

CARE-Lung: Calibrated, Cross-Validated Cycle-to-Patient Aggregation with Conformal Referral for Respiratory Sound Screening

Arnob Aich Anurag¹, Shamiul Islam¹, Nuren Tasnim Risha¹, Md.Šadman Hossain¹, Trisha Sarker¹, Sifat Rahman Ahona¹, and Juena Ahmed Noshin¹

¹Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka, Bangladesh
23-51390-1, 23-51353-1, 22-48474-3, 23-55242-3, 23-51420-1@student.aiub.edu
ahona,juena@aiub.edu
Corresponding author

Abstract. Screening systems for respiratory sounds are typically tested on a small held-out split for a single analysis, so model rankings carry a variance that is rarely quantified and are rarely checked against any baseline stronger than chance. To address this we developed CARE-Lung, a patient-level screening pipeline for the ICBHI 2017 respiratory sound dataset. The system works in four steps: it calibrates a classifier that flags abnormal sounds in short breathing cycles, summarises each patient’s cycles into a compact statistical profile, estimates from that profile how likely the patient is to be abnormal, and adds a safety net that can flag uncertain cases as refer rather than forcing a binary call. Instead of testing on 26 patients, we evaluate all 126 through patient-level cross-validation and report patient-grouped bootstrap intervals throughout. CARE-Lung reaches AUROC 0.832 (95% CI 0.758–0.897), PR-AUC 0.958 and sensitivity 0.720, with calibration error dropping more than threefold against a count-ratio baseline. Three controls then bound that result. Predicting every patient abnormal already gives 0.794 accuracy and 0.885 F1, above CARE-Lung’s 0.730 and 0.809, so those are the wrong headline metrics. A baseline using the recording device alone and no audio reaches AUROC 0.784, and the paired difference against CARE-Lung is +0.047 with a 95% interval of $[-0.047, +0.141]$, so the acoustic pipeline cannot be shown to beat knowing which stethoscope was used. With device held constant, AUROC falls to 0.707. What survives is the methodology rather than the headline: within a single device, distributional aggregation still beats hard counting, calibration still improves threefold, and referral still cuts selective risk from 0.375 to 0.100. The main remaining source of error is the microphone used to collect the data rather than the patients themselves.

Keywords: Respiratory sound analysis · COPD screening · Device confounding · Calibration · Cross-validation · Conformal prediction · Uncertainty quantification

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

Cite as: Anurag, A.A., et al.: CARE-Lung: Calibrated, Cross-Validated Cycle-to-Patient Aggregation with Conformal Referral for Respiratory Sound Screening. In: Mahmud, M., et al. (eds.) Applied Intelligence and Informatics. AII 2026. CCIS, vol. 3206, pp. 165–179. Springer, Cham (2026).

1 Introduction

Respiratory health disorders represent a significant public health dilemma worldwide, and early identification is critical because delayed diagnosis increases patient risk. Auscultation interpretation relies on clinician judgement as well as on listening-environment noise, so automated machine learning screening of respiratory sounds has emerged as a viable research direction, with the open-access ICBHI database among its primary resources [17]. Numerous studies demonstrate the ability of deep learning models to detect abnormal lung sounds with high confidence [8]. However, many classify sound at the cycle level rather than the patient level, which fails to incorporate patient-specific context, and many patients have cycles that resist definitive classification. There is therefore a need to integrate cycle-level predictions into a single patient-level result, and that result must be reliable, because high-confidence errors carry medical risk. Prior studies have emphasised robustness, valid evaluation, noise handling, and clinical applicability [14,18]. This study proposes CARE-Lung, whose core components are respiratory cycle classification, aggregation of outputs into patient-level decisions, and confidence calibration to identify uncertain cases. Unlike traditional methods designed to generate predictions for all patients, CARE-Lung is designed specifically for safe screening. We argue that accuracy at the cycle level is not enough, that a model must recognise its own limitations under ambiguity, and that it must report how uncertain its own reported numbers are, which on a 126-patient cohort is a first-order concern rather than a footnote. The contributions of this study are as follows:

1. **Calibrated pipeline.** A patient-independent pipeline over all 6,898 annotated breathing cycles in ICBHI, combining a confidence-calibrated cycle classifier, a statistical patient summary and a referral layer that flags uncertain cases rather than forcing a hard decision. Every within-fold subset, threshold and per-patient prediction is released for audit.
2. **Patient-level cross-validation with grouped uncertainty.** All 126 patients are evaluated through stratified 5-fold cross-validation that pools out-of-fold predictions, and every metric carries a patient-grouped bootstrap interval. Paired tests show that CARE-Lung’s margin is decisive only over the non-learned baselines.
3. **Aggregator ablation.** Compared against pre-specified patient summaries fixed in advance, a three-dimensional summary matches the 14-dimensional descriptor, isolating what the aggregation contributes.
4. **Calibration and referral study.** CARE-Lung reduces expected calibration error 3.4 times relative to the count-ratio baseline, and class-conditional conformal analysis shows that marginal referral misallocates its error budget on an imbalanced cohort.
5. **Device-confound audit.** A device-only baseline, a paired test against it, and device-controlled evaluation show that the headline AUROC cannot be distinguished from what stethoscope identity alone provides, while the aggregation, calibration and referral gains persist with device held fixed.

The remainder of the paper is organised as follows. Section 2 surveys related work. Section 3 describes the CARE-Lung methodology. Section 4 reports experimental results. Finally, Section 5 concludes the paper.

2 Related Work

The ICBHI respiratory-sound database is widely used to benchmark automated auscultation [17]. Deep learning has shown strong cycle-level performance [8], while Mahmud and collaborators have also demonstrated automatic lung-health screening directly from respiratory sounds [13]. Related physiological-sensing work shows that learned representations can extract clinically relevant patterns from noisy, repeated measurements [5], motivating aggregation beyond isolated cycles.

Robust validation remains central to clinical translation. Co-tuning and stochastic normalisation improve lung-sound classification under dataset and device variation [14], while reviews identify small cohorts, uncertainty, interpretability and privacy as unresolved deployment issues [18]. Mahmud-related clinical studies further demonstrate explainable critical-care prediction [3], prognostic modelling under severe patient risk [15], and preventive healthcare built around privacy, explainability and inclusion [12]. These works support evaluation designs that treat confidence and clinical context as part of the prediction problem rather than as post-hoc additions.

A complementary line concerns multimodal and patient-aware modelling. Explainable multimodal learning has been used to integrate heterogeneous behavioural and physiological evidence [16], while explainable digital-human modelling provides a patient-centred mechanism for combining longitudinal critical-care observations [2]. Broader XAI work emphasises that high-performing black-box predictions require interpretable support before use in high-stakes settings [6]. These ideas motivate CARE-Lung’s distributional patient summary and explicit calibration of cycle probabilities before aggregation.

Trustworthy deployment also requires attention to infrastructure, privacy and human oversight. Distributed healthcare architectures can support scalable patient monitoring [4]; explainable and privacy-preserving personalisation has been proposed for clinically sensitive AI [11]; and inclusive healthcare frameworks emphasise accessibility and accountable decision support [1]. Recent work extends these principles to privacy-preserving multi-agent clinical AI [10] and explainable multimodal digital-human models with uncertainty-aware adaptation [9]. Edge-oriented hierarchical profiling aggregates per-cycle predictions into a count-ratio patient score [7], which we retain as the lightweight baseline. Taken together, prior work motivates calibrated patient-level aggregation, grouped validation and referral rather than forced binary decisions.

Table 1. Dataset profile and evaluation consequences.

Property	Evidence	Consequence
Small cohort	126 subjects	Out-of-fold evaluation; grouped bootstrap intervals
Many cycles/patient	6,898 cycles, 920 recordings	Cycles treated as repeated noisy observations
Cycle-label balance	3,642 vs. 3,256 abnormal	Low cycle AUROC; aggregation required
Patient imbalance	100 abnormal vs. 26 healthy	Trivial all-abnormal baseline must be reported
Disease heterogeneity	COPD, URTI, pneumonia, etc.	Burden screening, not disease-specific diagnosis
Device imbalance	Meditron only mixed-class device	Device-stratified analysis and confound audit

3 Methodology

3.1 Dataset and Task

We employed the ICBHI 2017 respiratory sound database: 920 recordings from 126 subjects and 6,898 annotated respiratory cycles (3,642 normal; 3,256 with crackles, wheezes, or both). Diagnoses were binarised as healthy vs. abnormal (100 abnormal, 26 healthy), and the task is framed as abnormal respiratory-burden screening rather than COPD diagnosis. Table 1 links dataset structure to evaluation design. Two protocols are used: a *fixed patient-independent split* (seed 42: 60 train, 20 calibration, 20 conformal-calibration, 26 test) for cycle-level calibration analyses, and *patient-level stratified 5-fold cross-validation* (Section 3.4) for headline comparisons.

3.2 Cycle Features and Calibration

For each annotated respiratory cycle, 31 features were extracted: duration, RMS, mean absolute amplitude, standard deviation, zero-crossing rate, spectral centroid, bandwidth, rolloff, flatness, spectral entropy, dominant frequency, seven band-energy ratios, and 13 MFCC-like cepstral coefficients from a mel filterbank. No device identifier was used as a model input. A class-balanced random forest was trained on these features; isotonic regression was then fitted on a disjoint calibration cohort to map raw forest scores to calibrated probabilities $\hat{p}_{\text{cyc}}$. Expected calibration error is

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|. \quad (1)$$

On the fixed split, isotonic calibration reduced cycle-level ECE from 0.198 to 0.147 (Fig. 1). Cycle-level discrimination remained limited ($\text{AUROC} \approx 0.53$), indicating that patient-level performance depends on aggregating multiple calibrated posteriors: an individual 2–3s cycle from an abnormal patient frequently contains no audible adventitious event, so the cycle label is a noisy proxy for

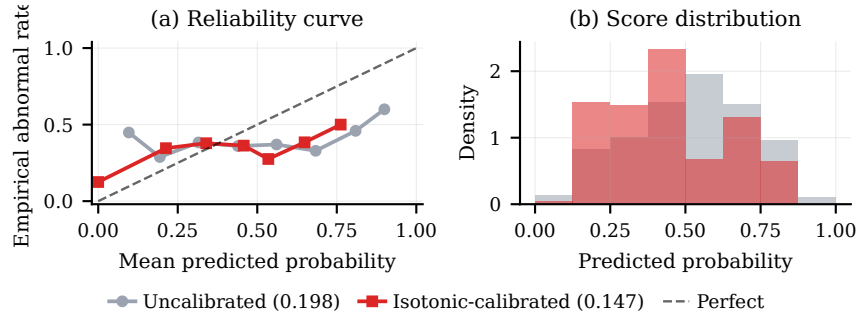


Fig. 1. Cycle-level reliability before and after isotonic calibration.

the patient label. Figure 2 summarises the full pipeline: cycle-level feature extraction and isotonic calibration, distributional cycle-to-patient aggregation, and split-conformal decisioning with an INCONCLUSIVE-REFER output.

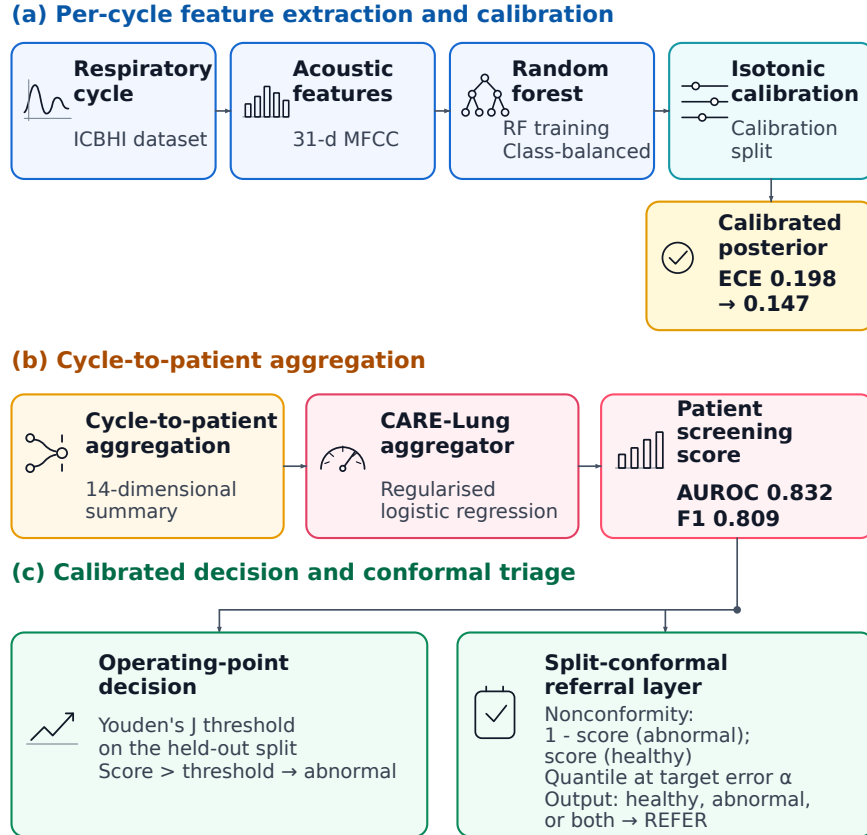


Fig. 2. CARE-Lung end-to-end pipeline.

3.3 Cycle-to-Patient Aggregation

We chose to aggregate because both the rate of respiratory events and the uncertainty of cycle classification need to be captured, rather than just averaging them. Calibrated cycle probabilities across all of a patient’s recordings are summarised into a 14-dimensional descriptor: hard count ratio, mean probability, entropy-weighted abnormal fraction, standard deviation, six percentiles (min, 25th, median, 75th, 90th, max), mean predictive entropy, count of high-confidence abnormal cycles, log cycle count, and device count. The descriptor is fixed in advance and no feature selection is performed at any point. Five aggregation strategies are compared:

- **Count-ratio.** Hard fraction of cycles classified abnormal, the lightweight edge-oriented baseline [7].
- **Confidence-weighted fraction.** Entropy-weighted mean probability, discounting low-confidence cycles.
- **CARE-Lung (proposed).** Class-balanced ℓ_2 -regularised logistic regression over the full 14-dimensional descriptor.
- **Gradient-boosted.** Shallow gradient-boosting over the same descriptor with sigmoid recalibration.
- **SVM + spectral summaries.** RBF-kernel SVM over the same descriptor.

CARE-Lung++ averages CARE-Lung and SVM scores and recalibrates through a 1-D sigmoid fitted on the same cohort as the base aggregators (Section 4.3). Binary thresholds are chosen by maximising Youden’s statistic on a held-out threshold-selection cohort, never on the evaluation cohort.

3.4 Patient-Level Stratified 5-Fold Cross-Validation

A single 26-patient test split (only 5 healthy patients) yields high-variance estimates: moving one healthy patient shifts specificity by 20 percentage points. We therefore run patient-level **stratified 5-fold cross-validation** (seed 42) over all 126 patients, pooling out-of-fold predictions so every patient contributes exactly one prediction from a model that excluded that patient from all training, calibration, aggregation, and threshold-selection steps. Within each outer fold the ~ 101 remaining patients are split into three disjoint subsets, each with one role: RF-train (60%) fits the cycle classifier, calibration (20%) fits the isotonic map and the sigmoid recalibrator, and threshold-selection (20%) fixes the operating point. Patient-level aggregators are fitted on RF-train $\cup$ calibration. Table 2 gives the exact composition of every subset in every fold. The conformal layer is calibrated on a further disjoint cohort, so no cohort is reused across roles.

Because the patient is the unit of analysis, all interval estimates use a patient-grouped bootstrap: we resample the 126 patients with replacement, and a drawn patient carries all of its cycles, its descriptor and its single pooled prediction together, so no patient is split across a resample. We draw 4,000 replicates and report 2.5th/97.5th percentiles. Model comparisons use the paired form, evaluating ΔAUROC on identical resamples so that shared cohort variance cancels.

Table 2. Within-fold subset composition: patients (abnormal/healthy) and cycles.

Fold	RF-train	Calibration	Threshold	Test (scored)
0	60 (48/12), 3,669	20 (16/4), 1,129	20 (16/4), 803	26 (20/6), 1,297
1	60 (48/12), 3,380	20 (16/4), 802	21 (16/5), 933	25 (20/5), 1,783
2	60 (48/12), 3,108	20 (16/4), 961	21 (16/5), 1,150	25 (20/5), 1,679
3	60 (48/12), 3,686	20 (16/4), 1,068	21 (16/5), 1,158	25 (20/5), 986
4	60 (48/12), 3,868	20 (16/4), 1,173	21 (16/5), 704	25 (20/5), 1,153

Per-fold means and standard deviations are reported alongside, since the bootstrap interval describes cohort sampling variability while the fold spread describes sensitivity to the training partition. We release the split manifests, per-fold thresholds, per-patient out-of-fold scores and predictions, the cycle feature cache, and all analysis scripts with fixed seeds, so that every reported number is traceable to a released file.

3.5 Conformal Referral

For patient score p , binary split-conformal prediction uses nonconformity score $1-p$ for abnormal patients and p for healthy patients. At target error level α , conformal quantile $\hat{q}$ is computed on a disjoint conformal-calibration cohort. A test patient receives: *screen-negative* if $p \leq \hat{q}$; *screen-positive* if $1-p \leq \hat{q}$; or *INCONCLUSIVE-REFER* if $1-\hat{q} \leq p \leq \hat{q}$ (a valid band when $\hat{q} \geq 0.5$).

This marginal guarantee holds, on average, across both classes, which, in a 100:26 cohort, can be met while systematically under-covering the minority behaviour of interest. We therefore also evaluate a Mondrian (class-conditional) variant computing separate quantiles from abnormal and healthy calibration patients, targeting $1 - \alpha$ coverage within each class. Both are assessed by selective risk, the error rate among non-referred patients, together with the referral rate. To evaluate over all 126 patients rather than 26, we use a cross-conformal procedure: 400 replicates, each splitting the pooled out-of-fold scores into a class-stratified calibration half and an evaluation half, with patients kept intact. In the intended pathway, CARE-Lung acts as a pre-clinical triage aid where auscultation is already performed but specialist review is scarce. Screen-negative patients continue routine care; screen-positive patients are queued for clinician review and, where indicated, spirometry; and *INCONCLUSIVE-REFER* patients enter the same queue flagged as model-uncertain rather than model-positive, so reviewer attention follows uncertainty rather than score alone. Where a missed abnormal patient is the dominant cost, the sensitivity-oriented CARE-Lung point applies; where reviewer capacity binds, CARE-Lung++ is preferable.

4 Results

4.1 Cross-Validated Patient-Level Comparison

Tables 3 and 4 report pooled out-of-fold metrics for all 126 patients. CARE-Lung leads on accuracy, sensitivity, F1, AUROC, and PR-AUC. SVM summaries

achieve higher precision and lower ECE; CARE-Lung++ achieves higher specificity, both at lower sensitivity. Figure 3 shows pooled ROC and PR curves. CARE-Lung has AUROC 0.832 and PR-AUC 0.958, with AUROC margins of 0.034–0.175 over non-ensemble baselines.

Table 3. Cross-validated patient-level screening metrics.

Model	Acc.	Prec.	Sens.	F1
<i>All-abnormal baseline</i>	0.794	0.794	1.000	0.885
Count-ratio	0.540	0.904	0.470	0.618
Conf.-weighted	0.643	0.899	0.620	0.734
CARE-Lung	0.730	0.923	0.720	0.809
Grad.-boosted	0.698	0.908	0.690	0.784
SVM summaries	0.690	0.930	0.660	0.772
CARE-Lung++	0.722	0.958	0.680	0.795

Table 4. Cross-validated ranking and calibration metrics.

Model	Spec.	AUROC	PR-AUC	ECE
<i>All-abnormal baseline</i>	0.000	0.500	0.794	n/a
<i>Device identity only</i>	n/a	0.784	0.939	n/a
Count-ratio	0.808	0.694	0.898	0.565
Conf.-weighted	0.731	0.657	0.892	0.384
CARE-Lung	0.769	0.832	0.958	0.165
Grad.-boosted	0.731	0.779	0.925	0.169
SVM summaries	0.808	0.798	0.947	0.075
CARE-Lung++	0.885	0.821	0.955	0.161

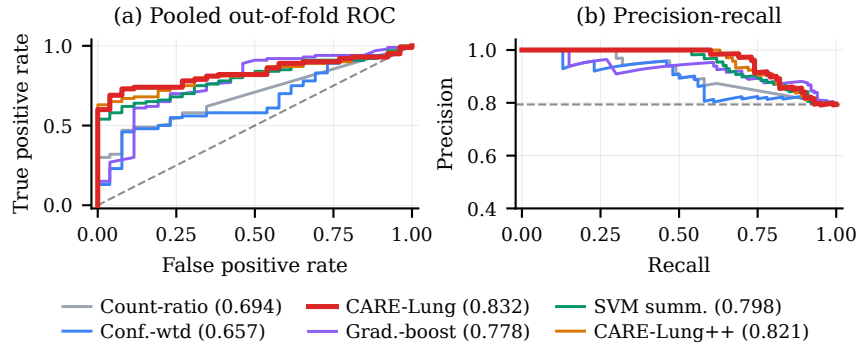


Fig. 3. Pooled out-of-fold ROC and precision–recall curves.

The baseline rows bound these numbers from the other side. Because the cohort is 79% abnormal, predicting every patient abnormal already yields 0.794 accuracy and 0.885 F1, above CARE-Lung’s 0.730 and 0.809, so accuracy and F1 are the wrong headline metrics on this cohort and we do not claim them as gains. What the model buys over that baseline is discrimination (AUROC 0.500 $\rightarrow$ 0.832), ranking under imbalance (PR-AUC 0.794 $\rightarrow$ 0.958) and the ability to identify healthy patients at all (specificity 0.000 $\rightarrow$ 0.769). The device-identity baseline is more demanding still, and Section 4.5 takes it up. Table 5 reports per-fold stability with grouped bootstrap intervals and paired AUROC differences. CARE-Lung achieves the lowest standard deviations in F1 (0.027)

and sensitivity (0.040) among learned aggregators. The intervals also discipline the ranking: its advantage is decisive over the two non-learned baselines, where the paired interval excludes zero, but the intervals against gradient boosting, SVM summaries and CARE-Lung++ all contain zero. Those four models are statistically indistinguishable in ranking quality here, so the honest claim is that distributional aggregation as a family beats hard counting.

Table 5. Fold-to-fold stability, bootstrap intervals, and paired Δ AUROC against CARE-Lung.

Model	F1 mean $\pm$ sd	AUROC mean $\pm$ sd	AUROC CI	Δ AUROC [CI]
Count-ratio	0.605 $\pm$ 0.141	0.698 $\pm$ 0.118	0.60–0.78	+0.137 [.06, .22]
Conf.-weighted	0.732 $\pm$ 0.044	0.708 $\pm$ 0.093	0.55–0.76	+0.174 [.07, .28]
CARE-Lung	0.810 $\pm$ 0.027	0.843 $\pm$ 0.091	0.76–0.90	n/a
Grad.-boosted	0.767 $\pm$ 0.130	0.801 $\pm$ 0.109	0.67–0.87	+0.053 [−.03, .15]
SVM summaries	0.770 $\pm$ 0.054	0.804 $\pm$ 0.074	0.72–0.87	+0.033 [−.01, .08]
CARE-Lung++	0.796 $\pm$ 0.030	0.829 $\pm$ 0.082	0.75–0.89	+0.011 [−.01, .03]

The fixed 26-patient split gives CARE-Lung an AUROC of 0.914 while the pooled estimate is 0.832, and the gap is a variance effect rather than a contradiction. AUROC is estimated from discordant pairs: the fixed split contributes $21 \times 5 = 105$, the pooled evaluation $100 \times 26 = 2,600$. Bootstrapping the fixed split at the patient level gives a 95% interval of 0.727–1.000 against 0.758–0.897 for the pooled estimate, so the two are consistent and the fixed-split number is simply three times less precise. The same mechanism explains the per-fold spread: with five healthy patients per test fold, specificity moves in steps of 0.20 and ranges from 0.40 to 1.00 (0.760 ± 0.294), while F1 is stable at 0.809 ± 0.027 . To test whether the descriptor’s size drives performance, we compare it against pre-specified patient summaries of increasing dimensionality (Table 6). Each summary is fixed before any model is fitted, no feature selection is performed, and every variant uses identical folds, fitting cohorts, logistic regression and threshold procedure. The 14-dimensional descriptor clearly beats a one-dimensional mean probability, confirming that a plain average of calibrated cycle scores is insufficient, but it does not beat a three-dimensional summary of mean, standard deviation and log cycle count, which reaches AUROC 0.843 with a paired difference of -0.011 (CI -0.052 to $+0.029$). The informative content is dispersion plus evidence volume rather than the specific 14 coordinates; we retain the full descriptor as the reference configuration but do not claim its dimensionality is responsible for the reported performance.

Table 6. Pre-specified aggregator ablation under identical folds.

Patient summary	Dim.	Sens.	F1	AUROC [CI]	Δ AUROC [CI]
Mean probability	1	0.630	0.741	0.727 [0.62, 0.83]	+0.104 [.01, .21]
Mean, SD	2	0.550	0.679	0.748 [0.64, 0.85]	+0.084 [−.01, .18]
Mean, SD, log cycles	3	0.760	0.840	0.843 [0.77, 0.91]	−0.011 [−.05, .03]
Mean, SD, median, max	4	0.590	0.715	0.768 [0.67, 0.86]	+0.063 [−.02, .14]
Quantile summary	6	0.700	0.800	0.825 [0.75, 0.89]	+0.007 [−.01, .03]
Full descriptor (CARE-Lung)	14	0.720	0.809	0.832 [0.76, 0.90]	n/a

4.2 Calibration Quality

healthy	20 (0.77)	6 (0.23)
abnormal	28 (0.28)	72 (0.72)
	pred. healthy	pred. abnormal

Fig. 4. CARE-Lung pooled out-of-fold confusion matrix ($n = 126$).

Figure 4 shows the pooled CARE-Lung confusion matrix: 28 of 100 abnormal patients are misclassified as healthy, identifying false negatives as the principal residual error. Figure 5 ranks all six aggregators by pooled out-of-fold ECE. The count-ratio baseline is poorly calibrated (ECE 0.565) because a hard cycle fraction is not a calibrated posterior. Confidence-weighting reduces ECE to 0.384 but remains inferior to learned distributional aggregation. CARE-Lung reduces ECE to 0.165 (95% CI 0.115–0.233) by fitting a regularised map over calibrated-cycle distribution statistics, a $3.4\times$ improvement. SVM summaries achieve the lowest ECE (0.075) but at lower sensitivity (0.660 vs. 0.720).

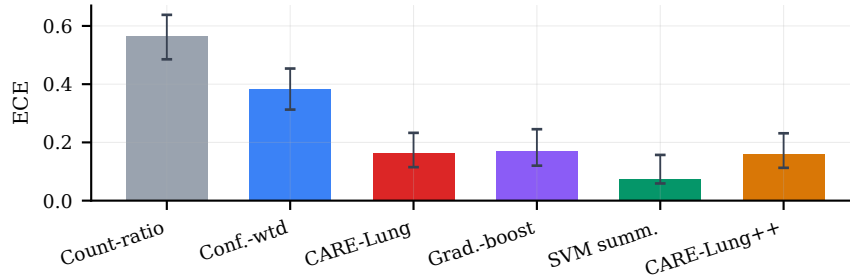


Fig. 5. Pooled out-of-fold ECE by aggregator.

4.3 CARE-Lung++: Higher-Specificity Operating Point

CARE-Lung++ averages the CARE-Lung and SVM scores, one linear over distributional descriptors and one an RBF-kernel margin, and recalibrates via a 1-D sigmoid. The trade-off is that specificity increases from 0.769 to 0.885 and precision from 0.923 to 0.958, while sensitivity decreases from 0.720 to 0.680 and AUROC from 0.832 to 0.821. Figure 6 visualises this directly. The specificity intervals overlap substantially and the paired AUROC difference is $+0.011$, so on 26 healthy patients this is a second operating point on the same calibrated score family rather than a better model, obtained without retraining the cycle classifier.

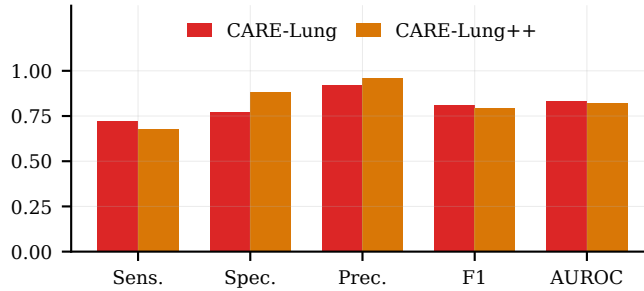


Fig. 6. CARE-Lung vs. CARE-Lung++, pooled out-of-fold ($n = 126$).

4.4 Conformal Triage

Table 7. Cross-conformal referral over all 126 patients.

α	$\hat{q}$	Cov. abnormal	Cov. healthy	Refer rate	Selective risk
<i>Marginal split-conformal</i>					
0.05	0.830	0.941	1.000	0.557	0.097
0.10	0.749	0.884	0.998	0.409	0.149
0.20	0.597	0.783	0.933	0.158	0.214
0.30	0.452	0.698	0.725	0.003	0.227
<i>Mondrian (class-conditional)</i>					
0.05	n/a	0.958	0.923	0.412	0.078
0.10	n/a	0.902	0.923	0.327	0.132
0.20	n/a	0.806	0.848	0.115	0.204
0.30	n/a	0.713	0.700	0.003	0.225

Table 7 and Fig. 7 report the referral trade-off over all 126 patients. Marginal coverage tracks its target closely (0.907 at $\alpha=0.10$ against 0.90), but the class-conditional breakdown shows the average conceals an asymmetry: healthy patients are covered at 0.998 and abnormal ones at 0.884, below target. Marginal conformal spends its error budget on the majority class and refers healthy patients at nearly twice the rate of abnormal ones, the wrong direction for a screening tool whose costly error is a missed abnormal patient. The Mondrian variant corrects this, giving 0.902 abnormal and 0.923 healthy coverage while reducing referral from 0.409 to 0.327 and selective risk from 0.149 to 0.132.

Referral also has to earn its cost (Fig. 8). Under forced choice the pooled error rate is 0.254; deferring the least confident patients reduces it monotonically to 0.184 at 69% coverage, 0.105 at 45% and 0.044 at 36%, with the missed-abnormal rate falling in step from 0.260 to 0.045. The patients the layer declines to decide are disproportionately the ones it would have got wrong, though deployment-scale coverage needs a larger calibration cohort.

4.5 Device Confounding

We stratify CARE-Lung by stethoscope over all 126 patients in Fig. 9. The AKGC417L (32) and Litt3200 (9) cohorts hold only abnormal patients and are detected at sensitivity 1.00; LittC2SE (21, also all abnormal) reaches 0.62; Meditron is the only mixed cohort (64: 38 abnormal, 26 healthy), with sensitivity

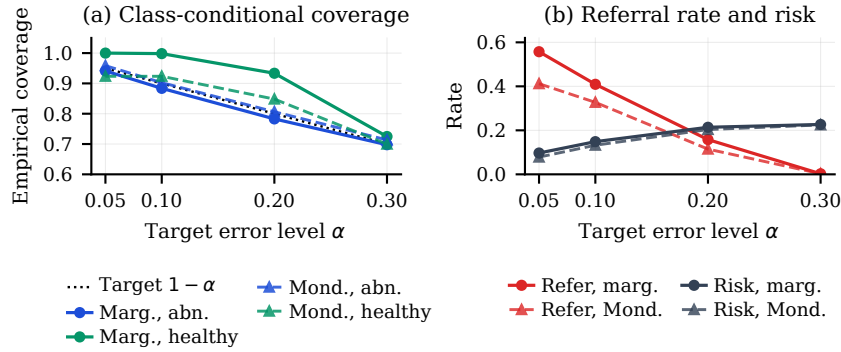


Fig. 7. Conformal coverage and referral rate vs. target error level.

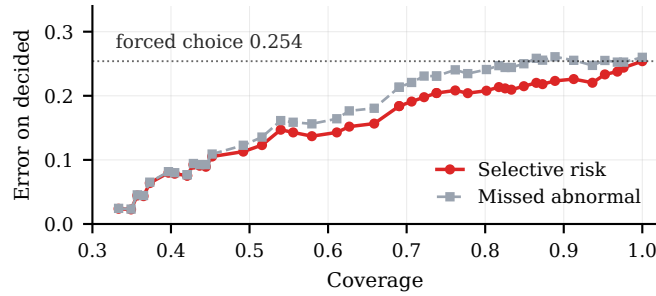


Fig. 8. Selective risk and missed-abnormal rate vs. coverage.

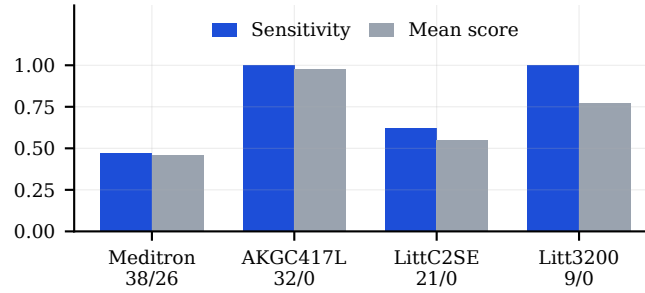


Fig. 9. Device-stratified CARE-Lung performance, pooled out-of-fold.

0.474 and specificity 0.769. Because all 26 healthy patients are recorded with Meditron, device shift and class shift are entangled in ICBHI. A device-leakage probe using cycle features alone reaches 0.811 accuracy against a 0.787 majority-device baseline, confirming that device signatures are encoded in the feature space even though no device identifier is an input. Reporting that errors differ by device does not say how much of the patient-level discrimination is attributable to hardware, so we add three controls. First, a device-only baseline scoring each patient by the abnormal rate of their stethoscope, fitted within each fold and

applied out-of-fold, uses no audio and still reaches AUROC 0.784 (95% CI 0.703–0.854). Second, a paired patient-grouped bootstrap of the AUROC difference against that baseline gives +0.047 with a 95% interval of $[-0.047, +0.141]$; the interval contains zero, so we cannot demonstrate that the acoustic model discriminates better than knowing which stethoscope was used. Third, restricting to Meditron holds the device constant, and CARE-Lung then reaches AUROC 0.707 (95% CI 0.576–0.830) against a prevalence of 0.594: genuine acoustic signal exists, but it is far weaker than the pooled figure suggests. The device-controlled setting is also where the methodological claims are properly tested, and two survive. Learned distributional aggregation still beats hard counting within a single device (0.707 vs. 0.579), and the calibration gain is undiminished: within Meditron, ECE falls from 0.478 to 0.157, a $3.0\times$ improvement matching the $3.4\times$ on the full cohort. Referral continues to work with device fixed, cutting selective risk from 0.375 under forced choice to 0.233 at 47% coverage and 0.100 at 16%. What does not survive is the headline discrimination number. Because the score tracks device, so does the abstention layer: at a fixed confidence band the referral rate is 0.000 for AKGC417L and 0.453 for Meditron, so a referral layer inherits whatever confounding the score contains and cannot be assumed safe merely because it abstains. The serialised model bundle is 9.86,MB with mean desktop inference time of 0.026,ms per cycle.

5 Conclusion

CARE-Lung brings together three ideas: isotonic-calibrated cycle classification, regularised distributional aggregation, and split-conformal referral into a unified pipeline for patient-level respiratory sound screening on ICBHI 2017. Individual breath cycles are hard to classify in isolation, with per-cycle AUROC around 0.53, but once calibrated score distributions are aggregated across a recording, patient-level AUROC reaches 0.832 (95% CI 0.758–0.897) and PR-AUC 0.958, alongside a $3.4\times$ reduction in calibration error over a count-ratio baseline. Three controls then place these numbers: accuracy and F1 fall below a trivial all-abnormal baseline and are not claimed as gains; a device-only baseline with no audio reaches AUROC 0.784 and the paired difference against CARE-Lung spans zero; and with device held constant the estimate falls to 0.707. What survives is the machinery rather than the headline, since within a single device the aggregation, calibration and referral gains all persist. The dataset sets clear limits on what we can claim. ICBHI 2017 covers only 126 patients with 26 healthy, so every specificity-type metric rests on that minority and the bootstrap intervals are correspondingly wide; device is confounded with class in a way no cross-validation scheme over this cohort can remove; and the conformal layer is calibrated on a few dozen patients, too few for the coverage guarantees a deployment would need. These results are best read as evidence for a calibrated, cross-validated screening methodology rather than a system ready for clinical deployment. The next steps follow from the confound rather than the results: device-balanced multi-site cohorts and external validation, device-invariant rep-

resentations, conformal calibration at a scale supporting genuine guarantees, and interpretable summaries of what the aggregation features capture.

Code and Data Availability. All code needed to reproduce these results is public, with fixed random seeds, at <https://github.com/Arnob10150/CARE-Lung>: the feature-extraction and cycle-classification pipeline, the cross-validation protocol, the grouped-bootstrap and cross-conformal analyses, the device-confound audit, and the scripts regenerating every figure and table. The repository also releases the derived artifacts the reported numbers depend on: per-fold split manifests with patient identifiers, fitted thresholds, and per-patient out-of-fold scores for every aggregator and ablation variant, so each reported value is traceable to a released file. The underlying audio is the third-party ICBHI 2017 Respiratory Sound Database, which we do not redistribute; it is openly available at <https://bhichallenge.med.auth.gr/>. The cached cycle-level feature matrix derived from it is included and suffices to reproduce every reported analysis. Reasonable requests for further material may be addressed to the corresponding author at ahonaaiub.edu.

References

1. Al Mamun, S., Kaiser, M.S., Mahmud, M.: An artificial intelligence based approach towards inclusive healthcare provisioning in society 5.0: A perspective on brain disorder. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **12960 LNAI**, 157 – 169 (2021)
2. Ali, M.S., Mahmud, M.: Towards a digital twin for critical care through explainable digital human modeling of icu patients. *Lecture Notes in Computer Science* **16720 LNCS**, 466 – 485 (2026)
3. Ali, M.S., Mahmud, M.: Towards explainable hospital mortality prediction using critical care data. *Lecture Notes in Computer Science* **16720 LNCS**, 449 – 465 (2026)
4. Asif-Ur-Rahman, M., Afsana, F., Mahmud, M., Shamim Kaiser, M., Ahmed, M.R., Kaiwartya, O., James-Taylor, A.: Toward a heterogeneous mist, fog, and cloud-based framework for the internet of healthcare things. *IEEE Internet of Things Journal* **6**(3), 4049 – 4062 (2019). <https://doi.org/10.1109/JIOT.2018.2876088>
5. Ghosh, S., Kim, S., Ijaz, M.F., Singh, P.K., Mahmud, M.: Classification of mental stress from wearable physiological sensors using image-encoding-based deep neural network. *Biosensors* **12**(12) (2022). <https://doi.org/10.3390/bios12121153>
6. Hassija, V., et al.: Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation* **16**(1), 45 – 74 (2024)
7. Jubair, H., et al.: Edge-based respiratory profiling: An explainable hierarchical framework for automated COPD screening. In: *Proc. Int. Conf. on Engineering and Innovation (ICEI)*, Abstract BME106 (2026)
8. Kim, Y., et al.: Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. *Scientific Reports* **11**, 17186 (2021). <https://doi.org/10.1038/s41598-021-96724-7>

9. Mahmud, M., Baslyman, M., et al.: Inclusive ai for digital human modelling: A perspective on explainable multimodal digital twins for accessibility and inclusion. In: International Conference on Human-Computer Interaction. pp. 1–16. Springer (2026)
10. Mahmud, M., Shaffi, N., Kaiser, M.S., et al.: Privacy-preserving multi-agent systems for human-centred inclusive healthcare ai. *AI Magazine* **47**(3), e70087 (2026). <https://doi.org/10.1002/aaai.70087>
11. Mahmud, M., et al.: Towards explainable and privacy-preserving artificial intelligence for personalisation in autism spectrum disorder. *Lecture Notes in Computer Science* **13309 LNCS**, 356 – 370 (2022)
12. Mahmud, M., et al.: Preventive healthcare through privacy-preserving, explainable and inclusive artificial intelligence. *Lecture Notes in Computer Science* **16340 LNCS**, 359 – 374 (2026)
13. Mukherjee, H., Sreerama, P., Dhar, A., Obaidullah, S.M., Roy, K., Mahmud, M., Santosh, K.: Automatic lung health screening using respiratory sounds. *Journal of Medical Systems* **45**(2) (2021). <https://doi.org/10.1007/s10916-020-01681-9>
14. Nguyen, T., Pernkopf, F.: Lung sound classification using co-tuning and stochastic normalization. *IEEE Transactions on Biomedical Engineering* **69**(9), 2872–2882 (2022). <https://doi.org/10.1109/TBME.2022.3156293>
15. Rahman, M.S., et al.: Machine learning-based prognostic model for 30-day mortality prediction in sepsis-3. *BMC Medical Informatics and Decision Making* **24**(1) (2024). <https://doi.org/10.1186/s12911-024-02655-4>
16. Rahman, M.A., Brown, D.J., Shopland, N., Burton, A., Mahmud, M.: Explainable multimodal machine learning for engagement analysis by continuous performance test. *Lecture Notes in Computer Science* **13309 LNCS**, 386 – 399 (2022)
17. Rocha, B.M., et al.: An open access database for the evaluation of respiratory sound classification algorithms. *Physiological Measurement* **40**(3), 035001 (2019). <https://doi.org/10.1088/1361-6579/ab03ea>
18. Xia, T., Han, J., Mascolo, C.: Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. *Experimental Biology and Medicine* **247**(22), 2053–2061 (2022). <https://doi.org/10.1177/15353702221115428>