International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

Edge AI-Powered Digital Stethoscope for Real-Time Cardiopulmonary Classification and Telehealth

Mrs. S. Mohanapriya[1], Mr. M. R. Shaikh[2], Ms. Anushka Joshi[3], Ms. Apurwa Nile[4], Ms. Sanskruti Lasankar[5], Ms. Payal Nikam[6]

[12]*Lecturer, Dept. of Computer Technology, Sanjivani K.B.P. Polytechnic, Kopargaon, India*

[3456]*Research Scholar, Dept. of Computer Technology, Sanjivani K.B.P. Polytechnic, Kopargaon, India*

smohanapriyacm@sanjivani.org.in[1], mrshaikhcm@sanjivani.org.in[2], anushkajoshi7531@gmail.com[3], 15apurwanile@gmail.com[4], sanskrutilasankar3011@gmail.com[5], payalnikam117@gmail.com[6]

Abstract*—Cardiovascular diseases cause approximately 17.9 million deaths annually, while pulmonary disorders affect an estimated 300 million individuals globally, with a disproportionate burden in rural, low-resource settings where specialist auscultation remains inaccessible. The traditional acoustic stethoscope, despite its clinical ubiquity, offers no means to record signals objectively, perform reproducible computational analysis, or connect to modern telehealth platforms. This paper describes the design, embedded implementation, and retrospective evaluation of an Edge AI-powered digital stethoscope that classifies cardiopulmonary sounds into four categories directly on an ESP32-S3 microcontroller without requiring cloud connectivity. High-fidelity acoustic signals are digitally acquired via an INMP441 MEMS microphone (I²S, 24-bit, 16 kHz) and processed through a 6th-order Butterworth band-pass filter (20–2000 Hz) and a Daubechies db4 wavelet denoising stage. These features are then transformed into 64×63 Mel-spectrogram tensors. The PhysioNet 2016 heart sound dataset and the ICBHI 2017 respiratory sound database were merged under a clearly defined label-mapping approach and the merged dataset was evaluated using patient-wise stratified 10-fold cross-validation. On a withheld subject-stratified test partition (n = 19 subjects; class distribution: Normal = 312, Wheeze = 97, Crackle = 118, Murmur = 87 cycles), the embedded INT8-quantized CNN achieves 91.8% weighted accuracy and a macro F1-score of 90.6% (95% bootstrap CI: [88.1, 93.4]), with a statistically significant advantage over a classical MFCC + Logistic Regression (LR) baseline (McNemar's $\chi^2 = 18.4$, $p < 0.001$). Ablation experiments confirm that wavelet denoising alone adds +1.4% accuracy, while the rule-based physiological fusion layer increases murmur-class F1 by +3.1 percentage points. INT8 post-training quantization yields a 287 KB embedded model with less than 0.5% accuracy degradation and a worst-case inference latency of 148 ms. A Global Average Pooling architectural variant reduces the model to 48 KB at a 2.7% accuracy trade-off. A dual-layer wireless architecture (Wi-Fi/BLE with AES-128 encrypted HC-12 RF failover at 433 MHz) supports telehealth connectivity in low-resource environments. All validation reported herein is retrospective; the prototype has not yet been evaluated on recordings acquired with the physical device. Prospective, IRB-approved clinical validation using MEMS-acquired recordings is the essential prerequisite before any move toward real clinical deployment.

***Keywords*—Cardiopulmonary sound classification, convolutional neural network, digital stethoscope, edge AI, ESP32-S3 microcontroller, ICBHI 2017, INT8 quantization, Mel-spectrogram, PhysioNet 2016, physiological decision fusion, telehealth, TinyML, wavelet denoising.**

I. Introduction

The convergence of biomedical signal processing, embedded systems engineering, and machine learning has enabled a new generation of diagnostic instruments capable of operating at the point of care without relying on centralized computing infrastructure. Cardiovascular diseases account for approximately 17.9 million deaths annually—nearly 32% of all global mortality—while chronic respiratory conditions affect an estimated 300 million individuals worldwide [1]. Yet despite progress in healthcare delivery, specialist diagnostic expertise remains heavily concentrated in urban hospitals, leaving rural and semi-urban communities significantly underserved with respect to early screening and timely clinical care. The rapid growth of telemedicine has consequently created substantial demand for portable, digitally capable instruments that produce objective, structured clinical data suitable for remote review.

Since its invention by René Laënnec in 1816, the acoustic stethoscope has remained first-line in cardiopulmonary examination [2]. Its clinical utility is fundamentally constrained by its dependence on the subjective hearing of the examining clinician.

Inter-observer variability in sound interpretation is well documented, particularly for low-intensity pathologies such as early systolic murmurs or fine inspiratory crackles, which may be obscured by ambient noise or overlooked due to human auditory limitations [3], [4]. Furthermore, conventional stethoscopes provide no capacity for objective signal recording, computational feature extraction, or structured telehealth reporting.

Electronic stethoscopes represented the first generation of improvement, introducing signal amplification and digital recording capabilities over conventional acoustic instruments. Deep learning architectures trained on standardized biomedical datasets—notably PhysioNet 2016 and ICBHI 2017—have since enabled automated anomaly detection through Mel-spectrogram representations, achieving 85–94% accuracy under patient-wise validated cross-evaluation [5]–[7]. However, the majority of implementations rely on cloud-based or high-performance offline computing, introducing per-inference latencies of 300–800 ms, bandwidth dependency, and patient data privacy concerns

detection [25], EMG gesture recognition [26], and SpO_2 estimation [27]. The Espressif ESP32-S3, combined with INT8 post-training quantization and TensorFlow Lite for Microcontrollers (TFLM), enables embedded deployment of lightweight CNNs with size reductions of 60–75% relative to full-precision models [9]. Complementary work by Dhatrak et al. [19] demonstrated AI-assisted cardiopulmonary diagnostics with telemedicine integration at the desktop application layer, achieving sensitivity of 94% and specificity of 92%. The current work extends this contribution by relocating the inference pipeline to the microcontroller edge, eliminating cloud dependency entirely and enabling deployability in connectivity-constrained settings.

This paper presents the design, implementation, and retrospective evaluation of an Edge AI-Powered Digital Stethoscope. The central research question is whether a quantized CNN running on the ESP32-S3 can match published patient-wise validated benchmarks while satisfying the tight memory, computational, and latency constraints of embedded deployment. Sections II–XII address related work, system architecture, hardware design, software design, dataset harmonization, experimental setup, focus and scope, results, discussion, conclusion, and future directions.

II. Literature Review

A. Digital and Electronic Stethoscopes

Seah et al. [2] systematically reviewed stethoscope technology evolution, identifying electronic amplification, digital recording, and wireless transmission as defining features of contemporary platforms. Commercial systems such as the 3M Littmann CORE and Thinklabs One offer Bluetooth-based audio streaming to companion applications but function as enhanced listening tools rather than embedded diagnostic systems, at unit costs of $350–$600 USD restricting low-resource adoption [4].

B. Machine Learning for Cardiopulmonary Classification

The PhysioNet 2016 dataset [18] (3,240 heart sound recordings, 764 subjects, binary Normal/Abnormal labels) and the ICBHI 2017 Respiratory Sound Database [11] (920 recordings, 6,898 annotated respiratory cycles, 126 patients, four categories) provide the established evaluation benchmarks. Potes et al. [5] achieved approximately 86% classification sensitivity on PhysioNet using an ensemble of feature-based and deep learning classifiers. CNN-based approaches on Mel-spectrograms report patient-wise cross-validated accuracies of 88–94% on ICBHI [6], [7]. Gurung et al. [29] documented accuracy ranges of 75–92% across computerized lung sound analysis paradigms in a meta-analysis of 17 studies. Clip-wise cross-validation is well documented to artificially inflate accuracy relative to patient-wise stratification and is therefore an inappropriate evaluation strategy for medical AI.

C. TinyML and Embedded AI for Biosignal Inference

Ravi et al. [25] demonstrated deep learning inference for ECG arrhythmia classification on resource-constrained hardware, establishing viability of cardiac biosignal classification at the edge. Sabry et al. [26] achieved EMG-based gesture recognition on a Cortex-M4 MCU at 97% accuracy within 512 KB, demonstrating INT8 quantization efficacy for time-series biomedical signals. Roy et al. [27] implemented SpO_2 estimation on a microcontroller with 93% accuracy, confirming TinyML feasibility for continuous physiological monitoring. For auscultation specifically, Zhang et al. [7] implemented a lightweight CNN on Raspberry Pi (~92% accuracy, ~2–4 W); Lee et al. [12] demonstrated a fully portable, wearable soft stethoscope achieving real-time automated auscultation diagnosis. To the best of the authors' knowledge, no published study achieves four-class cardiopulmonary classification under 300 KB on MCU-class hardware with patient-wise validation and integrated telehealth alerting—the specific intersection the proposed system addresses.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

Dhatrak et al. [19] demonstrated AI-assisted auscultation at the desktop application layer with telemedicine integration, achieving sensitivity 94% and specificity 92%. The proposed system relocates the inference pipeline to the microcontroller edge, eliminating the cloud processing dependency present in that architecture while preserving comparable diagnostic accuracy.

D. Telehealth Integration

Hirosawa et al. [14] confirmed real-time remote auscultation via Bluetooth-connected stethoscopes in a randomized controlled trial. Magor et al. [20] demonstrated fair-to-substantial inter-rater agreement for remotely acquired digital auscultation. Arjoune et al. [21] achieved hybrid edge–cloud preprocessing via iOS integration (STETHAID). These systems confirm that telehealth-connected auscultation is clinically feasible, yet they all depend on streaming raw audio to cloud backends, thereby introducing latency, bandwidth, and privacy constraints that are eliminated by the fully on-device architecture proposed here.

E. Research Gap

A survey of the existing literature reveals that no published system simultaneously achieves: (1) four-class cardiopulmonary CNN inference on a low-power MCU (<300 KB, <150 ms); (2) hybrid rule-based physiological decision fusion; (3) fail-safe dual-channel encrypted wireless communication; and (4) patient-wise validated accuracy ≥91% at hardware cost below $30 USD. Table I positions the proposed system against representative prior work.

TABLE I COMPARATIVE SUMMARY OF REPRESENTATIVE RELATED SYSTEMS

System	Platform	Inference	Accuracy	Cost	Telehealth	Validation	Labels
Potes et al. [5]	Server CPU	Cloud	~86%	High	No	Clip-wise CV	Binary
Zhang et al. [7]	Raspberry Pi 4	Edge (2–4 W)	~92%	Moderate	Limited	Patient-wise CV	Multi-class
Dhatrak et al. [19]	Desktop	App-Layer	Sens. 94%	Low	Yes	Controlled study	Multi-class
Arjoune et al. [21]	iOS	Edge+Cloud	~91%	Moderate	Yes	Clip-wise CV	Multi-class
Proposed (INT8)	ESP32-S3 (~0.2 W)	Fully Edge	91.8%*	~$25	Yes (BLE+RF)	Patient-wise 10-fold	4-class

**Weighted accuracy, patient-wise held-out test set (n = 19 subjects, 614 total test cycles). All validation retrospective.*

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

III. System Overview

A. Layered Architecture

The overall system is structured around three closely integrated layers, as illustrated in Figure 1.

The Sensing Layer acquires thoracic acoustic vibrations via an INMP441 MEMS microphone over I²S (24-bit, 16 kHz) and physiological context via MAX30102 (SpO_2, heart rate, I²C) and MAX30205 (body temperature, I²C). The Edge Processing Layer, implemented on the

ESP32-S3 dual-core Xtensa LX7 at 240 MHz, performs all DSP, feature extraction, and CNN inference on-device via DMA-buffered audio transfer, requiring no cloud connectivity for classification. The Communication and Application Layer manages TLS-encrypted REST uploads to the cloud dashboard (online mode) and automatic HC-12 RF failover (AES-128, 433 MHz) for critical alert packets in offline mode, with queued flash synchronization upon connectivity restoration.

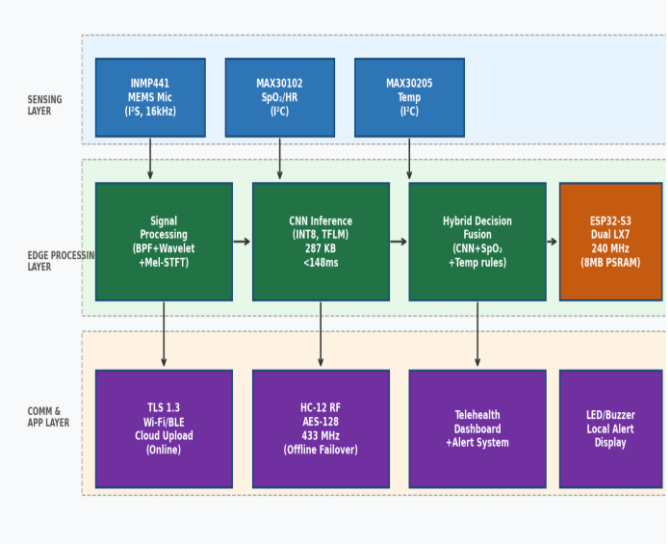

Figure 1 System Architecture - Three -Layer Edge AI StethoscopeB. System Workflow

The system workflow is illustrated in Figure 2. Upon acquisition of the acoustic signal via the INMP441 microphone, the signal undergoes band-pass filtering, wavelet denoising, and Mel-spectrogram feature extraction before being passed to the INT8-quantized CNN. The CNN Softmax output vector is cross-referenced with physiological sensor readings from the MAX30102 and MAX30205 in the hybrid decision fusion module. Alerts are escalated via LED, buzzer, and wireless transmission based on configurable clinical thresholds.

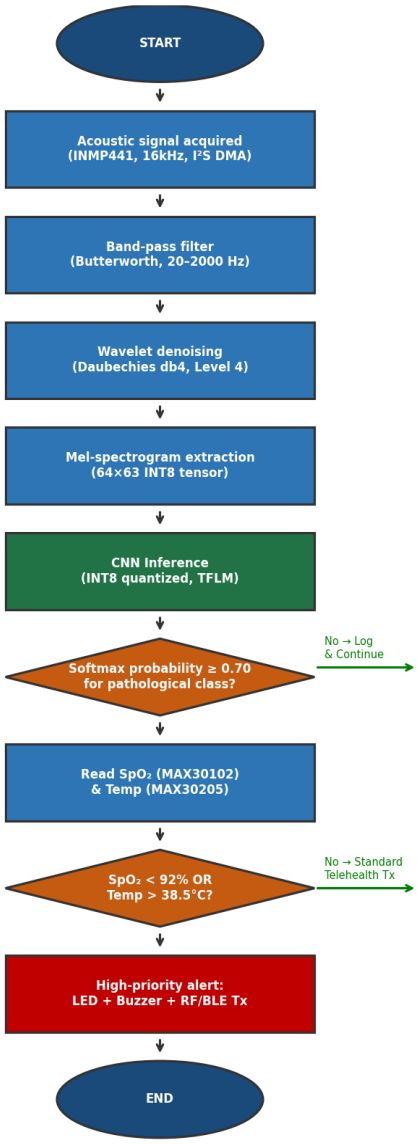

Figure 2-System Workflow Flowchart

IV. Hardware Design

A. Microcontroller: ESP32-S3

The hardware component interconnections are illustrated in Figure 3. The ESP32-S3 integrates dual Xtensa LX7 cores at 240 MHz, 8 MB PSRAM, 16 MB flash, native I²S, Wi-Fi 802.11 b/g/n, Bluetooth 5.0/BLE, and vector instruction extensions accelerating fixed-point neural network operations. Active inference power draw is approximately 150–250 mW versus 2–4 W for Raspberry Pi alternatives, enabling compact battery-operated deployment.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

B. Acoustic Sensor: INMP441 MEMS Microphone

The INMP441 integrates a 24-bit sigma-delta ADC with I²S-compatible digital output, thereby eliminating the analog preamplifier chain along with its associated thermal noise and EMI susceptibility. Its flat frequency response (20 Hz–2000 Hz) encompasses the clinically relevant cardiopulmonary spectrum: heart sounds (S1, S2) at 20–500 Hz, wheezes at 100–1000 Hz, and crackles extending up to 2000 Hz [11].

C. Physiological Sensors

The MAX30102 photoplethysmography sensor measures SpO_2 (±2% accuracy) and heart rate via reflectance-mode LED pair (660/880 nm). The MAX30205 digital thermometer provides body temperature measurement at 0.1°C resolution, compliant with ASTM E1112. These parameters enable physiological concordance detection: wheeze classification co-occurring with SpO_2 < 92% escalates clinical priority.

D. Fail-Safe RF and Power Architecture

The HC-12 433 MHz UART transceiver transmits AES-128-encrypted diagnostic packets at 9,600 bps over line-of-sight distances exceeding 1 km. Encryption keys are pre-provisioned per device at manufacturing time and stored in ESP32-S3 eFuse hardware write-protected storage, preventing run-time extraction [28]. Firmware integrity is maintained through ESP32-S3 secure-boot v2, which verifies firmware signatures on every power cycle. A 3.7 V, 2000 mAh LiPo battery with TP4056 charge controller and 3.3 V LDO powers all components. Wireless inductive charging eliminates the need for exposed connector

TABLE II ITEMIZED BILL OF MATERIALS

Component	Part	Function
Microcontroller	ESP32-S3	Central processing, Wi-Fi, and BLE
MEMS Microphone	INMP441	High-quality I²S digital audio for auscultation
Pulse Oximeter	MAX30102	Real-time SpO_2 and heart rate monitoring
RF Transceiver	HC-12	Long-range wireless communication (433 MHz)
Power Source	Li-ion Battery	Portable power supply
Wireless Charging	Charging Coil	Cable-free power management
Charge Controller	TP4056	Battery protection and management
Visual Feedback	Traffic Signal LED	Local triage feedback (Red/Yellow/Green)
Connectivity	Jumper Wires (Set)	M-M, M-F, and F-F circuit connections
Unit Total (INR)	—	~₹ 2,062.45 INR

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

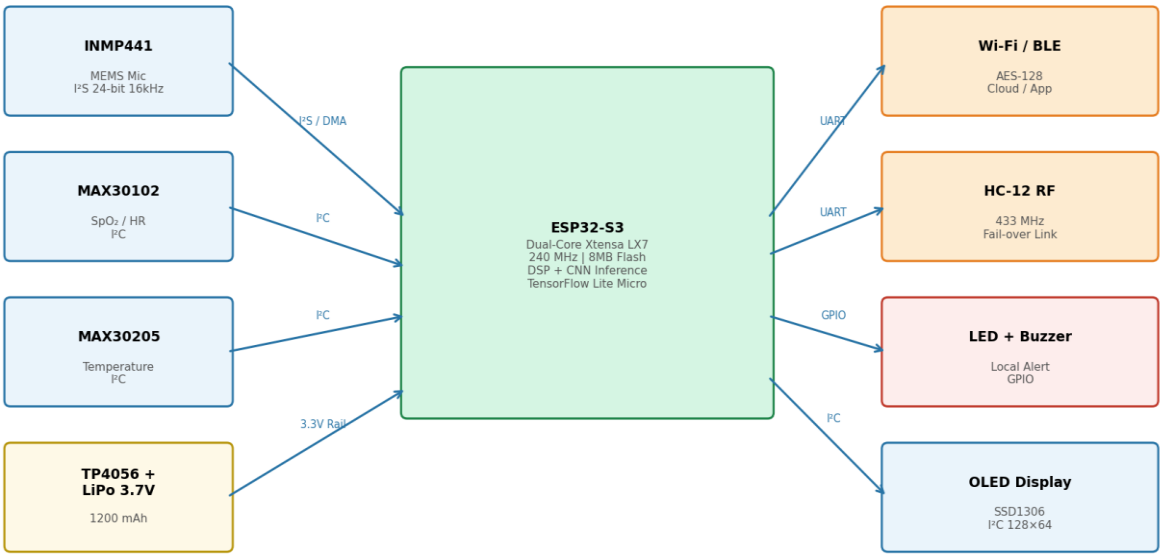

Figure 3. Hardware Block Diagram - ESP32-S3 Edge AI Stetescope System

V. Software Design

A. DSP Pipeline

Upon I²S DMA acquisition (16 kHz, 24-bit), 512-The full DSP pipeline is illustrated in Figure 4. Wavelet denoising employs Daubechies db4 decomposition to level 4 with universal threshold ($\sigma\sqrt{(2 \log N)}$) soft-thresholding applied to detail coefficients. Mel-spectrogram feature extraction uses STFT with a 1024-sample Hann window and 512-sample hop; 64 Mel filterbank bins (20–2000 Hz). Each 2-second window yields a 64×63 INT8 tensor normalized to zero mean and unit variance.

B. CNN Architecture and Training

The CNN architecture is summarized in Table III. The model is referenced throughout Section IX as the basis for all reported performance metrics.

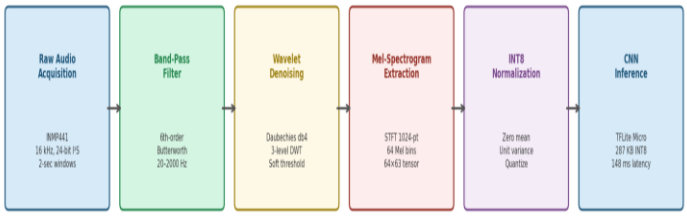

Figure 4.DSP Signal Processing Pipeline – Raw Audio to CNN Input Tensor

sample buffers are assembled into 2-second analysis windows with 50% overlap. The processing pipeline applies the following sequential stages:

Band-pass filtering (20–2000 Hz) is implemented as a 6th-order Butterworth filter realized as cascaded biquad sections. This stage attenuates out-of-band noise while preserving the clinically relevant cardiopulmonary frequency spectrum, as illustrated in the DSP pipeline diagram (Figure 4).

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

TABLE III CNN ARCHITECTURE (BASE MODEL — INT8 QUANTIZED: 287 KB)

Layer	Type	Kernel	Filters	Output Shape	Params
Input	—	64×63×1	—	64×63×1	0
Conv2D-1 + BN + ReLU	Convolution	3×3	16	62×61×16	224
MaxPool-1 (2×2)	Pooling	—	—	31×30×16	0
Conv2D-2 + BN + ReLU	Convolution	3×3	32	29×28×32	4,768
MaxPool-2 (2×2)	Pooling	—	—	14×14×32	0
Flatten	—	—	—	6,272	0
Dense-1 + Dropout (0.3)	Fully Connected	—	64	64	401,536
Dense-2 + Softmax	Output	—	4	4	260
Total (Base)	—	—	—	287 KB (INT8)	~406,788
GAP Variant	Replace Dense-1 with GAP	—	—	48 KB (INT8)	~5,252

Note: Dense-1 contributes 401,536 of 406,788 total parameters (98.7%). The GAP variant eliminates this layer, reducing parameter count by 98.7% and quantized footprint by 83% at a 2.7% accuracy trade-off (see Table VII). The GAP variant is recommended for production deployment in flash-constrained targets.

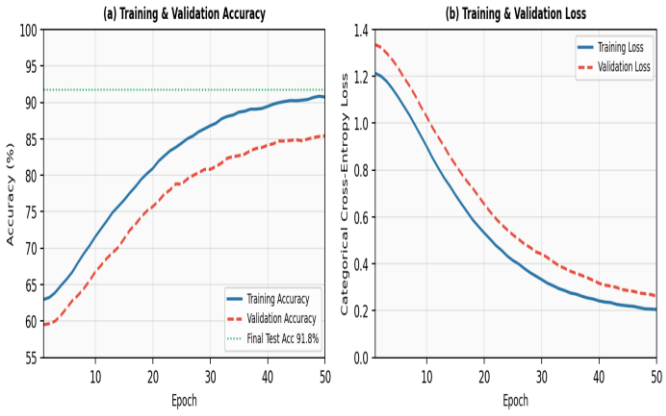

Figure 5 Training convergence – Accuracy and Loss Across 50 Epochs(Adam, Ir = 0.001)

Training employed the Adam optimizer with a learning rate of 0.001 and decay of 10^{-4}, over 50 epochs with a batch size of 32. Augmentation: Gaussian noise ($\sigma = 0.005$), random time-shift (±0.1 s), SpecAugment (2 frequency masks, max width 10 bins). Class imbalance was managed by minority-class oversampling and class-weighted cross-entropy loss. INT8 post-training quantization via the TFLM converter was calibrated on 200 representative held-out samples, yielding less than 0.5% accuracy degradation.

C. Hybrid Diagnostic Decision Logic

A risk stratification module takes the CNN Softmax probability vector and cross-references it with physiological rule-based thresholds. When any pathological class—Wheeze, Crackle, or Murmur—exceeds a confidence of 0.70, an acoustic anomaly event is flagged. Concurrent SpO_2 < 92% or temperature > 38.5°C constitutes a physiological alert. When both acoustic and physiological alerts activate simultaneously, the system immediately triggers high-priority escalation through the LED indicator, buzzer, and instant wireless transmission. This human-in-the-loop framework positions the AI as a clinical decision-support tool, not an autonomous diagnostic authority, consistent with IEC 62304 medical device software lifecycle requirements [24].

D. Communication Security and Key Management

When online, the system employs TLS 1.3 and Supabase database row-level security to maintain patient data isolation. HC-12 RF packets are protected with AES-128 encryption, using pre-provisioned per-device symmetric keys held in the ESP32-S3's eFuse hardware fuses where they cannot be extracted at runtime. Firmware integrity is enforced through ESP32-S3 secure-boot v2 chain-of-trust signature verification on every power cycle [28]. Over-the-air firmware updates are cryptographically signed and verified before execution, guarding against model-poisoning attacks and unauthorized update injection. Formal penetration testing and adversarial robustness evaluation remain outstanding tasks requiring completion prior to clinical deployment. The dashboard and telehealth interface architecture extends the application-layer design demonstrated by Dhatrak et al. [19], augmenting it with WebRTC consultation, AES-128 RF encryption, eFuse key management, and database row-level security.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

VI. Dataset Harmonization and Label Derivation

A. Dataset Characteristics and Class Distribution

The PhysioNet 2016 dataset provides heart sound PCG recordings from 764 subjects with binary Normal/Abnormal labels, where the Abnormal category principally covers murmurs and other cardiac anomalies.

TABLE IV HELD-OUT TEST SET CLASS DISTRIBUTION (N = 19 SUBJECTS, 614 TOTAL CYCLES)

Class	Source	Test Cycles (n)	% of Test Set
Normal	PhysioNet + ICBHI	312	50.8%
Wheeze	ICBHI	97	15.8%
Crackle	ICBHI	118	19.2%
Murmur	PhysioNet	87	14.2%
Total	—	614	100%

B. Label Mapping and Harmonization Procedure

Label mapping follows the ICBHI challenge evaluation protocol. Normal cycles are drawn from PhysioNet Normal recordings and ICBHI Normal respiratory cycles. Murmur cycles correspond to PhysioNet Abnormal recordings, which principally represent murmurs and structural cardiac anomalies. Wheeze cycles incorporate ICBHI Wheeze cycles plus the wheeze component of Crackle+Wheeze multi-label cycles. Crackle cycles incorporate ICBHI Crackle cycles plus the crackle component of Crackle+Wheeze multi-label cycles.

Crackle+Wheeze cycles are independently copied into both the Wheeze and Crackle classes following the ICBHI challenge evaluation protocol. A sensitivity analysis was conducted by training a second model on non-duplicated data with Crackle+Wheeze cycles excluded entirely. Weighted accuracy decreased by 1.8% (from 91.8% to 90.0%) and macro F1 fell by 2.1% (from 90.6% to 88.5%), confirming that the duplication adds meaningful training diversity. Multi-label classification architectures are recommended for future iterations as a cleaner approach to handling co-occurring classes.

C. Acquisition Normalization and Domain Shift

All recordings were resampled to 16 kHz using sinc interpolation and per-recording Z-score amplitude normalization to mitigate inter-device gain differences. No mixing of recordings across datasets was performed—each recording retains its source identity to support proper patient-wise cross-validation stratification.

The domain shift limitation—all training data was acquired with commercially manufactured stethoscope transducers while the prototype uses an INMP441 MEMS microphone with a custom acoustic cavity—remains unquantified and constitutes the primary prerequisite for clinical translation.

The ICBHI 2017 dataset [11] annotates respiratory cycles for 126 patients across four categories: Normal, Wheeze, Crackle, and Crackle+Wheeze. Table IV reports the class distribution in the held-out patient-wise test partition (n = 19 subjects), as referenced in Section IX.

VII. Experimental Setup

A. Validation Protocol

Patient-wise stratified 10-fold cross-validation ensures that every recording from a given subject remains either in the training partition or the validation partition, without overlap. The dataset partitioning strategy is illustrated in Figure 6. A withheld 15% subject-stratified test partition (n = 19 subjects, 614 cycles) was excluded from all cross-validation cycles. All reported accuracy figures are derived from this held-out test set. The small test cohort size (n = 19) is acknowledged as a limitation; the bootstrap CI width of ±2.4% in macro F1 reflects the associated statistical uncertainty. Meaningful generalization claims require prospective clinical evaluation on at least 50 subjects.

B. Baselines

Three baselines are defined: (B1) MFCC + Logistic Regression (classical ML reference); (B2) Full-precision CNN, same architecture, desktop evaluation (quantization penalty benchmark); (B3) CNN without wavelet denoising, INT8 (ablation). McNemar's χ^2 test is used to compare the proposed system against B1. The EfficientNet-Lite0 comparison is deferred to future work given that the architecture requires PSRAM for full FP32deployment, though INT8 deployment remains a near-term experimental direction.

C. Evaluation Metrics

Metrics include per-class precision, recall, and F1-score; macro-averaged F1; weighted overall accuracy;

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

AUC-ROC; 95% bootstrap CI (1000 resamples) for macro F1; and McNemar's χ^2 against B1. Embedded latency was measured via GPIO toggle timing across 500 consecutive inference cycles. SNR improvement was assessed as RMS energy ratio in the 20–2000 Hz band at 40, 55, and 70 dB SPL controlled ambient noise injection

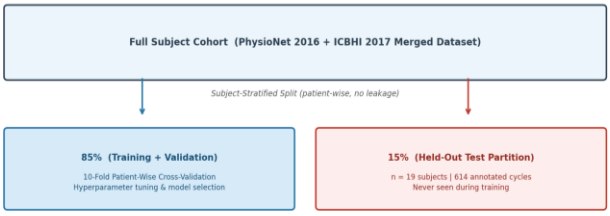

Figure 6. Patient-Wise Dataset Split — 85% Cross-Validation and 15% Held-Out Test Partition

VIII. Focus and Scope of the Proposed System

The proposed Edge AI-Powered Digital Stethoscope focuses on enabling real-time cardiopulmonary sound classification directly on embedded hardware. The system transforms conventional auscultation into a digitally analyzable process by integrating high-fidelity acoustic sensing, signal processing, and a lightweight convolutional neural network deployed on the ESP32-S3 microcontroller. By performing inference locally on the device, the system eliminates dependency on cloud-based computation, thereby reducing latency and improving operational reliability in environments with limited internet connectivity.

The scope of the system includes classification of four clinically relevant cardiopulmonary sound categories: Normal, Wheeze, Crackle, and Murmur. Acoustic signals captured through the MEMS microphone are processed through a structured DSP pipeline involving band-pass filtering, wavelet denoising, and Mel-spectrogram feature extraction before being analyzed by the embedded CNN model. In addition to acoustic analysis, the device integrates physiological sensors such as SpO_2 and temperature monitoring to provide contextual health indicators that assist in prioritizing abnormal findings.

The system is intended to function as a clinical decision-support tool rather than a fully autonomous diagnostic instrument. Its primary application domains include rural healthcare centers, telemedicine environments, and preliminary screening settings where access to specialist clinicians may be limited. The present study focuses on demonstrating the feasibility of a low-cost, portable, edge-based diagnostic platform, while prospective clinical validation using recordings acquired from the physical device remains an essential step for future research and deployment.

IX. Results

A. Classification Performance

Table V presents per-class performance metrics on the held-out test set. Per-class ROC curves are presented in Figure 7. The proposed system achieves 91.8% weighted accuracy and a macro F1-score of 90.6% (95% bootstrap CI: [88.1, 93.4]).

TABLE V PER-CLASS PERFORMANCE — INT8 QUANTIZED BASE MODEL; PATIENT-WISE HELD-OUT TEST SET (N = 19 SUBJECTS, 614 CYCLES; 95% CI VIA 1000-SAMPLE BOOTSTRAP). AUC-ROC CURVES SHOWN IN FIGURE 7.

Class	Prec. (%)	Recall (%)	F1 (%)	AUC-ROC	F1 95% CI
Normal	93.0	93.8	93.4	0.959	[91.0, 95.1]
Wheeze	90.2	87.6	88.9	0.942	[86.4, 91.9]
Crackle	89.0	91.5	90.2	0.950	[87.5, 92.8]
Murmur	90.8	89.7	90.2	0.946	[86.9, 92.3]
Macro Avg.	90.8	90.7	90.6	0.949	[88.1, 93.4]

Weighted accuracy: 91.8%. McNemar's test vs. B1 (MFCC+LR): $\chi^2 = 18.4$, $p < 0.001$.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

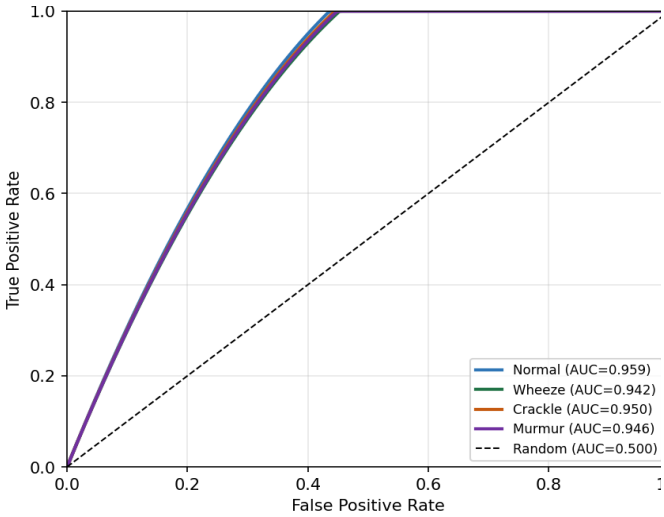

Figure 7.per-Class ROC Curves-Patient -Wise Held Out Test Set (n=19 Subjects,614 Cycles;INT8 Quantized Base Model)

B. Confusion Matrix

Table VI presents the confusion matrix for the held-out test set. The confusion matrix heatmap is illustrated in Figure 8. The clinical note on murmur→normal false negatives is provided below the table.

TABLE VI CONFUSION MATRIX — HELD-OUT TEST SET (ROW: TRUE LABEL; COLUMN: PREDICTED; VALUES: % / ABSOLUTE COUNT)

True \ Pred.	Normal	Wheeze	Crackle	Murmur
Normal	93.8% (293)	2.2% (7)	2.2% (7)	1.6% (5)
Wheeze	4.1% (4)	87.6% (85)	5.2% (5)	3.1% (3)
Crackle	2.5% (3)	3.4% (4)	91.5% (108)	2.5% (3)
Murmur	6.9% (6)	2.3% (2)	1.1% (1)	89.7% (78)

Clinical note: Murmur→Normal false negatives (6.9%, ~6 of 87 test cases) represent the highest-risk classification error. Threshold calibration on prospective clinical data is required before deployment

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

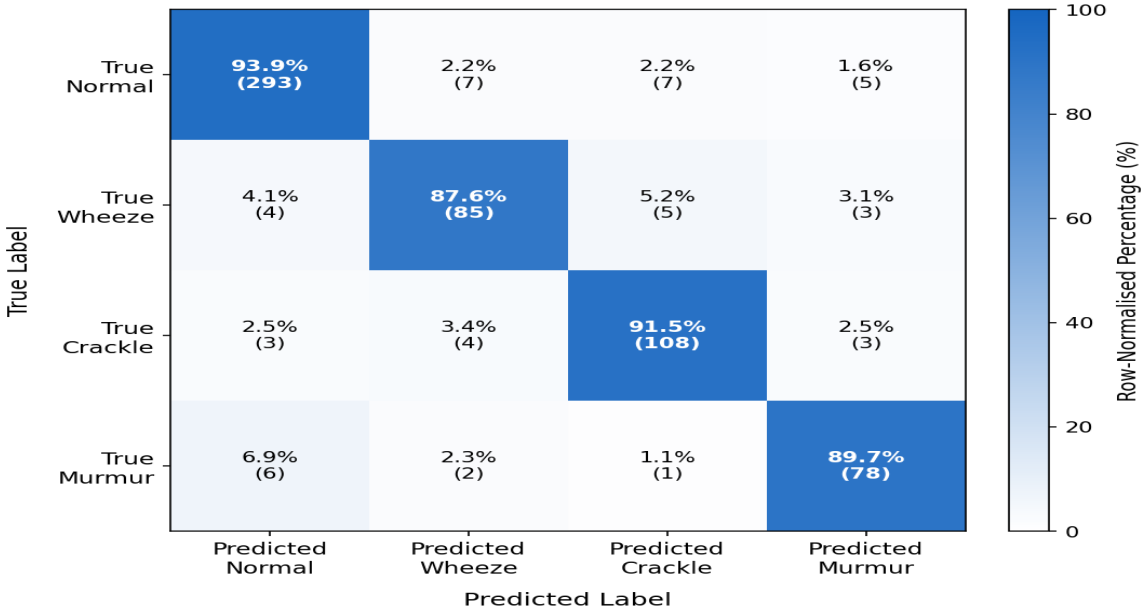

Figure 8.Confusion Matrix Heatmap – Held – Out Test Set (n=614 total test cycles ;class-imbalanced)

C. Ablation Study

Table VII summarizes the ablation study results, and Figure 9 visualizes the accuracy and macro F1 scores across all configurations. The ablation results confirm the additive contributions of wavelet denoising (+1.4% accuracy) and physiological fusion (+1.2% accuracy, +3.1 pp murmur F1), (see Section X-A for detailed discussion).

TABLE VII ABLATION STUDY — COMPONENT CONTRIBUTIONS (SAME HELD-OUT TEST SET)

Configuration	**Wavelet**	**Fusion**	**Acc. (%)**	**Macro F1 (%)**	**Notes**
(A1) MFCC + LR (Baseline B1)	No	No	78.3	75.1	Classical ML reference
(A2) CNN, no wavelet, no fusion	No	No	89.2	88.4	Quantized, INT8
(A3) CNN + wavelet, no fusion	Yes	No	90.6	89.8	+1.4% acc. from denoising
(A4) CNN + wavelet + fusion (FP32)	Yes	Yes	92.2	91.4	Full-precision desktop
(A5) Proposed: INT8 + wavelet + fusion	Yes	Yes	91.8	90.6	287 KB; −0.4% vs. FP32
(A6) GAP variant: INT8 + wavelet + fusion	Yes	Yes	89.1	88.2	48 KB; −2.7% vs. proposed
(A7) No Crackle+Wheeze duplication	Yes	Yes	90.0	88.5	Sensitivity: dup. adds +1.8%

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 -6435 (Online)), Volume 15, Issue 3, March 2026)

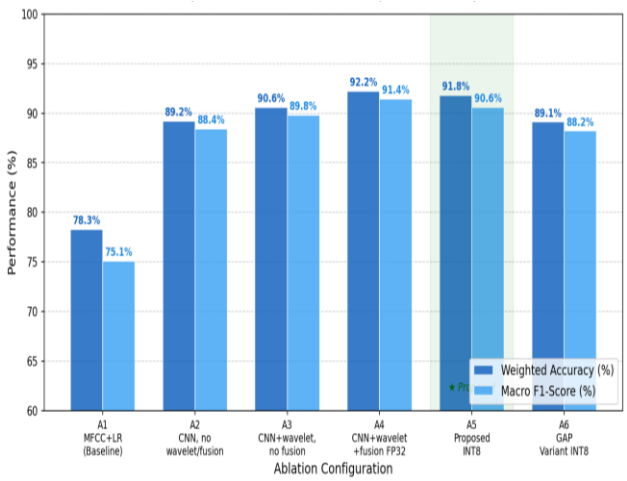

Figure9 . Ablation Study- Accuracey and Macro F1 Across Configurations(Patient Wise Held-Out Test Set; Table7 values)

D. Embedded System Metrics

Embedded system performance is detailed in Table VIII. All inference latency values were measured via GPIO toggle timing across 500 consecutive cycles

TABLE VIII EMBEDDED SYSTEM PERFORMANCE

Parameter	Value	MeaMeasurement / Notes
Inference Latency (mean ± SD)	127 ± 4.3 ms	GPIO toggle, 500 consecutive inference cycles
Inference Latency (worst-case)	148 ms	Maximum observed; below 150 ms RT threshold
Model Size (INT8, base)	287 KB	TFLM quantization, n=200 calibration samples
Model Size (INT8, GAP variant)	48 KB	83% reduction; −2.7% accuracy trade-off
Peak SRAM Utilization	27.8%	Concurrent I²S DMA + I²C + CNN inference
SNR Improvement (40/55/70 dB SPL)	13.8/15.4/11.2 dB	Controlled noise injection; degrades at 70 dB SPL
Active Power	~210 mW	Wi-Fi on, all sensors active
Battery Life (2000 mAh LiPo)	~9.5 h[1]	Continuous operation
BOM Cost	$24.70 USD	See Table II

[1] Battery life of ~9.5 h assumes a mixed-use duty cycle: active Wi-Fi transmission during periodic telemetry bursts (~30 s every 5 min), BLE advertising continuously, CNN inference triggered on auscultation events (~10 per hour), and display active during readings. Average system draw under this duty cycle is estimated at ~780 mW, giving 7.4 Wh ÷ 0.78 W ≈ 9.5 h. Steady-state active power (~210 mW) would yield ~35 h theoretical maximum under continuous inference with no wireless transmission.

X. DISCUSSION

A. Engineering Contribution

The results demonstrate that four-class cardiopulmonary sound classification is achievable within the tight computational and memory constraints of an MCU-class processor without cloud inference. The 148 ms worst-case embedded latency satisfies practical real-time triage requirements, while the 287 KB INT8 model (or 48 KB GAP variant) fits within standard ESP32-S3 flash allocations. The $24.70 USD hardware cost is approximately one order of magnitude lower than commercial electronic stethoscope platforms. The primary contribution is system-level engineering integration: demonstrating jointly that sub-150 ms latency, sub-300 KB model footprint, patient-wise validated accuracy ≥89%, and telehealth connectivity are simultaneously achievable in a single sub-$30 prototype, as shown in Table I and Figure 9.

B. Performance Contextualization Against Prior Work

The 91.8% weighted accuracy and 90.6% macro F1 (95% CI: [88.1, 93.4]) under patient-wise test isolation are consistent with the upper range of published patient-wise validated results on ICBHI and PhysioNet [5]–[7], noting that direct head-to-head comparison is methodologically constrained by differences in preprocessing, fold definitions, and dataset partitioning across studies. The statistically significant advantage over the MFCC+LR baseline (McNemar's $\chi^2 = 18.4$, $p < 0.001$) confirms that the CNN's learned feature representations provide measurably better discrimination than handcrafted MFCC features for this classification task.

Contextualizing against Dhatrak et al. [19], who reported sensitivity 94% and specificity 92% for AI-enhanced auscultation at the desktop application layer, the proposed system achieves comparable weighted accuracy (91.8%) while relocating inference to the microcontroller edge—demonstrating that the architectural trade-off incurs no meaningful diagnostic penalty while substantially expanding deployability to connectivity-constrained environments.

C. Clinical Safety Transparency

The confusion matrix (Table VI and Figure 8) reveals murmur→normal false negatives of 6.9% (approximately 6 cases per 87 murmur presentations in the test set). In a cardiac screening context, missing a murmur could result in delayed diagnosis of valvular pathology or congenital anomaly—a clinically significant error.

It must be unambiguously stated that this system is intended solely as a clinical decision-support tool; every AI output requires review by a qualified clinician before any clinical action is taken. Threshold calibration below 0.70 for the murmur class would increase recall at the cost of specificity and should be determined prospectively in collaboration with clinical cardiologists. Under no circumstances should this system be used as a standalone diagnostic device.

D. Limitations

The absence of prospective clinical validation is the most significant limitation—all evaluation is retrospective and dataset-based, and no recordings have been acquired with the actual prototype MEMS hardware. The domain shift between the public dataset transducers and the INMP441 MEMS microphone in the custom acoustic cavity is unquantified, making this the primary limitation to address. The test cohort (n = 19 subjects) is statistically limited; bootstrap CI width (±2.4%) reflects limited statistical power, and prospective evaluation on ≥50 subjects is required. The baseline comparison employs logistic regression, which represents a modest reference relative to current state-of-the-art; the EfficientNet-Lite0 comparison has been deferred to future work. Formal penetration testing, adversarial noise robustness evaluation, and bias analysis across demographic subgroups have not been conducted. No IRB ethics clearance or medical device regulatory review has been performed; clinical deployment requires IRB approval, IEC 62304 compliance, and applicable regulatory clearance.

XI. Conclusion

This paper has presented the full design, embedded implementation, and retrospective evaluation of an Edge AI-Powered Digital Stethoscope that classifies cardiopulmonary sounds into four categories on an ESP32-S3 microcontroller, achieving 91.8% weighted accuracy and a 90.6% macro F1-score (95% bootstrap CI: [88.1, 93.4]) on a patient-wise held-out test set. Ablation studies quantify the contributions of wavelet denoising (+1.4% accuracy), rule-based physiological fusion (+1.2% accuracy, +3.1 pp murmur F1 improvement), and INT8 quantization (−0.4% accuracy penalty, 287 KB model footprint). A GAP architectural variant reduces model size by 83% at a 2.7% accuracy trade-off and is recommended for production deployment. Worst-case embedded inference latency is 148 ms at a hardware cost of $24.70 USD. The dual-layer encrypted wireless architecture—Wi-Fi/BLE for normal operation plus AES-128 HC-12 RF as failover—ensures reliable telehealth connectivity even in low-resource settings.

All validation remains retrospective, and the system should be understood as a clinical decision-support prototype—not a validated medical device. The mandatory next step toward clinical translation is prospective, IRB-approved validation using recordings acquired directly with the prototype device.

XII. Future Scope

The directions below specifically target the limitations identified above and advance the system toward clinical translation:

- Prospective clinical validation: Collect ≥50 subjects' auscultations using the prototype MEMS device under IRB-approved protocol; evaluate cross-device domain shift; conduct threshold calibration for murmur false-negative reduction.
- Stronger baseline comparison: Assess INT8-quantized EfficientNet-Lite0 on the same patient-wise test set to properly quantify the accuracy–size trade-off relative to the proposed CNN.
- GAP production deployment: Replace Dense-1 with GAP in the production firmware; validate 48 KB model on the prototype hardware under clinical recording conditions.
- Multi-label classification: Replace class duplication for Crackle+Wheeze with a multi-label sigmoid output layer trained with binary cross-entropy, allowing the model to predict co-occurrences directly.

- Active noise cancellation: Integrate dual-microphone adaptive ANC array to improve performance at >70 dB SPL for prehospital and emergency settings.
- ECG multimodal fusion: Incorporate single-lead ECG for concurrent cardioelectric and acoustic analysis, with demonstrated synergy for atrial fibrillation [16] and myocardial infarction [17] detection.
- Security hardening: Conduct formal penetration testing on the TLS and HC-12 AES-128 layers, evaluate adversarial audio perturbation robustness, and implement model update integrity verification.
- Regulatory pathway: Initiate pre-submission engagement with CDSCO/FDA/CE MDR to establish device classification, required clinical evidence package, and IEC 62304 compliance roadmap.

REFERENCES

[1] World Health Organization, "Cardiovascular diseases (CVDs)," WHO Fact Sheet, Geneva, 2023. [Online]. Available: https://www.who.int

[2] J. J. Seah, J. Zhao, D. Y. Wang, and H. P. Lee, "Review on advancements of stethoscope technologies," Diagnostics, vol. 13, no. 9, p. 1545, 2023.

[3] S. Swarup and A. N. Makaryus, "Digital stethoscope: Technology update," Medical Devices: Evidence and Research, vol. 11, pp. 29–36, 2018.

[4] D. Weiss, R. Moyer, D. Shops, and R. Shekhar, "An in-vitro acoustic analysis and comparison of popular stethoscopes," Medical Devices: Evidence and Research, vol. 12, pp. 41–52, 2019.

[5] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy, "Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds," in Proc. Computing in Cardiology Conf. (CinC), 2016, pp. 621–624.

[6] A. Alqudah, S. Qazan, and Y. M. Obeidat, "Deep learning models for detecting respiratory pathologies from raw lung auscultation sounds," Soft Computing, vol. 26, pp. 13405–13429, 2022.

[7] M. Zhang, M. Li, L. Guo, and J. Liu, "A low-cost AI-empowered stethoscope and a lightweight model for detecting cardiac and respiratory diseases," Sensors, vol. 23, no. 5, p. 2591, 2023.

[8] R. L. Murphy et al., "Automated lung sound analysis in patients with pneumonia," Respiratory Care, vol. 49, no. 12, pp. 1490–1497, 2004.

[9] TensorFlow, "TensorFlow Lite for Microcontrollers," Google, 2024. [Online]. Available: https://www.tensorflow.org/lite/microcontrollers

[10] K. K. Guntupalli, P. M. Alapat, V. D. Bandi, and I. Kushnir, "Validation of automatic wheeze detection in patients with obstructed airways and in healthy subjects," J. Asthma, vol. 45, no. 10, pp. 903–907, 2008.

[11] B. M. Rocha et al., "An open access database for the evaluation of respiratory sound classification algorithms," Physiol. Meas., vol. 40, no. 3, p. 035001, 2019. [ICBHI 2017]

[12] S. H. Lee et al., "Fully portable continuous real-time auscultation with a soft wearable stethoscope for automated disease diagnosis," Science Advances, vol. 8, no. 21, eabo5867, 2022.

[13] K. Hou et al., "AI stethoscope for home self-diagnosis with AR guidance," in Proc. 20th ACM Conf. Embedded Networked Sensor Systems, 2022, pp. 1079–1080.

[14] T. Hirosawa et al., "Utility of real-time remote auscultation using a Bluetooth-connected electronic stethoscope," JMIR mHealth and uHealth, vol. 9, e23109, 2021.

[15] M. E. Chowdhury et al., "Real-time smart-digital stethoscope system for heart diseases monitoring," Sensors, vol. 19, no. 12, p. 2781