and scalable experimentation on the MAX78000 neural accelerator, we developed a fully automated multi-stage pipeline that integrates model training, post-training quantization, evaluation, and statistical summarization across multiple random seeds. The objective of this framework is to support deterministic benchmarking of hardware-aware model configurations while minimizing manual intervention during model development. A summary of the pipeline stages is provided in Table 5.

(1) Training: In the first stage, the model is trained deterministically across multiple random seeds to quantify performance variability. All experiments were conducted using the Adam optimizer with a learning rate of 5×10^{-4}, weight decay of 10^{-4}, and a batch size of 64. Training was run for 75 epochs with a learning rate step schedule, and the best-performing floating-point checkpoint per seed was selected based on validation Macro-F1. A validation split of 15% was held out from the training fold for checkpoint selection; the remaining data were used for training. Four random seeds (1, 3, 4, 6) were used per fold to quantify performance variability, yielding 20 runs per experimental configuration (5 folds $\times$ 4 seeds). Each training run uses a predefined augmentation strategy combining waveform perturbation and SpecAugment for temporal and spectral masking, described further in Section 5.5.1. The preprocessing module automatically generates Mel-spectrogram representations, applies channel selection, and normalizes the input features. Training jobs are executed in parallel across CPU cores to accelerate experimentation, and the best-performing checkpoint from each run is saved using a standardized naming scheme for downstream processing.

(2) Quantization: In the second stage, each floating-point checkpoint is passed through an automated post-training quantization routine implemented using the quantization tools provided by the ai8x-synthesis framework. Multiple clipping policies are evaluated, including MAX, AVGMAX, SCALE, and STDDEV, to determine appropriate dynamic ranges for activations and weights. For the SCALE policy, several scaling coefficients are explored to balance range compression against numerical resolution. Similarly, for the STDDEV policy, multiple standard-deviation multipliers are tested to constrain activations within controlled variance bounds. The resulting quantized models are stored as 8-bit representations compatible with the MAX78000 accelerator, and the pipeline logs the clipping policy, quantization parameters, and generated checkpoints for each seed. The motivation for evaluating multiple clipping policies is that optimal quantization ranges depend on the activation distribution of the trained model, which can vary across seeds and configurations; systematic enumeration ensures the best-performing policy is selected for each run.

Table 5. Automated pipeline for MAX78000 model development and evaluation.

Stage	Description
Training	Input: training and testing data together with a random seed. Processing: data augmentation, spectrogram generation, channel configuration, and model optimization (Adam, lr = 5×10^{-4}, 75 epochs). Output: best floating-point model checkpoint.
Quantization	Input: floating-point model and quantization settings. Processing: clipping and post-training quantization using MAX, AVGMAX, SCALE, and STDDEV policies. Output: best quantized model checkpoint.
Evaluation	Input: trained or quantized model and testing data. Processing: deterministic inference, metric computation (Macro-F1, balanced accuracy, class-wise F1, AUC), and confusion-matrix generation. Output: evaluation logs and class-wise performance summaries.
Summarization	Input: floating-point and quantized evaluation logs. Processing: metric parsing and aggregation across seeds as mean $\pm$ standard deviation. Output: summary tables and CSV files for analysis.

(3) Evaluation: Each floating-point and quantized model is evaluated using a deterministic configuration to eliminate variability arising from data ordering, augmentation, and dataset partitioning. Evaluation is performed on the same held-out test set used for the corresponding floating-point model, enabling direct comparison of quantization effects. During inference, the pipeline computes Macro-F1, balanced accuracy, class-wise F1-scores,

and AUC and generates confusion matrices for class-wise performance analysis. Intermediate activations and representative sample inputs are also logged to support interpretability and to facilitate verification during hardware deployment. These traces are useful for diagnosing quantization-induced degradation and for ensuring consistency between software inference and on-device execution.

(4) Summarization: After training and evaluation are completed for all seeds, the final stage parses the floating-point and quantized log files, extracts the test-set metrics, and aggregates the results across runs. Mean and standard deviation are computed for each metric to characterize both expected performance and run-to-run stability. The aggregated results are exported as structured CSV files for downstream analysis of accuracy–efficiency trade-offs and hardware–software co-design decisions.

The automated pipeline provides an end-to-end framework for reproducible evaluation of quantized neural networks on the MAX78000. By ensuring that each stage, from training to summarization, is executed deterministically and logged systematically, the framework supports large-scale parameter sweeps and multi-seed benchmarking with minimal manual effort.

4.3. Quantization and Testing

To evaluate the effect of reduced numerical precision on model performance and hardware compatibility, we investigated two complementary 8-bit quantization strategies for deployment on the MAX78000 neural accelerator: post-training quantization (PTQ) and quantization-aware training (QAT). Both approaches target integer inference for weights and activations to enable efficient execution on the MAX78000 AI accelerator.

(1) Post-Training Quantization (PTQ): In the PTQ setting, a trained floating-point checkpoint is quantized after training without additional gradient updates, using the quantization utilities provided in the *ai8x-synthesis* framework. Multiple clipping policies are evaluated, including MAX, AVGMAX, SCALE, and STDDEV, each of which defines the activation range used during calibration. PTQ is computationally efficient and well suited for rapid hardware prototyping; however, its performance depends on the clipping policy, particularly when activation distributions are non-uniform or contain outliers.

Figure 6 summarizes the distribution of accuracy change (ΔAccuracy = Q8 $-$ Float) after PTQ across eight clipping configurations. The y-axis shows the per-seed accuracy change relative to the floating-point baseline; all values are negative, confirming that quantization consistently reduces accuracy across all tested policies. This universal degradation reflects the fundamental sensitivity of the compact network to reduced numerical precision: with only 57k parameters and activations concentrated in a small number of feature channels, even modest clipping errors in a single layer can propagate and affect the final classification boundary.

Among the tested methods, the SCALE strategy provides the most consistent performance with the lowest variance across scaling factors, as illustrated by the narrow interquartile ranges of SCALE_0.85 and SCALE_0.95 in Figure 6. This behavior is explained by the activation distribution of the trained network: the compact architecture produces relatively uniform per-layer activation ranges, for which a scaling coefficient of 0.85 effectively clips approximately the top 15% of the activation range, removing outlier activations that would otherwise force the quantization grid to be unnecessarily coarse. This controlled range compression therefore preserves the effective numerical resolution for the majority of activations, which dominate classification performance. By contrast, MAX and AVGMAX determine the clipping range from the observed maximum or average-maximum activation during calibration, making them sensitive to any single high-energy outlier activation which widens the quantization range and reduces effective resolution for the majority

of activations, explaining their greater variability across seeds visible in Figure 6. The STDDEV-based method shows competitive behavior at moderate clipping settings, but performance becomes less stable as the clipping range is relaxed excessively, likely because over-broad clipping reduces effective quantization resolution.

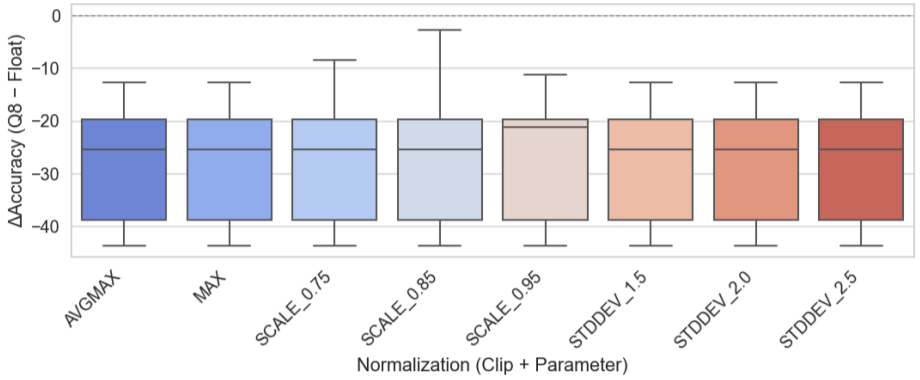


Figure 6. Distribution of accuracy change (ΔAccuracy = Q8 − Float) after post-training quantization (PTQ) across eight clipping configurations: AVGMAX, MAX, SCALE_0.75, SCALE_0.85, SCALE_0.95, STDDEV_1.5, STDDEV_2.0, and STDDEV_2.5. All values are negative, confirming that quantization consistently reduces accuracy relative to the floating-point baseline. Each box plot shows the distribution across seeds within the corresponding configuration. SCALE_0.85 and SCALE_0.95 exhibit the smallest spread and least degradation, indicating that moderate range compression is most effective at preserving model accuracy under 8-bit constraints. AVGMAX and MAX show wider distributions, reflecting greater sensitivity to layer-wise activation outliers.

Figure 7 compares floating-point and quantized F1-scores for the top-performing PTQ configurations across all evaluated seeds and clipping settings. Each point represents one (floating-point, quantized) pair from a specific seed and clipping configuration. Points lying on or near the dashed diagonal indicate cases where quantized performance closely tracks the floating-point baseline, reflecting minimal quantization-induced degradation, while points below the diagonal indicate performance loss after quantization. The clustering of SCALE-based configurations in the upper-right region (Float F1 $\approx$ 0.6–0.8), closest to the diagonal, confirms that SCALE clipping achieves the most favorable balance between quantization resolution and activation preservation among the tested policies. The AVGMAX and STDDEV points appearing at low float F1 values ($\approx$0.2) do not represent quantization failures; rather, as noted in the caption of Figure 7, they correspond to random seeds in which the floating-point model itself converged poorly during training, making the quantization comparison uninformative for those particular runs. This pattern further underscores the importance of reporting results across multiple seeds rather than from a single training run, and confirms that the poor quantized performance observed for these configurations is attributable to training instability rather than to the quantization policy itself. These results confirm that PTQ with SCALE clipping at a coefficient of 0.85 achieves stable deployment performance on the MAX78000 with relatively low implementation complexity and was, therefore, selected as the quantization strategy for all subsequent experiments.

(2) Quantization-Aware Training (QAT): To examine whether quantized performance could be improved further, we also evaluated QAT within the MAX78000-compatible training framework. In this approach, quantization operators are inserted directly into the training graph so that the network is optimized while simulating 8-bit inference during the forward pass, with gradient updates still computed in floating-point precision. Hardware-aware mechanisms, including fixed-point scaling and learned output shifts, are incorporated to preserve compatibility with the MAX78000 accelerator.

In our experiments, QAT did not provide a consistent improvement over PTQ. Across multiple seeds and quantization settings, the resulting quantized models generally failed to exceed the best PTQ configurations in either classification performance or stability. This outcome is likely attributable to the relatively compact network architecture and the constrained activation ranges observed in this task, which reduce the potential advantages of QAT while introducing additional quantization noise during optimization. Consequently, PTQ was selected as the final quantization strategy for deployment, as it offered a more favorable trade-off between simplicity, reproducibility, and performance on the MAX78000 platform.

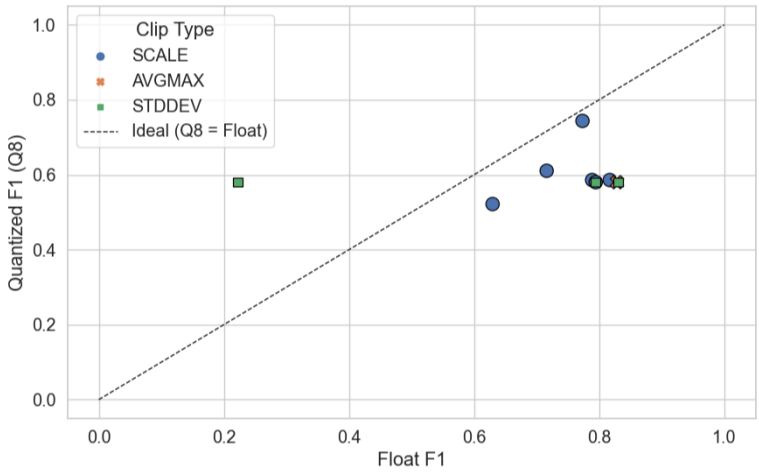


Figure 7. Comparison of floating-point (x-axis) and quantized F1-scores (y-axis) for the top-performing PTQ configurations, separated by clip type: SCALE (blue circles), AVGMAX (orange squares), and STDDEV (green squares). Each point represents one (floating-point, quantized) pair from a specific seed and clipping configuration. The dashed diagonal indicates the ideal case where quantized performance exactly matches the floating-point baseline. SCALE-based points cluster in the upper-right region (Float F1 $\approx$ 0.6–0.8), closest to the diagonal, confirming their superior quantization stability. AVGMAX and STDDEV points at lower float F1 values ($\approx$0.2) indicate configurations where the floating-point model itself performed poorly, making quantization comparison less meaningful for those runs.

5. Training and Evaluation Workflow

The development workflow for the proposed model was organized as a multi-stage pipeline to support reproducible quantization, evaluation, and deployment analysis on the MAX78000 platform. For each experimental configuration, the model was trained across four deterministic random seeds (1, 3, 4, 6) under leave-one-subject-out (LOSO) cross-validation with five folds, yielding 20 runs per configuration. The resulting floating-point checkpoints were quantized using the SCALE clipping policy with a scaling coefficient of 0.85, selected based on the PTQ analysis described in Section 4. Each quantized model was then evaluated on the corresponding held-out fold using a deterministic test protocol to enable consistent comparison across configurations.

Because the primary objective of this work is on-device deployment, only the 8-bit quantized (Q8) results are reported in the final comparison. Performance is summarized using Macro-F1, balanced accuracy, class-wise F1-scores for the cough and non-cough classes, and AUC, with results aggregated across seeds and reported as mean $\pm$ standard deviation. Macro-F1 and balanced accuracy serve as the primary deployment metrics because they account for class imbalance and are computed at a fixed decision threshold consistent with on-device binary inference. AUC is included as a threshold-independent complement to these metrics. This evaluation protocol allows identification of the most stable and best-performing quantized configuration for deployment on the MAX78000 hardware.

It should be noted that all results are based on segment-level binary classification of pre-segmented one-second windows. This evaluation protocol is appropriate for assessing the per-segment detection capability of the deployed model, but does not directly evaluate continuous cough monitoring performance. Deployment in a continuous monitoring scenario would require an additional event detection layer, such as a sliding-window segmentation and onset detection strategy, which is beyond the scope of this feasibility study and is identified as a direction for future work.

5.1. Baseline Model Configuration

The baseline configuration serves as the reference point for all subsequent experiments. In this setup, a single-channel Mel-spectrogram input is used with a sampling rate of 16 kHz, a 1-second segment duration, and a spectrogram resolution of 128×128, generated using 128 Mel filters, an FFT size of 1024, and a hop length of 118. The model is trained using class weights of [1.5, 1.0] for the cough and non-cough classes respectively, and no data augmentation is applied at this stage. This configuration was chosen as the starting point because it satisfies the MAX78000 memory constraints (fc_inputs $\times$ dim^2 = $16 \times 8 \times 8 = 1024$) while providing sufficient spectral and temporal resolution for bone-conduction cough signals.

The experimental workflow is organized into three sequential phases, each focused on tuning a different aspect of the pipeline: data preprocessing (Section 5.2), feature representation and model architecture (Section Feature Extraction: Spectrogram, Temporal Derivatives, and MFCC Comparison), and data augmentation (Section 5.5.1). At the end of each phase, the best-performing configuration is incorporated into the baseline, which then serves as the starting point for the next phase. This iterative refinement ensures that each subsequent stage builds upon the most effective version of the model and training pipeline. Because the primary goal is on-device deployment, configuration selection at each stage is based on quantized (Q8) performance rather than floating-point performance.

5.2. Data Preprocessing

This stage establishes the signal preprocessing configuration used throughout training and evaluation. Three preprocessing parameters were systematically varied to study their effects on classification performance and hardware compatibility: segment duration (0.5–1.0 s), sampling frequency (4–16 kHz), and spectrogram input resolution (32×32 to 128×128). The rationale for varying each parameter is as follows. Segment duration determines the amount of temporal context available to the classifier; since a typical human cough lasts approximately 0.5 s, longer windows provide pre- and post-cough context at the cost of increased input size. Sampling frequency determines the spectral bandwidth of the input; the V2S200D bone-conduction sensor exhibits peak sensitivity near 7 kHz, which according to the Nyquist criterion requires a minimum sampling rate of 14 kHz to preserve this informative frequency range. Input resolution controls the trade-off between spectral detail and on-chip memory usage, with higher resolutions preserving finer spectro-temporal structure but increasing the feature map dimensions and risk of exceeding MAX78000 memory limits. The final preprocessing configuration was selected to balance predictive performance with the memory and deployment constraints of the MAX78000 platform.

5.2.1. Segment Duration

A typical human cough lasts approximately 0.5 s under natural conditions. To assess whether a longer input window improves classification by capturing pre- and post-cough context, the segment duration was varied from 0.5 s to 1.0 s. Since the training dataset consists of 1 s audio samples, shorter segments are produced by extracting the center

portion of each clip rather than trimming from the onset of high energy. This center-extraction strategy ensures balanced temporal coverage across all segment lengths and prevents the model from learning to rely on energy-onset cues that may not generalize across cough styles and intensities.

Table 6 summarizes the quantized (Q8) performance across the three evaluated durations. The 1.0 s configuration achieved the best results across all reported metrics. The performance drop for shorter durations is consistent with the hypothesis that the 1 s window captures not only the main cough burst but also the preceding breath and the trailing decay, both of which contribute discriminative spectro-temporal structure. The 0.5 s window, which approximately matches the cough burst duration alone, showed the largest degradation in cough F1 (0.5629 ± 0.1049) and the highest variance, suggesting that this duration provides insufficient context for stable quantized classification. Accordingly, the 1.0 s segment duration was selected for all subsequent experiments. From a deployment perspective, this choice confirms that the MAX78000 must process a full second of bone-conduction audio per inference cycle a constraint that directly informed the memory budget allocation and input pipeline design described in Section 6.

Table 6. Comparison of quantized (Q8) performance across different segment durations. Results are reported as mean ± standard deviation across seeds. The selected configuration is indicated by the gray shaded row.

Time (s)	MacroF1	BalancedAcc	$F1_{cough}$	$F1_{non\text{-}cough}$	AUC
1.0	0.79 ± 0.04	0.79 ± 0.04	0.72 ± 0.06	0.87 ± 0.02	0.89 ± 0.03
0.75	0.73 ± 0.04	0.73 ± 0.04	0.63 ± 0.06	0.83 ± 0.02	0.84 ± 0.04
0.5	0.68 ± 0.05	0.69 ± 0.06	0.56 ± 0.10	0.79 ± 0.03	0.79 ± 0.04

5.2.2. Input Size Variation

To investigate the effect of spectrogram resolution on model performance, four input configurations were evaluated by varying the number of Mel bins and time resolution: 32×32, 64×64, 128×128, and 256×256. For each configuration, the FFT size and hop length were adjusted proportionally to maintain consistent temporal coverage. Lower-resolution inputs reduce the number of model parameters and computational complexity, whereas higher-resolution inputs preserve finer spectral detail at the cost of increased memory and compute requirements.

The MAX78000 hardware imposes a constraint on the fully connected layer dimensions, requiring that the total number of flattened input features (fc_inputs $\times$ dim^2) remain within the allowable on-chip memory limit, where dim denotes the final spatial dimension of the feature map after convolution and pooling. For the 128×128 configuration this condition is satisfied ($16 \times 8 \times 8 = 1024$). Scaling to 256×256 increases the final feature map size to $16 \times 16 \times 16 = 4096$, which exceeds the MAX78000 limit and prevents deployment. The 256×256 configuration is therefore included in Table 7 for reference only and was not considered a feasible deployment option.

As shown in Table 7, the 128×128 configuration achieved the best quantized performance across all metrics. The 32×32 configuration produced lower scores, consistent with the expected loss of fine spectral and temporal cues at reduced resolution. The 64×64 configuration showed an even larger degradation after quantization despite its lower memory footprint, a counterintuitive result that warrants explanation. At 64×64 resolution, the hop length is 236 samples at 16 kHz, corresponding to a time step of approximately 14.75 ms per spectrogram column. This resolution is coarse enough to blur the short-duration transient structure of cough events, yet fine enough to retain some spectral detail, resulting in a feature representation that is neither compact and robust like 32×32 nor sufficiently detailed like 128×128. Under 8-bit quantization, this intermediate representation is particularly

problematic: the activations span a moderately wide dynamic range without the strong, consistent energy patterns present at higher resolution, making the clipping boundaries harder to set accurately during post-training quantization. In contrast, the 32 × 32 configuration produces highly compressed, coarse features whose activation distributions are more uniform and therefore more amenable to stable 8-bit quantization, explaining why it outperforms 64 × 64 despite its lower floating-point resolution. The AUC values follow a different trend, increasing monotonically from 0.86 at 32 × 32 to 0.88 at 64 × 64 to 0.89 at 128 × 128, because AUC is a threshold-independent metric that reflects the overall discriminative capacity of the model, which benefits from higher resolution regardless of the threshold-dependent sensitivity–specificity balance that drives MacroF1 and F1-score degradation. The 128 × 128 configuration provides the best balance between discriminative feature resolution, quantization stability, and MAX78000 hardware feasibility and was therefore selected for all subsequent experiments.

Table 7. Spectrogram configuration and quantized (Q8) performance across input sizes. The 128 × 128 configuration achieved the best overall performance and was selected for subsequent experiments; the selected configuration is indicated by the gray shaded column. The 256 × 256 configuration exceeded the MAX78000 memory limit and was not deployable.

Feature	32 × 32	64 × 64	128 × 128	256 × 256
n_{mels}	32	64	128	256
n_{fft}	512	1024	1024	2048
Hop length	472	236	118	59
Parameters	55,856	56,240	57,776	231,104
MacroF1 (Q8)	0.64 ± 0.04	0.48 ± 0.03	0.79 ± 0.04	—
BalancedAcc (Q8)	0.66 ± 0.03	0.58 ± 0.02	0.79 ± 0.04	—
$F1_{\text{cough}}$ (Q8)	0.47 ± 0.07	0.31 ± 0.01	0.72 ± 0.06	—
$F1_{\text{non-cough}}$ (Q8)	0.82 ± 0.02	0.65 ± 0.05	0.87 ± 0.02	—
AUC (Q8)	0.86 ± 0.03	0.88 ± 0.03	0.89 ± 0.03	—

5.2.3. Sampling Frequency Analysis

To investigate the effect of sampling rate on model performance and computational efficiency, the input audio was evaluated at three sampling frequencies: 4 kHz, 8 kHz, and 16 kHz. For each configuration, the analysis window was fixed at 64 ms, while the FFT size and hop length were scaled proportionally, as summarized in Table 8. Lower sampling rates reduce computational cost and memory usage, which is advantageous for resource-constrained embedded systems. However, aggressive downsampling limits the available spectral bandwidth and may remove high-frequency components that contribute to cough discrimination.

The bone-conduction microphone used in this work exhibits peak sensitivity near 7 kHz (Figure 5). According to the Nyquist criterion, preserving this frequency range requires a minimum sampling rate of 14 kHz. The 16 kHz configuration therefore provides sufficient bandwidth to capture the most informative spectral content of the bone-conduction cough signal while remaining compatible with MAX78000 deployment constraints, whereas the 8 kHz and 4 kHz configurations truncate spectral content below and well below this resonance, respectively.

As shown in Table 8, the 16 kHz configuration achieved the best overall quantized MacroF1 and AUC, and the highest non-cough F1-score. However, it should be noted that the differences between 16 kHz and 8 kHz across most metrics are small and fall within the reported standard deviation ranges: MacroF1 (0.79 ± 0.04 vs. 0.78 ± 0.04), balanced accuracy (0.79 ± 0.04 vs. 0.79 ± 0.05), and cough F1-score (0.72 ± 0.06 vs. 0.71 ± 0.07) are all within mutual SD overlap and should not be interpreted as meaningful performance differences in isolation. The more reliable distinction between 16 kHz and 8 kHz lies in

the non-cough F1-score (0.87 ± 0.02 vs. 0.86 ± 0.02) and AUC (0.89 ± 0.03 vs. 0.87 ± 0.02), where the differences are more consistent across seeds and better supported by the Nyquist argument: the 8 kHz configuration truncates spectral content below the 7 kHz resonance of the V2S200D sensor, which reduces the discriminative information available for non-cough classification. The 4 kHz configuration yielded a marginally higher balanced accuracy and cough F1-score than 16 kHz, but at the cost of a substantially lower non-cough F1-score (0.83 ± 0.05 vs. 0.87 ± 0.02) and lower AUC (0.86 ± 0.03 vs. 0.89 ± 0.03), indicating a less balanced and less discriminative overall classifier. This pattern suggests that the loss of spectral content above 2 kHz at 4 kHz sampling disproportionately impairs non-cough discrimination, likely because several non-cough events such as speech and clothing rubbing have distinctive spectral characteristics in the 2–7 kHz range that help separate them from coughs. Taken together, 16 kHz provides the most reliable trade-off between signal fidelity, quantization robustness, and deployment feasibility and was therefore selected as the baseline sampling frequency for the subsequent experiments. Critically, this choice is not merely a performance preference but a physically motivated necessity: operating below 14 kHz on this BC sensor would sacrifice the resonance band that carries the most discriminative cough-specific energy, and no amount of model optimization can recover information that was never captured at the sensor level.

Table 8. Sampling frequency comparison with preprocessing parameters and quantized (Q8) performance. Results are reported as mean ± standard deviation across seeds. The selected configuration is indicated by the gray shaded row.

Feature	16 kHz	8 kHz	4 kHz
n_{fft}	1024	512	256
Hop length	118	59	30
Window size (ms)	64	64	64
MacroF1 (Q8)	0.79 ± 0.04	0.78 ± 0.04	0.78 ± 0.05
BalancedAcc (Q8)	0.79 ± 0.04	0.79 ± 0.05	0.80 ± 0.04
$F1_{\text{cough}}$ (Q8)	0.72 ± 0.06	0.71 ± 0.07	0.72 ± 0.05
$F1_{\text{non-cough}}$ (Q8)	0.87 ± 0.02	0.86 ± 0.02	0.83 ± 0.05
AUC (Q8)	0.89 ± 0.03	0.87 ± 0.02	0.86 ± 0.03

In summary, Phase 1 established the foundation for deployment-oriented model optimization through a systematic exploration of segment duration, sampling frequency, and input size. A segment duration of 1 s, a sampling frequency of 16 kHz, and an input size of 128×128 consistently provided the best overall trade-off between temporal context, spectral fidelity, and hardware compatibility. Under this configuration, the quantized model achieved a MacroF1 of 0.79 ± 0.04, balanced accuracy of 0.79 ± 0.04, cough F1-score of 0.72 ± 0.06, non-cough F1-score of 0.87 ± 0.02, and AUC of 0.89 ± 0.03. Shorter input windows and lower-resolution representations consistently resulted in reduced performance, highlighting the trade-off between computational efficiency and the preservation of discriminative cough features under 8-bit quantization constraints.

5.3. Feature and Model Architecture

Phase 2 builds upon the optimized preprocessing configuration established in Phase 1 and focuses on improving the model's representational capacity through feature-level and architectural refinements. Three factors are investigated in this phase: the spectral feature representation (Mel-spectrogram vs. MFCC); the input channel configuration (one, two, or three channels); and class-weight optimization.

The rationale for examining these factors is as follows. The choice of spectral representation determines which acoustic properties of the bone-conduction signal are made available to the classifier. Mel-spectrograms preserve the full spectral energy distribution

and are well suited to capturing the wideband, transient nature of cough events, while MFCCs emphasize perceptually weighted spectral shape and are more compact. The number of input channels determines whether temporal-derivative features (Δ, Δ^2) are included alongside the base spectrogram; these derivatives capture the rate and acceleration of spectral change, which may help characterize the onset and decay of cough bursts. However, multi-channel inputs also increase memory requirements and may reduce quantization robustness on hardware-constrained platforms. Class weighting addresses the 2:1 non-cough/cough imbalance in the training set by assigning a higher loss penalty to cough misclassifications, which is important for maximizing sensitivity to the target event.

The objective of this phase is to identify the most informative and quantization-robust input configuration for bone-conduction cough detection, while preserving compatibility with the memory and computational constraints of the MAX78000 platform. The best configuration identified in Phase 2 is carried forward as the starting point for Phase 3 (data augmentation, Section 5.5.1).

Feature Extraction: Spectrogram, Temporal Derivatives, and MFCC Comparison

To capture both spectral and temporal characteristics of cough sounds, multiple feature representations and channel configurations were evaluated, as illustrated in Figure 8. For Mel-based inputs, the single-channel configuration (1-ch) uses only the log-Mel spectrogram and represents the spectral energy distribution over time. The two-channel configuration (2-ch) augments the Mel spectrogram with its first temporal derivative (Δ), which emphasizes short-term spectral changes associated with transient events such as cough onsets. The three-channel configuration (3-ch) further includes the second temporal derivative (Δ^2), which captures the acceleration of spectral variation and provides additional information about the temporal evolution of cough bursts and decays.

$$\Delta_t = \frac{\sum_{n=1}^{N} n \cdot (E_{t+n} - E_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{1}$$

$$\Delta_t^2 = \frac{\sum_{n=1}^{N} n \cdot (\Delta_{t+n} - \Delta_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{2}$$

where E_t denotes the log-Mel energy at frame t, and N is the temporal window size (typically $N = 2$). These derivatives capture the *velocity* and *acceleration* of spectral energy variation across time and enrich the representation of cough temporal dynamics. The same one-, two-, and three-channel configurations were also evaluated for MFCC-based inputs.

As shown in Table 9, derivative-based representations improved floating-point performance, particularly for Mel features, where the 2-channel configuration achieved the strongest float-domain results among the Mel variants. However, this benefit did not carry over to 8-bit deployment. After quantization, the single-channel Mel representation achieved the best overall performance across MacroF1, balanced accuracy, and AUC. The progressive degradation in MacroF1 from Mel (1-ch) to Mel (3-ch), from 0.79 to 0.72 to 0.60, indicates that additional derivative channels introduce quantization sensitivity that outweighs their representational benefit under 8-bit constraints. This is likely because derivative features exhibit wider activation ranges and more non-uniform distributions than the base spectrogram, making them harder to clip accurately during PTQ.

Among the MFCC variants, the single-channel configuration achieved the strongest results, with a competitive cough F1-score (0.74 ± 0.04) and the highest balanced accuracy (0.81 ± 0.04) among all evaluated configurations. Despite this, Mel (1-ch) was selected as the final feature representation because it provided the best overall balance between

MacroF1, non-cough discrimination, and AUC, which are collectively more informative for the target deployment scenario than cough sensitivity alone.

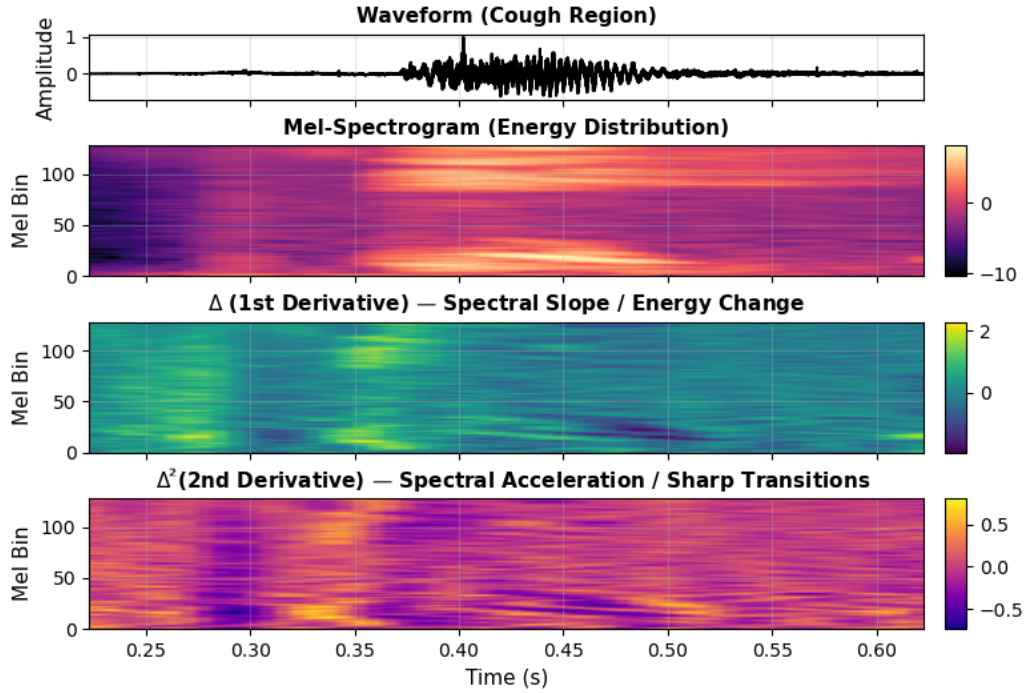


Figure 8. Waveform and Mel-based feature representations of a single cough segment, including the log-Mel spectrogram (**top**), first temporal derivative Δ (**middle**), and second temporal derivative Δ^2 (**bottom**). The Mel spectrogram captures the wideband energy burst of the cough event, while Δ highlights the onset and offset transitions, and Δ^2 captures sharp spectral accelerations. Although these derivative channels enrich the floating-point representation, they reduce quantization robustness under 8-bit deployment constraints.

Table 9. Comparison of quantized (Q8) performance across Mel and MFCC feature representations with different channel configurations. Results are reported as mean ± standard deviation across seeds. The selected configuration is indicated by the gray shaded row.

Feature	MacroF1	BalancedAcc	$F1_{cough}$	$F1_{non\text{-}cough}$	AUC
Mel (1 ch)	0.79 ± 0.04	0.79 ± 0.04	0.72 ± 0.06	0.87 ± 0.02	0.89 ± 0.03
Mel (2 ch)	0.72 ± 0.04	0.73 ± 0.03	0.59 ± 0.06	0.85 ± 0.02	0.89 ± 0.03
Mel (3 ch)	0.60 ± 0.05	0.64 ± 0.03	0.40 ± 0.08	0.80 ± 0.03	0.89 ± 0.03
MFCC (1 ch)	0.79 ± 0.03	0.81 ± 0.04	0.74 ± 0.04	0.84 ± 0.03	0.88 ± 0.02
MFCC (2 ch)	0.73 ± 0.05	0.76 ± 0.04	0.65 ± 0.08	0.82 ± 0.03	0.88 ± 0.03
MFCC (3 ch)	0.71 ± 0.03	0.72 ± 0.03	0.60 ± 0.07	0.82 ± 0.02	0.88 ± 0.03

5.4. Class Weighting Strategy

To address the class imbalance between cough and non-cough samples in the training dataset, a class-weighted cross-entropy loss was adopted. The cough class contains $N_{\text{cough}} = 1045$ samples and the non-cough class contains $N_{\text{non-cough}} = 2090$ samples, yielding a 2:1 ratio. The total loss is defined as

$$L = w_{\text{cough}} \cdot L_{\text{cough}} + w_{\text{non-cough}} \cdot L_{\text{non-cough}} \tag{3}$$

where $w_{\text{cough}} > w_{\text{non-cough}}$ assigns a greater penalty to errors on the underrepresented cough class. Class-weight configurations are reported as $[w_{\text{cough}}, w_{\text{non-cough}}]$.

Three weighting configurations were evaluated: [1.5, 1], [2, 1], and [3, 1]. As summarized in Table 10, increasing the cough-class weight produced modest but consistent improvements in MacroF1, balanced accuracy, and cough F1-score. The [3, 1] configuration

achieved the best results across these metrics. The gains relative to [1.5, 1] are small MacroF1 increases from 0.79 to 0.80 and cough F1 from 0.72 to 0.74 and should be interpreted as incremental improvements rather than substantial changes. Notably, the AUC remained stable across all three configurations at 0.89, indicating that class weighting primarily shifts the decision threshold toward greater cough sensitivity without meaningfully changing the overall discriminative capacity of the model. Accordingly, the [3, 1] configuration was selected for the subsequent experiments as it provided the best balance between cough sensitivity and overall quantized performance. The stability of AUC across all three configurations is a practically important finding for deployment: it confirms that the underlying model has consistent discriminative capacity regardless of the weighting scheme, and that the [3, 1] configuration simply shifts the operating point to favour recall which is the clinically preferable direction for a health monitoring application where missed cough events carry greater cost than occasional false alarms.

Table 10. Comparison of class-weight configurations under quantized (Q8) evaluation. Results are reported as mean ± standard deviation across seeds. The selected configuration is indicated by the gray shaded row.

Class Weight	MacroF1	BalancedAcc	$F1_{cough}$	$F1_{non\text{-}cough}$	AUC
[1.5, 1]	0.79 ± 0.04	0.79 ± 0.04	0.72 ± 0.06	0.87 ± 0.02	0.89 ± 0.03
[2, 1]	0.80 ± 0.05	0.80 ± 0.05	0.73 ± 0.07	0.87 ± 0.03	0.89 ± 0.03
[3, 1]	0.80 ± 0.04	0.81 ± 0.04	0.74 ± 0.06	0.87 ± 0.03	0.89 ± 0.03

5.5. Data Augmentation

5.5.1. SpecAugment Augmentation

To improve model generalization and robustness to unseen cough patterns, SpecAugment was applied during training by randomly masking contiguous regions along both the frequency and time axes of the Mel spectrogram. Several masking configurations were explored: $(1,2)$, $(1,4)$, and $(1,8)$ for frequency and time masking widths, together with a no-mask baseline $(0,0)$. For each masking width, two augmentation variants were evaluated: *Standard SpecAugment*, in which masking is applied without restricting the masked region, and *Selective SpecAugment*, in which masking is constrained to preserve the most informative bone-conduction frequency region. This strategy encourages the model to rely on more global and invariant spectral–temporal cues rather than localized features, thereby helping to mitigate overfitting while remaining compatible with the real-time and memory constraints of the MAX78000 deployment.

Table 11 summarizes the top ten Mel-frequency bins ranked by their mean log-energy and corresponding temporal derivatives (Δ, Δ^2). The highlighted bins (ranks 3–5) correspond to frequencies between approximately 7–8 kHz, where both the log-energy and its temporal derivatives show prominent activity. This region coincides with the resonance peak of the bone-conduction sensor near 7 kHz. Standard SpecAugment could occasionally mask this spectrally prominent region, thereby suppressing discriminative cues important for cough characterization. To mitigate this, *Selective SpecAugment* excludes Mel bins corresponding to the 6–8 kHz range (approximately Mel bins 96–128) from masking. By preserving this high-energy resonance band, the model retains essential spectral information associated with bone-conducted cough bursts while still benefiting from temporal masking and lower-frequency spectral masking.

As shown in Table 12, the no-mask baseline $(0,0)$ achieved the highest MacroF1 (0.80 ± 0.04) and non-cough F1-score (0.87 ± 0.03), while *Selective SpecAugment* $(1,4)$ achieved the best balanced accuracy (0.81 ± 0.05) and cough F1-score (0.74 ± 0.06). The AUC values are broadly similar across all configurations (0.89–0.90), indicating that augmentation choices

primarily affect the threshold-dependent sensitivity–specificity balance rather than the overall discriminative capacity of the model. More aggressive masking at $(1, 8)$, for both standard and selective variants, led to lower MacroF1 and cough F1, suggesting that excessive removal of spectral–temporal information degrades quantized robustness.

The selection of *Selective SpecAugment* $(1, 4)$ over the no-mask baseline warrants explicit justification since the no-mask configuration achieves higher MacroF1 and non-cough F1 on this dataset. The rationale is generalization rather than in-sample performance: the no-mask baseline is evaluated on the same five participants used for training and may overfit to participant-specific spectral patterns. *Selective SpecAugment* introduces controlled spectral and temporal variability that encourages the model to rely on more global and invariant cues, which is expected to benefit generalization to new participants in the planned cohort expansion. The preservation of the 6–8 kHz resonance band in the selective variant ensures that the most physically meaningful BC cough features those concentrated at the V2S200D sensor resonance are never masked, so the augmentation introduces variability only where the model can afford to be uncertain. The *Selective SpecAugment* $(1, 4)$ configuration was therefore selected for deployment as the configuration most likely to support robust generalization under the hardware and data constraints of the present study.

Table 11. Top 10 dominant Mel bins ranked by mean log-energy and temporal derivatives. Highlighted bins (ranks 3–5) fall within the 7–8 kHz resonance band of the V2S200D sensor and were excluded from masking in the Selective SpecAugment strategy.

Rank	Mel Bin	Center Freq. (Hz)	Mel (log E)	Δ	Δ^2
1	0	0	1.50	0.11	−0.61
2	1	21	0.93	−0.63	−1.32
3	121	7630	0.38	−1.29	−1.95
4	120	7190	0.35	−1.29	−1.96
5	122	8090	0.35	−1.30	−1.97
6	119	6760	0.29	−1.33	−1.99
7	123	8330	0.26	−1.37	−2.04
8	2	43	0.21	−1.43	−2.10
9	118	6340	0.19	−1.47	−2.13
10	124	8570	0.13	−1.56	−2.23

Table 12. Comparison of quantized (Q8) performance across SpecAugment masking configurations. Results are reported as mean ± standard deviation across seeds. The selected configuration is indicated by the gray shaded row.

Mask Configuration	MacroF1	BalancedAcc	$F1_{cough}$	$F1_{non\text{-}cough}$	AUC
$(0, 0)$ No Mask	0.80 ± 0.04	0.81 ± 0.04	0.74 ± 0.06	0.87 ± 0.03	0.89 ± 0.03
$(1, 2)$ Standard	0.80 ± 0.04	0.81 ± 0.04	0.73 ± 0.06	0.86 ± 0.03	0.89 ± 0.03
$(1, 4)$ Standard	0.80 ± 0.04	0.81 ± 0.04	0.74 ± 0.06	0.86 ± 0.02	0.90 ± 0.03
$(1, 8)$ Standard	0.79 ± 0.05	0.79 ± 0.05	0.72 ± 0.07	0.86 ± 0.03	0.90 ± 0.03
$(1, 2)$ Selective	0.79 ± 0.04	0.80 ± 0.05	0.73 ± 0.06	0.86 ± 0.03	0.89 ± 0.04
$(1, 4)$ Selective	0.80 ± 0.05	0.81 ± 0.05	0.74 ± 0.06	0.86 ± 0.03	0.89 ± 0.03
$(1, 8)$ Selective	0.79 ± 0.04	0.80 ± 0.05	0.72 ± 0.06	0.86 ± 0.03	0.89 ± 0.03

5.5.2. Waveform Augmentation

In addition to spectrogram-level augmentation, waveform-level perturbations were explored, including random amplitude scaling, temporal shifts, and low-level Gaussian noise injection, to simulate variations in skin contact and sensor coupling. These perturbations did not provide a consistent improvement in quantized performance across seeds and folds. Consequently, waveform augmentation was not included in the final experimental

sweep, and greater emphasis was placed on spectrogram-level strategies, which showed clearer and more consistent benefits under deployment-oriented evaluation.

This outcome is consistent with the physical properties of the BC sensing modality. Unlike air-conduction microphones, which are susceptible to ambient acoustic variability, bone-conduction sensors couple mechanically to the body surface and therefore exhibit relatively stable signal characteristics across acquisition sessions. As a result, the primary source of variability in this dataset is spectral and temporal in nature, related to differences in cough style, intensity, and duration across participants, rather than low-level waveform amplitude or noise. This explains why spectrogram-level augmentation, which directly targets spectral and temporal variability, was more effective than waveform-level perturbations for this modality.

5.6. Final Model Design

The final training configuration uses 1 s audio segments sampled at 16 kHz, producing 128×128 single-channel Mel-spectrogram inputs, a class-weight configuration of $[3, 1]$, and *Selective SpecAugment* with a $(1, 4)$ frequency–time masking configuration. This combination was selected based on the cumulative results of Phases 1–3, with each design choice motivated by the hardware and task constraints described in the preceding sections.

As summarized in Table 13, the final configuration produced modest but consistent improvements in quantized performance relative to the baseline. MacroF1 increased from 0.79 to 0.80, balanced accuracy from 0.79 to 0.81, and cough F1-score from 0.72 to 0.74. The non-cough F1-score remained essentially unchanged (0.87 vs. 0.86), and the AUC was stable at 0.89 for both configurations. These gains are real but small and should be interpreted as incremental refinements rather than large performance steps; their practical significance lies in the improved cough sensitivity and class balance, which are more directly relevant to the deployment objective than aggregate accuracy.

Notably, the improvement is more pronounced under quantized evaluation than in the floating-point setting, indicating that the benefit of the final design is primarily realized during 8-bit deployment. Since the primary objective of this work is on-device inference on the MAX78000, the final model was selected based on its Q8 performance rather than its floating-point baseline. This outcome highlights the importance of optimizing the full training and augmentation pipeline directly for quantized embedded deployment rather than relying solely on floating-point performance as the selection criterion.

Table 13. Quantized (Q8) performance comparison between the baseline and the final optimized model. Results are reported as mean ± standard deviation across seeds. The selected configuration is indicated by the gray shaded row.

Model	MacroF1	BalancedAcc	$F1_{cough}$	$F1_{non\text{-}cough}$	AUC
Baseline	0.79 ± 0.04	0.79 ± 0.04	0.72 ± 0.06	0.87 ± 0.02	0.89 ± 0.03
Final Model	0.80 ± 0.05	0.81 ± 0.05	0.74 ± 0.06	0.86 ± 0.03	0.89 ± 0.03

6. Embedded AI Deployment

The final stage of this work focuses on the deployment of the optimized quantized model on the MAX78000 platform, as illustrated in Figure 1d, to assess its practical feasibility for low-power on-device cough detection. In addition to classification performance, this stage evaluates deployment-oriented metrics, including inference latency, energy consumption, and memory usage, which are critical constraints for always-on embedded sensing systems. The results reported here should be interpreted as an initial hardware feasibility assessment on a benchtop prototype rather than as evidence of readiness for con-

tinuous real-world monitoring, given the five-participant cohort and controlled acquisition conditions of the underlying dataset.

6.1. MAX78000 Neural Accelerator and Custom Hardware Design

The MAX78000 is a low-power microcontroller that integrates a hardware CNN accelerator for energy-efficient inference in embedded applications. The device combines an Arm Cortex-M4 core for general-purpose system control with a dedicated CNN engine that supports highly parallel 8-bit neural network execution. By storing weights and activations in on-chip memory, the accelerator reduces reliance on external memory access and lowers inference power consumption. This architecture makes the MAX78000 well suited for always-on sensing tasks such as audio event detection and wearable health monitoring. The platform provides standard peripheral interfaces, including I^2S, SPI, and I^2C, which support integration with microphones, inertial sensors, and auxiliary embedded components.

To support the present study, we developed a custom hardware prototype built around the MAX78000 for multimodal respiratory sensing and on-device inference. The prototype consists of a four-layer printed circuit board integrating the MAX78000 together with a power management IC (PMIC), microphone interface circuitry, an audio codec, and an SD-card interface for data logging. The board also includes expansion headers for connecting an external IMU, enabling synchronized acquisition of motion and vibration signals during multimodal data collection. This modular design facilitates early-stage debugging and iterative hardware refinement through accessible test points and flexible sensor interfacing. Figure 9 shows the embedded AI hardware setup used in this work.

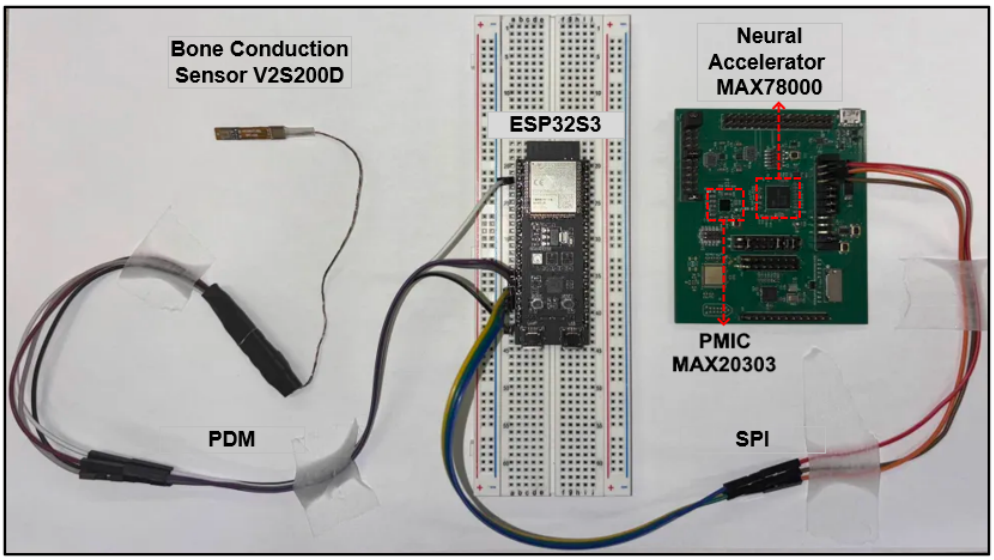


Figure 9. Embedded AI hardware setup consisting of the bone-conduction sensor (V2S200D, Syntiant Corp., Sunnyvale, CA, USA), an ESP32-S3 interface, and the MAX78000 neural accelerator platform (Analog Devices, Inc., Wilmington, MA, USA). The ESP32-S3 (Espressif Systems, Shanghai, China) streams PDM audio to the MAX78000 over SPI for real-time inference. The current implementation is a benchtop prototype; a miniaturized 3 cm × 3 cm wearable version is under development.

It should be noted that the current implementation is a benchtop prototype and has not been miniaturized for wearable use. A compact version of the board, with an approximate size of 3 cm × 3 cm, is currently under development for future integration into a chest-mounted wearable patch. The results reported in this section therefore reflect on-device inference performance under laboratory conditions and should not be interpreted as validated wearable system performance.

6.2. Selection of the Final Deployed Model and C Code Generation

Model selection was based on quantized performance together with hardware feasibility on the MAX78000 platform. Although some higher-dimensional feature configurations achieved stronger software-level performance, the final deployed model was selected from the set of configurations that could be synthesized and executed within the memory constraints of the device. The final deployment configuration uses a single-channel bone-conduction Mel-spectrogram input with 1 s segments, 16 kHz sampling frequency, and a 128×128 input resolution, trained with class weights $[3, 1]$ and *Selective SpecAugment* with a $(1, 4)$ masking configuration.

The 2-channel 128×128 Mel + Δ representation was not adopted for final deployment because it exceeded the available memory budget of the MAX78000 during preprocessing and CNN input allocation. The selected single-channel configuration satisfied the hardware limits of the target platform while preserving strong quantized classification performance.

Under quantized evaluation, the deployed model achieved a MacroF1 of 0.80 ± 0.05, balanced accuracy of 0.81 ± 0.05, cough F1-score of 0.74 ± 0.06, non-cough F1-score of 0.86 ± 0.03, and AUC of 0.89 ± 0.03. The full deployment configuration and on-device metrics are summarized in Table 14.

The selected checkpoint was synthesized using the ai8x-synthesis framework (version 20240705, commit 6e7a784), which converts the trained PyTorch (version 2.3.0) model into fixed-point C code compatible with the MAX78000 software development kit. The synthesis flow generates quantized network parameters, C source files, and validation vectors required for bit-exact verification between software inference and on-device execution.

Table 14. Final deployed model configuration and quantized (Q8) performance on the MAX78000 platform. Results are reported as mean ± standard deviation across seeds.

Feature	Final Deployed Model
Sensor modality	Bone-conduction (BC)
Feature representation	Single-channel Mel spectrogram
Input shape	1×128^2
Input size	16 KB
Preproc. RAM	70 KB
Total RAM (128 KB)	90 KB
CNN L0 (32 KB)	∼16.5 KB
Segment duration	1 s
Sampling frequency	16 kHz
Input resolution	128×128
Class weights	$[3, 1]$
Data augmentation	Selective SpecAugment $(1, 4)$
Latency (50 MHz)	15–16 ms
Energy/inference	∼20 µJ
MacroF1 (Q8)	0.80 ± 0.05
BalancedAcc (Q8)	0.81 ± 0.05
$\text{Precision}_{\text{cough}}$ (Q8)	0.71 ± 0.07
$\text{Recall}_{\text{cough}}$ (Q8)	0.79 ± 0.11
F1_{cough} (Q8)	0.74 ± 0.06
$\text{F1}_{\text{non-cough}}$ (Q8)	0.86 ± 0.03
AUC (Q8)	0.89 ± 0.03

To further examine modality-specific behavior, the final deployed BC model was compared with a matched air-conduction reference model using fine-grained class-wise prediction rates, as summarized in Table 15. The two modalities achieved similar true cough detection rates ($78.8 \pm 10.7\%$ for BC vs. $77.3 \pm 13.1\%$ for AC), and their overall AUC values were also comparable (BC: 0.89 ± 0.03; AC: 0.90 ± 0.04). A detailed interpretation of these

modality-specific results, including the non-uniform trade-off profile and its implications for wearable deployment, is provided in Section 7.2.

Table 15. Class-wise prediction rates (%) for the air-conduction (AC) reference model and the final deployed bone-conduction (BC) model. Each row corresponds to the true fine-grained event label; each column pair reports the percentage predicted as cough and non-cough. Results are reported as mean ± standard deviation across seeds. The cough row is indicated by the gray shaded row.

True Event	Air-Conduction (AC)		Bone-Conduction (BC)	
	Pred. Cough	Pred. Non-Cough	Pred. Cough	Pred. Non-Cough
arm_move	8.8 ± 8.1	91.2 ± 8.1	10.8 ± 10.1	89.2 ± 10.1
cough	77.3 ± 13.1	22.7 ± 13.1	78.8 ± 10.7	21.2 ± 10.7
researcher	5.5 ± 7.5	94.5 ± 7.5	6.8 ± 6.6	93.2 ± 6.6
rubbing	10.9 ± 8.9	89.1 ± 8.9	4.8 ± 5.6	95.2 ± 5.6
sneeze	40.4 ± 27.4	59.6 ± 27.4	51.6 ± 24.3	48.4 ± 24.3
speech	8.0 ± 8.0	92.0 ± 8.0	3.2 ± 3.3	96.8 ± 3.3
throat_clear	52.3 ± 29.5	47.7 ± 29.5	73.8 ± 16.0	26.2 ± 16.0

7. Discussion

7.1. Comparison with Prior Edge-Deployed Cough Detection Systems

Table 16 focuses specifically on cough detection systems that target edge or embedded deployment, and is intentionally limited to this subset of prior work. Broader comparisons with smartphone-based or cloud-dependent systems are omitted as their evaluation conditions and hardware constraints differ fundamentally from the deployment target of the present work. A wider survey of edge-AI cough detection datasets and algorithms, including the multimodal dataset from Orlandic et al. [16] and the Cough-E system [15], provides additional context for situating the present contribution within the broader field.

Table 16 summarizes the performance of cough detection systems designed for edge or embedded deployment. BCough achieves a cough F1-score of 0.74 and AUC of 0.89 with full hardware deployment on the MAX78000 at 15–16 ms latency and approximately 20 µJ per inference. It is important to note that the systems listed in Table 16 differ substantially in their evaluation conditions, sensor modalities, datasets, and classification protocols, which limits direct numerical comparison. As detailed in the Evaluation Protocol column of Table 16, Albini et al. [15] report sensitivity and precision under an event-based cross-validation protocol with a different class definition and dataset, while Wang et al. [11] use a random within-subject train/test split and Zhang et al. [31] use a leave-one-session-out protocol with IMU-triggered detection on earbud-mounted sensors all of which differ from the LOSO cross-validation used here. These differences in evaluation protocol, sensor modality, and hardware target mean that the numerical performance figures in Table 16 are not directly comparable and should be interpreted accordingly.

The cough F1-score of BCough is lower than some prior systems for several reasons. First, the model is explicitly constrained to approximately 57k parameters and fewer than 350k MAC operations to satisfy the Flash and SRAM limits of the MAX78000, which inherently limits representational capacity relative to models deployed on smartphones or cloud servers. Second, the LOSO cross-validation protocol with five participants produces higher variance and typically lower aggregate performance than within-subject or random-split evaluations used in some prior works. Third, bone-conduction sensing sacrifices spectral information in the 2–6 kHz band in exchange for noise robustness and privacy preservation, which reduces the discriminative features available to the classifier compared with full-bandwidth air-conduction recordings. These trade-offs are by design and reflect the deployment objectives of the present work rather than deficiencies in the approach. BCough

represents the end-to-end demonstration of bone-conduction-based cough detection that satisfies the Flash and SRAM constraints of a microcontroller-class CNN accelerator, bridging the gap between privacy-preserving BC sensing and ultra-low-power embedded AI inference within a single validated hardware–software co-design pipeline.

It should also be noted that IMU-only and BC–IMU multimodal baselines were not evaluated in the present study. Although synchronized IMU data were collected alongside bone-conduction audio for all participants (Section 2), the primary objective of this work is to establish the feasibility of bone-conduction audio alone as a sensing modality for embedded cough detection on the MAX78000. Characterizing the audio-only performance baseline under hardware deployment constraints is a necessary precondition before introducing additional modalities. Each additional sensing modality introduces preprocessing, synchronization, and memory overhead that must be validated against the MAX78000 hardware constraints before evaluation as illustrated by the 2-channel Mel + Δ audio configuration, which already exceeded the available memory budget during preprocessing and CNN input allocation (Section 6.2). Extending this to IMU fusion requires not only validating the combined memory footprint but also characterizing the synchronization latency between the bone-conduction audio pipeline and the 100 Hz IMU stream under real-time inference conditions neither of which has been validated on the current hardware prototype. Evaluating IMU-only and BC–IMU fusion baselines under the same LOSO protocol and deployment constraints therefore constitutes a necessary and well-defined next step to be pursued alongside the planned participant cohort expansion and hardware miniaturization described in Section 8.

Table 16. Comparison of cough detection systems designed for edge or embedded deployment. Systems differ substantially in evaluation conditions, sensor modalities, datasets, and classification protocols; direct numerical comparison should be interpreted with caution. LOSO = leave-one-subject-out; CV = cross-validation; AC = air-conduction; BC = bone-conduction; Sens. = sensitivity/recall; Prec. = precision; Acc. = accuracy.

Reference	HW-Aware	Evaluation Protocol	Performance	Sensors	Model
Albini, 2025 [15]	Yes	Event-based CV; 20 healthy subjects; different class definition	Sens. = 71.0% Prec. = 78.0% F1 = 77.7%	AC microphone, kinematic	Two XGBoost classifiers
Wang, 2022 [11]	No	Random train/test split; within-subject; earbud-mounted sensors	Acc. = 78.5% F1 = 77.0%	Four AC microphones	CNN
Zhang, 2022 [31]	No	Leave-one-session-out CV; IMU-triggered detection; earbud platform	AUC = 77.0%	AC microphone, kinematic	Multi-Nearest Center Classifier
BCough (ours)	Yes	LOSO CV; 5 healthy subjects; segment-level binary classification	Prec. = 71.0% Sens. = 79.0% F1 = 74.0% AUC = 89.0%	BC microphone	CNN

7.2. BC vs. AC Modality Trade-Offs

Table 17 presents the aggregate quantized performance of the final deployed BC model alongside a matched AC reference model trained and evaluated under identical conditions. At the aggregate level, both modalities perform similarly across all reported metrics. MacroF1 (0.81 vs. 0.80), balanced accuracy (0.81 vs. 0.81), and F1-scores for both classes are nearly identical, and the AUC values differ by only one percentage point (0.90 for AC vs. 0.89 for BC). These aggregate results alone would suggest that the two modalities are largely interchangeable.

Table 17. Aggregate quantized (Q8) performance comparison between the final deployed bone-conduction (BC) model and the matched air-conduction (AC) reference model. Results are reported as mean ± standard deviation across 5 LOSO folds and 4 seeds ($n = 20$).

Metric	AC (Q8)	BC Final (Q8)
MacroF1	0.81 ± 0.05	0.80 ± 0.05
BalancedAcc	0.81 ± 0.05	0.81 ± 0.05
$\text{Precision}_{\text{cough}}$	0.75 ± 0.10	0.71 ± 0.07
$\text{Recall}_{\text{cough}}$	0.77 ± 0.13	0.79 ± 0.11
F1_{cough}	0.75 ± 0.07	0.74 ± 0.06
$\text{F1}_{\text{non-cough}}$	0.87 ± 0.04	0.86 ± 0.03
AUC	0.90 ± 0.04	0.89 ± 0.03

However, the class-wise prediction rates in Table 15 reveal a non-uniform trade-off profile that is more informative than the aggregate metrics alone. The BC model substantially reduces false cough detections for speech ($3.2 \pm 3.3\%$ vs. $8.0 \pm 8.0\%$) and rubbing ($4.8 \pm 5.6\%$ vs. $10.9 \pm 8.9\%$). This reduction reflects the mechanical coupling of the BC sensor to the body surface, which attenuates externally sourced acoustic interference that would otherwise contaminate an air-conduction recording. At the same time, BC recall for true cough events (0.79 ± 0.11) is slightly higher than AC recall (0.77 ± 0.13), indicating that BC misses fewer actual cough events. This combination fewer false positives from speech and rubbing, and fewer missed coughs represents the primary practical advantage of BC sensing for wearable deployment, where bystander speech, clothing friction, and surface-contact artifacts are common failure modes for air-conduction microphones.

Furthermore, the deployed BC model achieves a cough recall of 0.79 ± 0.11, which exceeds its cough precision of 0.71 ± 0.07, indicating that the model is biased toward sensitivity over specificity for the cough class. This operating point is preferable for a health monitoring application where missed coughs are more costly than occasional false alarms.

Conversely, the BC model shows higher confusion for physiologically similar internal events. The cough prediction rate for throat-clear events increased from $52.3 \pm 29.5\%$ (AC) to $73.8 \pm 16.0\%$ (BC), and for sneeze events from $40.4 \pm 27.4\%$ to $51.6 \pm 24.3\%$. This pattern arises because throat clearing and sneezing generate body-conducted vibrations that are spectrally and temporally similar to cough transients, making them harder to distinguish via BC sensing alone. The underlying mechanism is the shared biomechanical vibration pathway in the neck and chest: both events produce short-duration, high-energy pressure transients that couple into the sternum through the same tissue routes as a cough. The resulting BC spectrograms are therefore difficult to distinguish from true cough events at the segment level. Notably, BC precision for the cough class (0.71 ± 0.07) is lower than AC (0.75 ± 0.10), primarily driven by this increased confusion with internal respiratory events.

Returning to the modality comparison, BC and AC exhibit complementary error profiles that together motivate future multimodal integration. BC is more robust to externally coupled interference, such as bystander speech and clothing friction, while AC better discriminates internally generated cough-like events such as throat clearing and sneezing. This complementary behavior strongly motivates multimodal fusion of BC and AC signals as a direction for future work, as combining both modalities could simultaneously address the limitations of each. Prior work from our group has demonstrated the effectiveness of audio–IMU fusion for robust cough detection [23,25], providing a foundation for extending this approach to BC–AC–IMU multimodal systems in future implementations.

8. Conclusions and Future Work

This paper presented BCough, a bone-conduction-based embedded AI platform for on-device cough detection that integrates sensing, model development, quantization, and hard-

ware deployment within a unified edge AI pipeline. By leveraging the privacy-preserving and body-coupled properties of bone-conduction sensing, the proposed approach reduces sensitivity to environmental interference and bystander speech without relying on conventional air-conduction microphones.

A systematic experimental study was conducted to optimize the full deployment pipeline, including preprocessing design, feature representation, class weighting, and augmentation strategy. The optimized configuration used 1 s segments, a sampling frequency of 16 kHz, a 128×128 single-channel Mel-spectrogram input, class weights of $[3, 1]$, and Selective SpecAugment with a $(1, 4)$ masking configuration. Under quantized evaluation, the final model achieved a MacroF1 of 0.80 ± 0.05, balanced accuracy of 0.81 ± 0.05, cough precision of 0.71 ± 0.07, cough recall of 0.79 ± 0.11, cough F1-score of 0.74 ± 0.06, non-cough F1-score of 0.86 ± 0.03, and AUC of 0.89 ± 0.03. These results demonstrate that careful co-design of preprocessing, feature selection, and training strategy is essential for robust embedded deployment.

The final model was successfully deployed on the MAX78000 neural accelerator, achieving real-time inference with a latency of 15–16 ms and an on-chip energy consumption of approximately 20 μJ per inference. A custom four-layer hardware prototype was developed to support this study, integrating the MAX78000, power management circuitry, bone-conduction sensing interface, audio codec, IMU connectivity, and SD-card logging.

These results confirm the feasibility of low-power, privacy-preserving cough detection using bone-conduction sensing and a microcontroller-class neural accelerator within an initial five-participant study. Two broader findings merit emphasis. The best floating-point configuration is not necessarily the best deployment configuration; the final embedded model must be selected according to quantized performance and hardware compatibility jointly. Additionally, the BC and AC modalities exhibit complementary error profiles BC is more robust to externally coupled interference, while AC better discriminates internally generated cough-like events which motivates multimodal fusion as a priority for future work. It should be noted that all evaluations were performed at the segment level on pre-segmented one-second windows under controlled acquisition conditions with five participants. Extension to continuous monitoring, diverse user populations, and real-world environments remains as future work.

Future work will extend this platform in several directions. Multimodal fusion of bone-conduction signals with IMU measurements will be investigated to improve robustness to cough-like confounding events, such as throat clearing and sneezing, as motivated by prior work from our group demonstrating the complementary role of inertial sensing in wearable cough detection [23,25]. Richer temporal modeling and event-counting strategies will be explored to support continuous cough monitoring rather than segment-level detection alone. Finally, a miniaturized battery-powered chest-patch implementation integrating sensing, embedded AI, and wireless operation is under development to enable long-term wearable respiratory monitoring in real-world settings.

Author Contributions: Conceptualization, M.S. and E.J.L.; methodology, M.S.; software, M.S., A.S. and S.D.; validation, M.S., A.S. and S.D.; formal analysis, M.S.; Hardware, M.S., S.D. and J.T.; investigation, M.S., J.d.l.V., A.S. and S.D.; data collection and curation, J.d.l.V.; writing—original draft preparation, M.S.; writing—review and editing, M.S. and E.J.L.; visualization, M.S.; supervision, E.J.L. and M.D.; project administration, E.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation (NSF), Alexandria, VA, USA, under awards IIS-1915599 and IIS-2037328.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of North Carolina State University for the study entitled "Synchronized Acoustic and Motion Sensing for Cough and Speech Analysis" (IRB Protocol 28453).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. No identifiable participant information is included in this manuscript.

Data Availability Statement: The data supporting the findings of this study will be made available in a public repository upon acceptance of the manuscript. Until then, the data are available from the corresponding author on reasonable request, subject to privacy and ethical restrictions.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-4, OpenAI, San Francisco, CA, USA) and Claude (Claude Sonnet 4.6, Anthropic, San Francisco, CA, USA) for language editing and manuscript refinement. The authors reviewed and edited all AI-assisted outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AC	Air-conduction
BC	Bone-conduction
CNN	Convolutional neural network
IMU	Inertial measurement unit
MFCC	Mel-frequency cepstral coefficients
PMIC	Power management integrated circuit
PTQ	Post-training quantization
QAT	Quantization-aware training
Q8	8-bit quantized inference
SRAM	Static random-access memory
STFT	Short-time Fourier transform

References

1. GBD 2015 Chronic Respiratory Disease Collaborators. Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015. *Lancet Respir. Med.* **2017**, *5*, 691–706. [CrossRef] [PubMed]
2. Alqudaihi, K.S.; Aslam, N.; Khan, I.U.; Almuhaideb, A.M.; Alsunaidi, S.J.; Ibrahim, N.M.A.R.; Alhaidari, F.A.; Shaikh, F.S.; Alsenbel, Y.M.; Alalharith, D.M.; et al. Cough sound detection and diagnosis using artificial intelligence techniques: Challenges and opportunities. *IEEE Access* **2021**, *9*, 102327–102344. [CrossRef] [PubMed]
3. Birring, S.S.; Fleming, T.; Matos, S.; Raj, A.A.; Evans, D.H.; Pavord, I.D. The Leicester Cough Monitor: preliminary validation of an automated cough detection system in chronic cough. *Eur. Respir. J.* **2008**, *31*, 1013–1018. [CrossRef]
4. Sun, X.; Lu, Z.; Hu, W.; Cao, G. SymDetector: Detecting sound-related respiratory symptoms using smartphones. In *UbiComp '15: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 97–108.
5. Drugman, T.; Urbain, J.; Bauwens, N.; Chessini, R.; Valderrama, C.; Lebecque, P.; Dutoit, T. Objective study of sensor relevance for automatic cough detection. *IEEE J. Biomed. Health Inform.* **2013**, *17*, 699–707. [CrossRef] [PubMed]
6. Kvapilova, L.; Boza, V.; Dubec, P.; Majernik, M.; Bogar, J.; Jamison, J.; Goldsack, J.C.; Kimmel, D.J.; Karlin, D.R. Continuous sound collection using smartphones and machine learning to measure cough. *Digit. Biomark.* **2020**, *3*, 166–175.
7. McGuinness, K.; Holt, K.; Dockry, R.; Smith, J. P159 Validation of the VitaloJAK 24-hour ambulatory cough monitor. *Thorax* **2012**, *67*, A131. [CrossRef]

8. Urban, C.; Kiefer, A.; Conradt, R.; Kabesch, M.; Lex, C.; Zacharasiewicz, A.; Kerzel, S. Validation of the LEOSound monitor for standardized detection of wheezing and cough in children. *Pediatr. Pulmonol.* **2022**, *57*, 551–559. [CrossRef] [PubMed]
9. Liaqat, D.; Liaqat, S.; Chen, J.L.; Sedaghat, T.; Gabel, M.; Rudzicz, F.; de Lara, E. CoughWatch: Real-world cough detection using smartwatches. In *Proceedings of the ICASSP 2021—IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 8333–8337.
10. Nemati, E.; Zhang, S.; Ahmed, T.; Rahman, M.M.; Kuang, J.; Gao, A. CoughBuddy: Multi-modal cough event detection using earbuds platform. In *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*; IEEE: New York, NY, USA, 2021; pp. 1–4.
11. Wang, Y.; Zhang, X.; Chakalasiya, J.M.; Xu, X.; Jiang, Y.; Li, Y.; Patel, S.; Shi, Y. HearCough: Enabling continuous cough event detection on edge computing hearables. *Methods* **2022**, *205*, 53–62. [CrossRef] [PubMed]
12. Singh, R.; Gill, S.S. Edge AI: A survey. *Internet Things Cyber-Phys. Syst.* **2023**, *3*, 71–92. [CrossRef]
13. Lee, Y.L.; Tsung, P.K.; Wu, M. Techology trend of edge AI. In Proceedings of the 2018 International Symposium on VLSI Design, Automation and Test (VLSI-DAT), Hsinchu, Taiwan, 16–19 April 2018; pp. 1–2. [CrossRef]
14. Lin, J.; Zhu, L.; Chen, W.M.; Wang, W.C.; Han, S. Tiny Machine Learning: Progress and Futures. *IEEE Circuits Syst. Mag.* **2023**, *23*, 8–34. [CrossRef]
15. Albini, S.; Orlandic, L.; Dan, J.; Thevenot, J.; Teijeiro, T.; Constantinescu, D.A.; Atienza, D. Cough-E: A multimodal, privacy-preserving cough detection algorithm for the edge. *IEEE J. Biomed. Health Inform.* **2025**, 1–15. [CrossRef]
16. Orlandic, L.; Thevenot, J.; Teijeiro, T.; Atienza, D. A Multimodal Dataset for Automatic Edge-AI Cough Detection. In *Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Sydney, Australia, 24–27 July 2023*; IEEE: New York, NY, USA, 2023; pp. 1–7. [CrossRef]
17. Analog Devices. MAX78000 Datasheet and Product, 2021. Available online: https://www.analog.com/en/products/max78000.html (accessed on 26 April 2026).
18. STMicroelectronics. STM32G4 Series, 2024. Available online: https://www.st.com/en/microcontrollers-microprocessors/stm32g4-series.html (accessed on 26 April 2026).
19. Wang, M.; Chen, J.; Zhang, X.-L.; Rahardja, S. End-to-End Multi-Modal Speech Recognition on an Air and Bone Conducted Speech Corpus. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 513–524. [CrossRef]
20. He, L.; Hou, H.; Shi, S.; Shuai, X.; Yan, Z. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables. In *MobiSys '23: Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 14–27. [CrossRef]
21. Liu, C.; Wang, X.; Xiao, J.; Zhou, J.; Wu, G. A Piezoelectric Micromachined Ultrasonic Transducer-Based Bone Conduction Microphone System for Enhancing Speech Recognition Accuracy. *Micromachines* **2025**, *16*, 613. [CrossRef] [PubMed]
22. Otoshi, T.; Nagano, T.; Izumi, S.; Hazama, D.; Katsurada, N.; Yamamoto, M.; Tachihara, M.; Kobayashi, K.; Nishimura, Y. A novel automatic cough frequency monitoring system combining a triaxial accelerometer and a stretchable strain sensor. *Sci. Rep.* **2021**, *11*, 9973. [CrossRef]
23. Chen, Y.; Barahona, J.; Eldho, I.; Yu, Y.; Muhammad, R.; Kutsche, B.; Hernandez, M.L.; Carpenter, D.; Bozkurt, A.; Lobaton, E. Robust multimodal cough and speech detection using wearables: A preliminary analysis. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; IEEE: New York, NY, USA, 2024; pp. 1–6.
24. Diab, M.S.; Rodriguez-Villegas, E. Feature evaluation of accelerometry signals for cough detection. *Front. Digit. Health* **2024**, *6*, 1368574. [CrossRef]
25. Chen, Y.; Xiang, F.; Hernandez, M.L.; Carpenter, D.; Bozkurt, A.; Lobaton, E. Robust Multimodal Cough Detection with Optimized Out-of-Distribution Detection for Wearables. *IEEE J. Biomed. Health Inform.* **2026**, *30*, 3485–3494. [CrossRef] [PubMed]
26. Syntiant. V2S Voice Vibration Sensor, 2024. Available online: https://www.syntiant.com/v2s (accessed on 26 April 2026).
27. Knowles Corporation. SPH0141LM4H-1 SiSonic Surface Mount MEMS Microphone, 2024. Available online: https://www.knowles.com/docs/default-source/default-document-library/mic-selection-guide-r5.pdf (accessed on 26 April 2026).
28. MbientLab Inc. MetaMotionS Documentation, 2021. Available online: https://mbientlab.com/metamotions/ (accessed on 26 April 2026).
29. Richard, J.; Zimpfer, V.; Roth, S. Effect of Bone-Conduction Microphone Location and Mouth Opening on Transfer Function between Oral Cavity Sound Pressure and Skin Acceleration. In Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023, Turin, Italy, 11–15 September 2023. [CrossRef]

30. Liu, C.; Wang, X.; Xie, Y.; Wu, G. Bone Conduction Pickup Based on Piezoelectric Micromachined Ultrasonic Transducers. In *2023 IEEE 36th International Conference on Micro Electro Mechanical Systems (MEMS)*; IEEE: New York, NY, USA, 2023; pp. 949–952. [CrossRef]
31. Zhang, S.; Nemati, E.; Dinh, M.; Folkman, N.; Ahmed, T.; Rahman, M.; Kuang, J.; Alshurafa, N.; Gao, A. Coughtrigger: Earbuds IMU Based Cough Detection Activator Using An Energy-Efficient Sensitivity-Prioritized Time Series Classifier. In *ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2022; pp. 1–5. [CrossRef]