Supplementary Material

A Low-Cost AI Stethoscope for Resource-Constrained Heart Murmur Screening

Dominik Ludera and Marta Kersten-Oertel

Department of Computer Science and Software Engineering,
Concordia University, Montréal, QC, Canada
dominik.ludera@mail.concordia.ca

This supplement expands the clinical-endpoint evaluation of Sec. 3.2 of the paper with full interval estimates and per-fold detail, reports the two supporting analyses referenced there (the two-class murmur configuration and the checkpoint-selection sensitivity analysis), and holds the prototype photograph, bill of materials and preprocessing ablation that the paper's length budget did not accommodate. All numbers derive from the same pooled out-of-fold predictions ($n = 941$ patients, five disjoint patient-level folds) used in the paper; intervals are Wilson score 95% intervals. Leading zeros are omitted throughout. NNR is the number needed to refer, 1/PPV.

1 Clinical-Endpoint Trade-offs

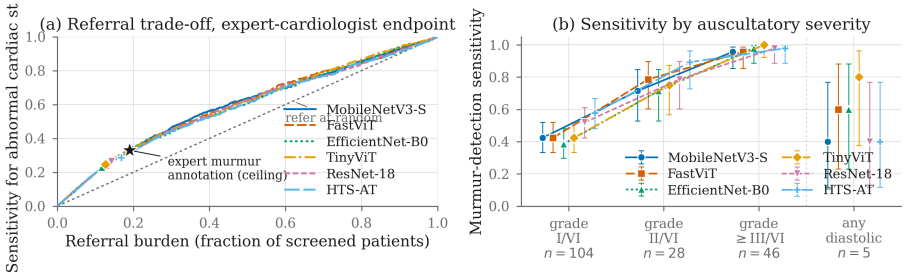

Fig. S1. (a) Sensitivity for abnormal cardiac status (endpoint E1) against referral burden, sweeping the decision threshold on the abnormality score. Markers give each model's deployed argmax operating point; the star gives the expert murmur annotation applied as a referral rule. Every model tracks the expert ceiling and none separates from the others, so the architecture-independence seen for murmur classification also holds at the clinical endpoint. (b) Murmur-detection sensitivity stratified by auscultatory severity, with Wilson 95% intervals. Systolic grades are ordinal and joined; diastolic murmurs form a separate stratum and are drawn detached. Sensitivity rises monotonically with grade for all six architectures.

Figure S1 accompanies Sec. 3.2 of the paper. Panel (a) shows that the models neither beat nor fall short of expert auscultation as a referral rule for abnormal

cardiac status: their threshold sweeps pass through the expert operating point, so the low sensitivity at this endpoint is a property of auscultation rather than of the models. Panel (b) shows where the residual errors sit — overwhelmingly in barely audible grade I/VI murmurs, which are also the ones most likely to be innocent.

2 Full Diagnostic Accuracy per Clinical Endpoint

Tables S1–S3 give sensitivity, specificity, PPV and NPV with 95% intervals for the three endpoints defined in Sec. 3.2 of the paper, under the deployed decision rule (predict *Present* → refer). The two italic rows are reference rules rather than models: the expert murmur annotation applied directly as a referral rule, which upper-bounds any murmur detector, and referring every screened patient.

Table S1. Endpoint E1, abnormal cardiac status per the examining cardiologist (455/941, 48.4% prevalence).

Referral rule	Sensitivity	Specificity	PPV	NPV	NNR
MobileNetV3-S (1.52M)	.235 [.199, .276]	.981 [.965, .990]	.922 [.859, .959]	.578 [.544, .611]	1.08
FastViT (3.26M)	.240 [.203, .281]	.981 [.965, .990]	.924 [.861, .959]	.580 [.546, .613]	1.08
EfficientNet-B0 (4.01M)	.231 [.194, .272]	.990 [.976, .996]	.955 [.898, .980]	.579 [.545, .612]	1.05
TinyViT (5.07M)	.246 [.209, .288]	.986 [.971, .993]	.941 [.884, .971]	.583 [.549, .616]	1.06
ResNet-18 (11.2M)	.264 [.225, .306]	.971 [.952, .983]	.896 [.832, .937]	.585 [.551, .618]	1.12
HTS-AT (27.6M)	.288 [.248, .331]	.944 [.920, .962]	.829 [.763, .880]	.586 [.551, .620]	1.21
Expert murmur annotation	.330 [.288, .374]	.940 [.916, .958]	.838 [.777, .885]	.600 [.565, .634]	1.19
Refer every patient	1.000 [.992, 1.000]	.000 [.000, .008]	.484 [.452, .515]	—	2.07

Table S2. Endpoint E2, pathological murmur, i.e. a murmur the cardiologist also judged abnormal (150/941, 15.9%).

Referral rule	Sensitivity	Specificity	PPV	NPV	NNR
MobileNetV3-S (1.52M)	.673 [.595, .743]	.981 [.969, .988]	.871 [.798, .920]	.941 [.922, .955]	1.15
FastViT (3.26M)	.693 [.615, .762]	.982 [.971, .989]	.881 [.811, .928]	.944 [.926, .958]	1.13
EfficientNet-B0 (4.01M)	.673 [.595, .743]	.989 [.979, .994]	.918 [.852, .956]	.941 [.923, .955]	1.09
TinyViT (5.07M)	.713 [.636, .780]	.985 [.974, .991]	.899 [.832, .941]	.948 [.930, .961]	1.11
ResNet-18 (11.2M)	.760 [.686, .821]	.975 [.961, .984]	.851 [.781, .901]	.955 [.939, .968]	1.18
HTS-AT (27.6M)	.780 [.707, .839]	.948 [.930, .962]	.741 [.667, .803]	.958 [.941, .970]	1.35
Expert murmur annotation	1.000 [.975, 1.000]	.963 [.948, .974]	.838 [.777, .885]	1.000 [.995, 1.000]	1.19
Refer every patient	1.000 [.975, 1.000]	.000 [.000, .005]	.159 [.137, .184]	—	6.27

The comparison across the three tables isolates where the endpoint, rather than the model, is the limiting factor. Sensitivity for E1 is low for every rule, including the expert annotation, because 262 of the 455 patients the cardiologist judged abnormal have no audible murmur; PPV, by contrast, is high throughout. Moving to E2 and E3—endpoints that a murmur can in principle indicate—sensitivity rises to 0.67–0.78 and then 0.92–0.98 while NPV reaches 0.995–0.999.

Table S3. Endpoint E3, non-innocent murmur, i.e. systolic grade ≥III/VI or any diastolic murmur (50/941, 5.3%).

Referral rule	Sensitivity	Specificity	PPV	NPV	NNR
MobileNetV3-S (1.52M)	.920 [.812, .968]	.921 [.902, .937]	.397 [.312, .488]	.995 [.988, .998]	2.52
FastViT (3.26M)	.940 [.838, .979]	.920 [.901, .936]	.398 [.315, .488]	.996 [.989, .999]	2.51
EfficientNet-B0 (4.01M)	.960 [.865, .989]	.930 [.912, .945]	.436 [.347, .530]	.998 [.991, .999]	2.29
TinyViT (5.07M)	.980 [.895, .996]	.921 [.902, .937]	.412 [.327, .502]	.999 [.993, 1.000]	2.43
ResNet-18 (11.2M)	.940 [.838, .979]	.902 [.881, .920]	.351 [.275, .435]	.996 [.989, .999]	2.85
HTS-AT (27.6M)	.940 [.838, .979]	.875 [.852, .896]	.297 [.232, .373]	.996 [.989, .999]	3.36
Expert murmur annotation	1.000 [.929, 1.000]	.855 [.831, .877]	.279 [.219, .349]	1.000 [.995, 1.000]	3.58
Refer every patient	1.000 [.929, 1.000]	.000 [.000, .004]	.053 [.041, .069]	—	18.82

3 Per-Fold Variability of the Clinical Endpoint

Table S4 reports fold-level sensitivity and PPV for endpoint E1, confirming that the pooled estimates in the paper are not driven by a single favourable partition: the fold-to-fold standard deviation is ≤0.036 for sensitivity and ≤0.068 for PPV.

4 Severity-Stratified Detection Sensitivity

Table S5 gives the numbers plotted in Fig. 3(b) of the paper, with interval estimates and stratum sizes, plus the two strata that separate murmurs the cardiologist judged abnormal from those judged normal. The last two rows quantify the asymmetry that supports the triage framing: murmurs accompanied by an abnormal diagnosis are detected 1.6–4.9× more often than murmurs accompanied by a normal one, the ratio being largest for EfficientNet-B0 (4.9×) and smallest for HTS-AT (1.6×). Note that CirCor assigns grade I/VI both to barely audible murmurs and to murmurs not captured at every auscultation site, so grade is an informative but imperfect severity proxy.

5 Operating Points Beyond the Deployed Threshold

The deployed rule is a fixed argmax, but a screening programme may prefer to trade precision for sensitivity. Table S6 reports, for each model, the smallest referral burden that attains a target sensitivity for abnormal cardiac status, together with the PPV at that operating point. Reaching a sensitivity of 0.50 requires referring 33.7–36.9% of those screened at PPV 0.657–0.719; reaching 0.75 requires referring 63.4–67.1% at PPV 0.542–0.573. The curves are nearly indistinguishable across architectures, mirroring the architecture-independence observed for murmur classification.

6 Two-Class Murmur Configuration

The *Unknown* label in CirCor marks recordings whose quality prevented the annotator from deciding, not a distinct clinical state. Table S7 reports the same

Table S4. Per-fold sensitivity and PPV for endpoint E1 (abnormal cardiac status), deployed decision rule.

	Sensitivity by fold					
Model	**0**	**1**	**2**	**3**	**4**	**mean±s.d.**
MobileNetV3-S	.226	.235	.276	.220	.217	.235±.021
FastViT	.215	.235	.265	.231	.250	.239±.017
EfficientNet-B0	.226	.235	.276	.198	.217	.230±.026
TinyViT	.280	.259	.255	.231	.207	.246±.025
ResNet-18	.269	.284	.306	.198	.261	.264±.036
HTS-AT	.323	.284	.327	.242	.261	.287±.033
	PPV by fold					
Model	**0**	**1**	**2**	**3**	**4**	**mean±s.d.**
MobileNetV3-S	.955	1.000	.931	.952	.800	.928±.068
FastViT	.952	.950	.929	.913	.885	.926±.025
EfficientNet-B0	.955	1.000	.931	.947	.952	.957±.023
TinyViT	.963	.955	.893	1.000	.905	.943±.039
ResNet-18	.926	.920	.882	.947	.828	.901±.042
HTS-AT	.789	.885	.842	.846	.800	.832±.034

patient-level five-fold protocol with *Unknown* merged into *Present*, for the four architectures we ran in both configurations. Macro F1 rises from 0.562–0.618 in the three-class setting to 0.770–0.810, confirming that the weak three-class *Unknown* performance reflects label scarcity rather than an architectural limitation. Specificity stays high (0.929–0.981) while murmur sensitivity remains the binding constraint (0.489–0.665), and HTS-AT again trades precision for sensitivity.

7 Checkpoint-Selection Sensitivity Analysis

Because the best checkpoint of each fold is chosen by held-out accuracy, that selection shares data with evaluation. Table S8 bounds the resulting optimism using the four checkpoints each run retains: the peak checkpoint exceeds the mean of the retained set by 0.38–1.44 percentage points, mean 0.79 pp. For reference, the entire spread of Challenge Scores across the six architectures is 2.1 pp, so the selection effect is of the same order as the between-architecture differences

Table S5. Murmur-detection sensitivity by stratum, with Wilson 95% intervals. n is the stratum size (identical across models).

Stratum	n	MobileNetV3-S	FastViT	EffNet-B0	TinyViT	ResNet-18	HTS-AT
Systolic grade I/VI	104	.423 [.333, .519]	.423 [.333, .519]	.385 [.297, .481]	.423 [.333, .519]	.519 [.424, .613]	.577 [.481, .667]
Systolic grade II/VI	28	.714 [.529, .847]	.786 [.605, .898]	.714 [.529, .847]	.750 [.566, .873]	.786 [.605, .898]	.893 [.728, .963]
Systolic grade ≥III/VI	46	.957 [.855, .988]	.957 [.855, .988]	.978 [.887, .996]	1.000 [.923, 1.000]	.978 [.887, .996]	.978 [.887, .996]
Any diastolic murmur	5	.400 [.118, .769]	.600 [.231, .882]	.600 [.231, .882]	.800 [.376, .964]	.400 [.118, .769]	.400 [.118, .769]
Murmur present, status abnormal	150	.673 [.595, .743]	.693 [.615, .762]	.673 [.595, .743]	.713 [.636, .780]	.760 [.686, .821]	.780 [.707, .839]
Murmur present, status normal	29	.241 [.122, .421]	.241 [.122, .421]	.138 [.055, .306]	.172 [.076, .345]	.241 [.122, .421]	.483 [.314, .656]

Table S6. Referral burden (%) and PPV at the cheapest operating point attaining each target sensitivity for endpoint E1. Cells read *burden / PPV*.

Model	Sens ≥.35	Sens ≥.40	Sens ≥.50	Sens ≥.60	Sens ≥.75
MobileNetV3-S	20.9 / .812	24.3 / .795	33.7 / .719	45.1 / .644	65.9 / .552
FastViT	20.7 / .821	24.9 / .778	35.1 / .691	46.0 / .630	63.4 / .573
EfficientNet-B0	20.1 / .847	25.9 / .746	36.1 / .671	49.3 / .588	66.4 / .547
TinyViT	20.4 / .833	25.0 / .774	35.9 / .675	47.2 / .615	65.1 / .558
ResNet-18	21.8 / .780	25.5 / .758	35.6 / .681	46.7 / .622	66.6 / .545
HTS-AT	22.0 / .773	25.9 / .746	36.9 / .657	48.6 / .597	67.1 / .542

it might be thought to manufacture. Since the identical procedure is applied to every architecture, it cannot account for a ranking among them.

8 Prototype and Bill of Materials

Figure S2 and Table S9 accompany Sec. 2.6 of the paper. The prototype is assembled entirely from off-the-shelf parts; prices are single-unit retail converted to USD, with multi-pack items charged per unit, and exclude shipping, taxes and the stethoscope head to which the microphone couples.

9 Preprocessing Ablation and Class-Weight Sensitivity

Table S10 accompanies Sec. 3.6 of the paper, which states the findings. Each block varies one setting against the baseline of 30 s windows, a 25–400 Hz band-pass, and multi-location restructuring, on a fixed held-out fold with identical

Table S7. Two-class murmur task (*Unknown* merged into *Present*), patient-level five-fold cross-validation, mean±s.d. over folds. Sensitivity and precision refer to the murmur class.

Model	Accuracy	Macro F1	Sensitivity	Precision	Specificity
EfficientNet-B0	.852±.017	.770±.035	.489±.070	.902±.029	.981±.006
FastViT	.850±.018	.772±.035	.509±.068	.866±.082	.971±.016
TinyViT	.855±.020	.780±.041	.530±.100	.870±.017	.971±.009
HTS-AT	.860±.006	.810±.006	.665±.035	.772±.034	.929±.014

Table S8. Held-out accuracy of the peak retained checkpoint versus the mean over the four retained checkpoints, averaged over the five folds.

Model	Peak checkpoint	Mean of retained	Difference (pp)
MobileNetV3-S	.857	.850	0.73
FastViT	.858	.850	0.79
EfficientNet-B0	.856	.849	0.63
TinyViT	.863	.855	0.78
ResNet-18	.847	.843	0.38
HTS-AT	.804	.790	1.44

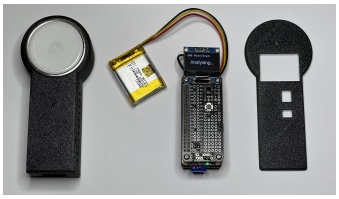

Fig. S2. Prototype integrating an Orange Pi Zero 2 W, OLED display, PiJuice power subsystem, and a custom 3D-printed enclosure.

hyperparameters. Note that single-location training attains the highest accuracy and Challenge Score while collapsing Unknown F1, which is why the deployed pipeline retains restructuring: the metric that matters for a three-class screener is minority-class recall, not aggregate accuracy.

Class-weight sensitivity was evaluated on TinyViT, fold 0. Relative to the selected inverse-frequency weights $[4, 10, 1]$ (Challenge Score 0.918, Macro F1 0.683, Unknown F1 0.300), lighter weighting $[2, 5, 1]$ raises the Challenge Score to 0.926 but halves Unknown F1 to 0.125, and heavier weighting $[6, 15, 1]$ lowers the Challenge Score to 0.906 at Unknown F1 0.235. The chosen weighting therefore maximises Macro F1 and minority-class recall simultaneously.

Table S9. Prototype bill of materials (single-unit retail prices converted to USD; multi-pack items charged per unit).

Component	USD	Component	USD
SBC (Allwinner H618, 4 GB)	$36	USB microphone module	$11
Power-management HAT	$43	Aluminium heatsink	$9
1200 mAh Li-ion battery	$17	0.96″ OLED display	$4
32 GB microSD card	$21	Prototyping HAT	$4
3D-printed PLA enclosure	$7	Push buttons (×2)	$1
		Indicative total	**∼$153**

Table S10. Preprocessing ablation (fixed held-out fold). Baseline: 30 s windows, 25–400 Hz bandpass, and multi-location restructuring.

Condition	Accuracy	Macro F1	Chall. Score	Unknown F1
Baseline (30s / BP / Multi)	0.873	**0.678**	0.907	**0.316**
Location: single-location only	**0.904**	0.661	**0.945**	0.125
Window: 10 s	0.852	0.588	0.891	0.125
Window: 15 s	0.857	0.561	0.905	0.000
Filter: none	0.857	0.558	0.902	0.000
Filter: 25–1000 Hz	0.857	0.557	0.899	0.000

10 Reproducibility Notes

Patients are assigned to five folds by deterministic round-robin over sorted patient identifiers with seed 42, so the splits are reconstructible without shipping index files. Segment-level logits are mean-pooled per patient before argmax; the continuous screening score used for Table S6 and Fig. 3(a) of the paper is the softmax of the pooled logits, summed over the *Present* and *Unknown* classes. Pooling the five held-out folds yields one prediction per patient for 941 of the 942 patients in the public partition; one patient was dropped during dataset preparation and is excluded from every analysis. Clinical reference labels are taken verbatim from the `training_data.csv` distributed with CirCor DigiScope: the `Outcome` column for cardiac status, and the `Systolic murmur grading` and `Diastolic murmur grading` columns for severity. The code that produces every table and figure in this supplement, together with the trained models, deployment scripts and hardware documentation, is available at `https://github.com/domludera/mirasol2026`.