Supplementary Information

A multi-corpus benchmark reveals limits of diagnostic preservation in heart-sound denoising

Eashita Wazed, Choonsung Shin, Hieyong Jeong

Mathematical verification of the preservation conditions

Proposition S1 (AUROC rank preservation). Let q_i be the original classifier score, $\widehat{q}_{id} = q_i + e_{id}$ the score after denoiser d, and $\gamma_{ij} = q_i - q_j$ for every positive–negative pair. If $|e_{id}| \le \eta$ for all recordings and $\min_{i,j} |\gamma_{ij}| > 2\eta$, preprocessing preserves every pair ordering and hence $\Delta U = 0$.

Proof. For every pair,

$$|e_{id} - e_{jd}| \le |e_{id}| + |e_{jd}| \le 2\eta < |\gamma_{ij}|.$$

The perturbation cannot change the sign of γ_{ij}, so every concordance term after preprocessing equals its original value. Their average, and therefore AUROC, is unchanged. The condition is sufficient rather than necessary because pair exchanges in opposite directions can cancel.

Main-text Eq. (2) also gives a direct data interpretation. With 150 positive and 150 negative recordings, $N_+N_- = 22{,}500$. Thus $22{,}500\Delta U$ is the net number of concordant-pair equivalents gained after preprocessing, with a tie contributing one half. The original-recording BMD-HS effect $\Delta U = -0.387$ therefore represents approximately 8,708 more pair-order losses than gains, subject to rounding.

Proposition S2 (fixed-threshold decision preservation). For fixed threshold t, preprocessing cannot change recording i's binary decision if $|e_{id}| < |q_i - t|$.

Proof. The score perturbation is smaller than the distance from the original score to the threshold. Therefore $q_i - t$ and $q_i + e_{id} - t$ have the same sign. Conversely, a changed decision requires a perturbation at least as large as its original threshold margin. A sensitivity change of -0.767 among 150 positives corresponds to approximately 115 fewer true-positive decisions at the fixed reference threshold.

Proposition S3 (non-identifiability of residual source). The observed pair $(x, D(x))$ does not uniquely identify how much of the transformation residual came from patient-derived cardiac content rather than extrinsic noise.

Proof. Suppose $x = s + n$. For any compatible waveform h, define $s' = s + h$ and $n' = n - h$. Then $s' + n' = x$, so the same observed input, denoised output and residual are compatible with infinitely many content/noise decompositions. A paired clean target or additional source-identification assumptions are required to distinguish noise removal from signal distortion. A residual-restoration intervention can test diagnostic contribution, while validated event annotation can strengthen localisation; neither alone identifies biological source. Main-text Eq. (3) localises observable attenuation to event-aligned time–frequency regions but cannot prove that every attenuated component was disease-causal cardiac signal.

Together, Propositions S1–S3 justify the inferential boundary: paired data can falsify strict diagnostic preservation and localise event-aligned attenuation, but a prospectively chosen non-inferiority margin is required to certify safety and an additional identification experiment is required for causal labelling of the residual.

Benchmark resource and training provenance

The resource framework treated an executable resource—a dataset, denoising implementation, interference source or generator, or diagnostic classifier—as the unit of benchmark construction rather than treating a publication title as sufficient evidence of executability. Four documented stages were used: candidate identification within the cited literature and archived experimental resources, scientific eligibility,

executable eligibility and complementary role assignment. This was not a systematic review and makes no claim of exhaustive worldwide coverage. Candidate identification was literature-guided and included clinical datasets, waveform-output denoisers, recorded-noise sources, reproducible generated perturbation mechanisms and diagnostic architectures. Scientific screening required human PCG waveforms, stable identifiers, mappable diagnostic labels and a grouping unit for datasets; waveform output for denoisers; a traceable recorded source or fully specified generated mechanism for interference; and a label-compatible diagnostic score for classifiers. Executable eligibility then required deterministic conversion or generation to the common interface, reproducible cohort construction, the strongest training–evaluation separation permitted by released metadata, and runnable frozen code and weights or sufficient detail for a traceable harmonised implementation. Reported performance, citation count, venue and the observed results of this benchmark were not selection criteria.

Eligible clinical corpora were assigned non-redundant roles instead of being pooled as interchangeable samples. CirCor provided a patient-disjoint development-distribution test; P16 provided heterogeneous external recordings and the abnormality-classifier source; BMD-HS provided patient-grouped multi-disease transfer and original recordings without a paired clean reference; and PaHS provided an independent original-recording challenge with a recoverable parent-recording grouping. The selected interference panel likewise represented environmental, physiological, hospital-ambient, swept-frequency, narrowband and impulsive mechanisms. This design supports contrasts across specified axes, not an estimate of field-wide method frequency or clinical prevalence.

Model provenance was separated from architectural inspiration. The fcn, unet and lunet systems were local architectural baselines adapted from published designs and trained on the common study partition, whereas the other seven learned denoisers were study-developed variants assembled from cited components. The four classical filters and identity were deterministic controls, and the four classifiers were fixed outcome probes rather than denoisers. The learned panel covered convolutional, recurrent, attention-based, complex time–frequency and two-stage context; reconstruction-only and task-consistency objectives; and standard versus impulse-matched exposure. Because capacity, architecture and objective were not varied factorially, family summaries are robustness checks rather than causal estimates of individual design choices.

Training provenance was linked to retained artefacts at two levels. For the seven task-aware and impulse-trained systems, configured epochs, batch size, learning rates, seed, mixture range and split fractions were read from checkpoint-embedded `train_args`; optimiser type, gradient clipping, objectives and minimum-validation-loss retention were verified from matching trainer control flow. For fcn, unet and lunet, the checkpoints stored model state, validation SI-SDR and architecture but not run arguments. Their values are therefore archived-trainer defaults: 50 epochs, AdamW, learning rate 2×10^{-4}, batch size 16, gradient clipping at 5.0 and selection by highest mean validation SI-SDR. For the classifiers, the identified production trainer wrote the exact directory, filename pattern and seven-field checkpoint schema found in all eight retained files. Checkpoint metadata independently confirmed architecture, source corpus, 64 mel bands, 157 frames, training-set size and two-class output encoding. The production source specified default settings of 30 epochs, Adam, learning rate 10^{-3}, batch size 32, cross-entropy, inverse-frequency replacement sampling, seed 42, no internal validation, scheduler or early stopping, and retention of the final configured epoch. A separate fold/smoke-test harness used a different storage schema and was not used as provenance for the evaluated checkpoints. Because classifier command-line arguments were not embedded in the retained checkpoints and a run-level command log was not preserved, an unrecorded classifier override cannot be independently excluded. The same limitation applies to the three reconstruction baselines, but not to the seven learned systems whose `train_args` were retained. Supplementary Tables S11 and S12 distinguish checkpoint-verified from source-linked configuration evidence.

Supplementary figures

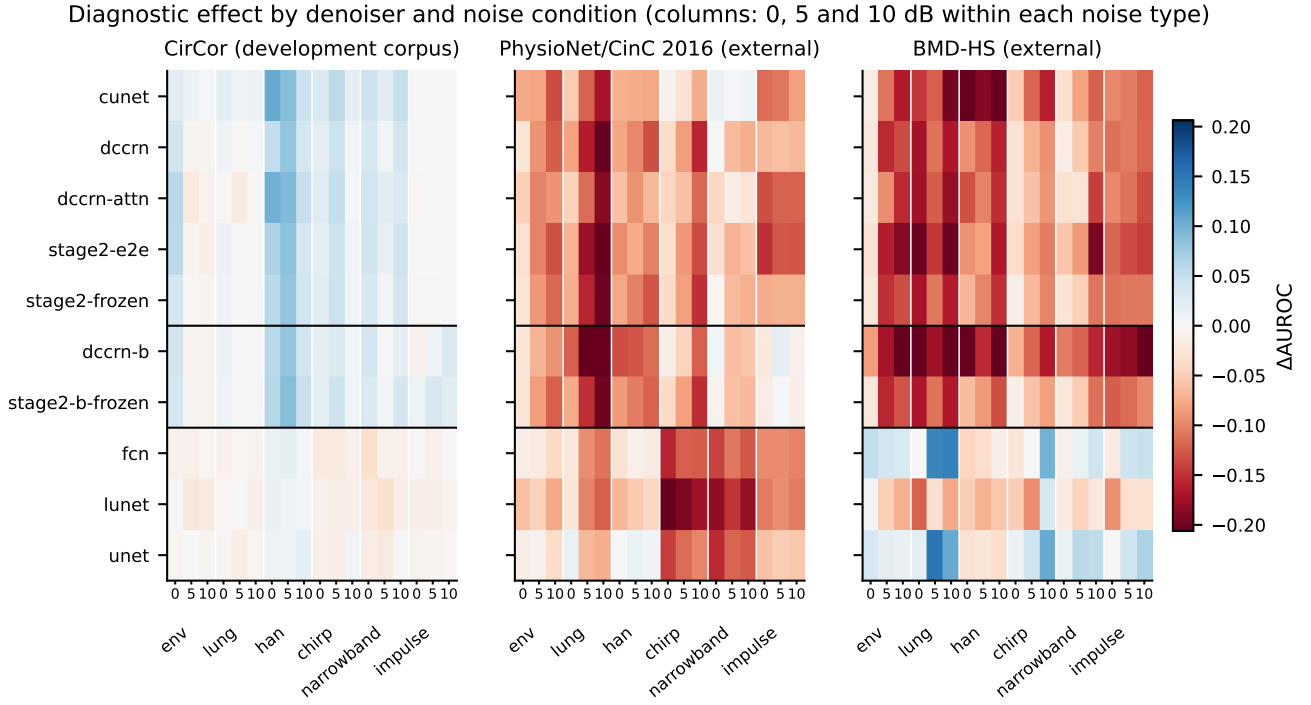


Figure S1: Diagnostic effect by learned denoiser and artificial-corruption condition. Rows are ten denoisers, grouped by horizontal rules; columns are six interference types, each at 0, 5 and 10 dB. Each cell is paired ΔAUROC relative to its undenoised corrupted input for the primary classifier and the same 300 recordings. The first five rows are task-aware, the next two impulse-trained and the final three reconstruction-only. Red denotes degradation and blue improvement.

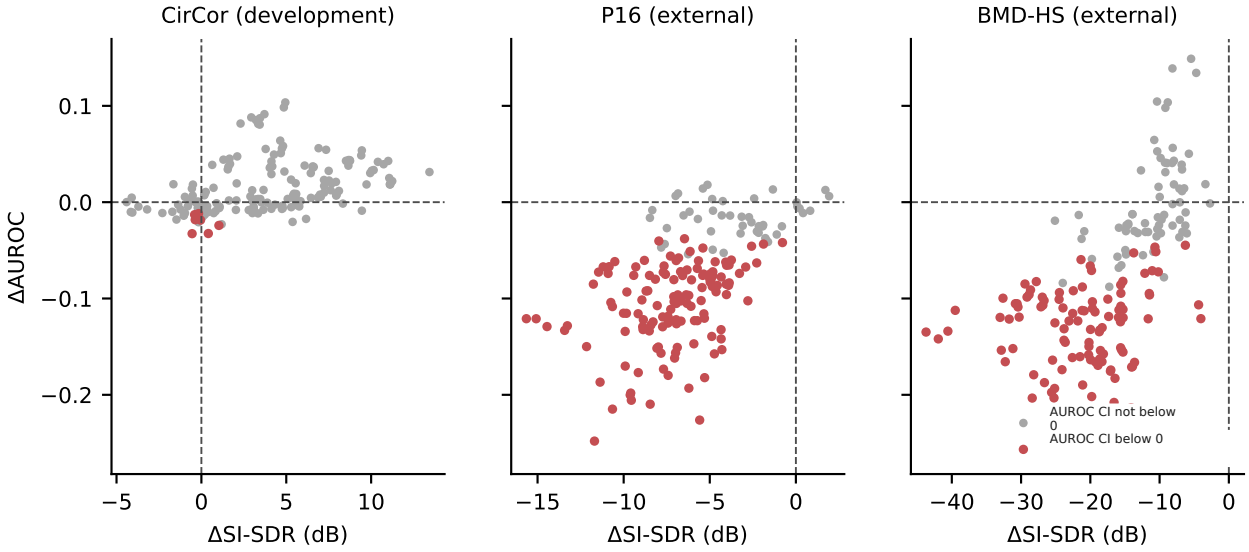


Figure S2: Reference-based fidelity and diagnostic utility. Change in SI-SDR versus paired ΔAUROC for each learned-denoiser–artificial-condition pair ($n = 180$ per corpus). Red denotes a diagnostic 95% bootstrap interval wholly below zero. The clean reference exists because the noise was added artificially; this analysis is unavailable for original recordings without artificial corruption.

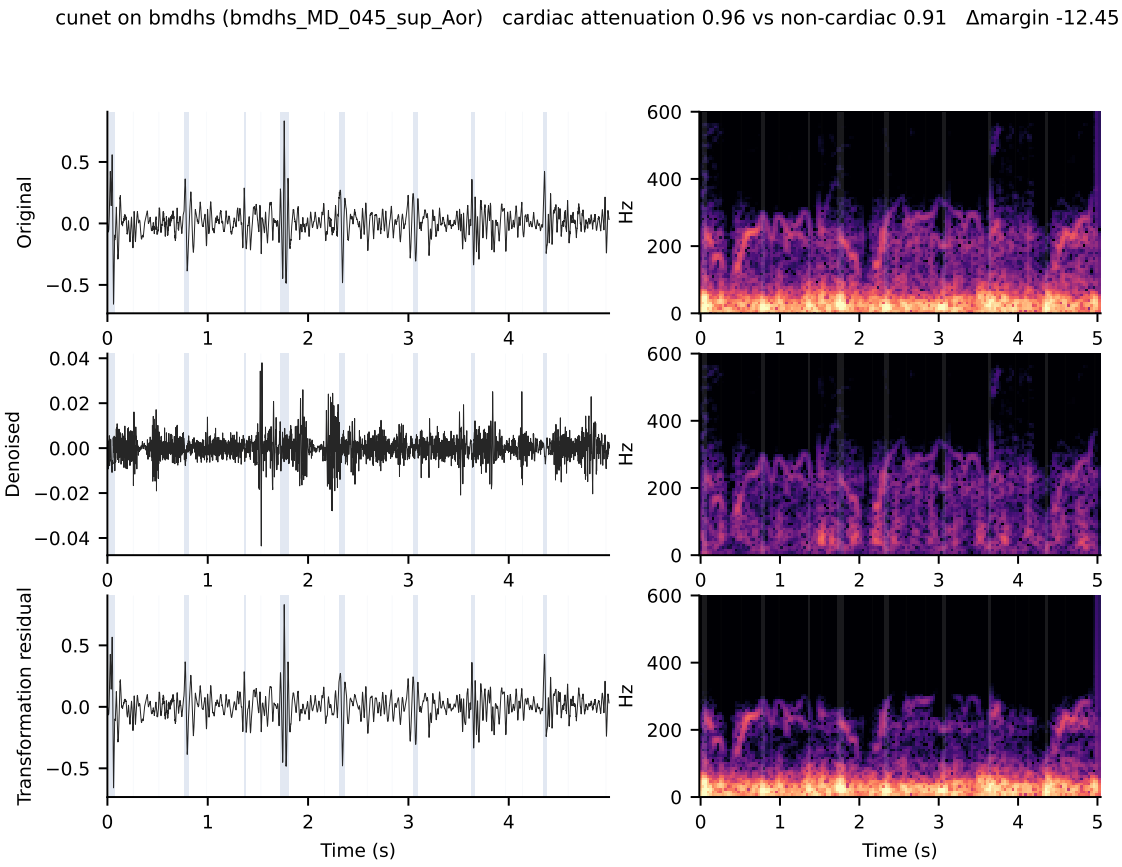


Figure S3: Rule-selected illustrative BMD-HS case. Original recording, cunet output and transformation residual with their spectrograms. Shading marks cardiac-event regions detected from the original recording. The case is the median correct-class margin loss among recordings changing from correct to incorrect classification, not a visually selected extreme. Panel-specific vertical scales should be considered when comparing waveforms.

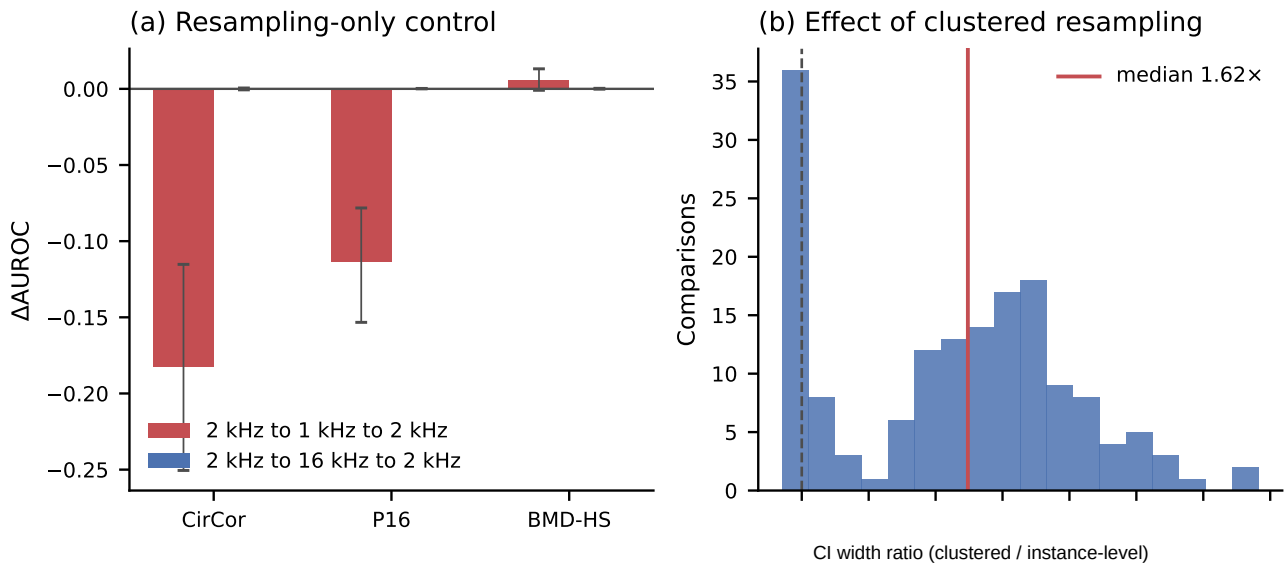


Figure S4: Resampling and clustered-inference controls. (a) Paired ΔAUROC for no-added-noise recordings passed through resample-out/resample-back paths without denoising; error bars are clustered 95% bootstrap intervals. P16 abbreviates PhysioNet/CinC 2016. (b) Ratio of patient-clustered to instance-level confidence-interval widths; the solid vertical line marks the median 1.62.

Table S1: Diagnostic effect of denoising by denoiser family and corpus, aggregated over the 18 artificially corrupted conditions. Values are the mean change in AUROC relative to the undenoised input; counts give the number of denoiser–condition comparisons whose clustered 95% confidence interval lies entirely below or above zero. Clustering uses patients for CirCor and BMD-HS and recordings for P16; the interval-based counts are descriptive.

	Mean ΔAUROC			Degraded / improved (of n)		
Denoiser family	CirCor	P16	BMD-HS	CirCor	P16	BMD-HS
Task-aware	+0.023	−0.087	−0.111	0 / 25 (90)	71 / 0 (90)	72 / 0 (90)
Impulse-trained	+0.022	−0.079	−0.133	0 / 10 (36)	24 / 0 (36)	28 / 0 (36)
Reconstruction-only	−0.006	−0.084	+0.004	7 / 0 (54)	39 / 0 (54)	8 / 14 (54)

Supplementary tables

Table S2: No-added-noise passthrough. Change in AUROC when each denoiser is applied to unmodified recordings without added artificial noise. Confidence intervals use patients as the clustering unit for CirCor and BMD-HS and recordings for P16.

		CirCor		P16		BMD-HS	
Family	Denoiser	Δ	95% CI	Δ	95% CI	Δ	95% CI
Task-aware	cunet	+0.002	[−0.003, +0.006]	−0.102	[−0.140, −0.069]	−0.380	[−0.458, −0.307]
	dccrn	−0.000	[−0.005, +0.004]	−0.134	[−0.179, −0.093]	−0.257	[−0.335, −0.182]
	dccrn_attn	−0.002	[−0.008, +0.004]	−0.141	[−0.185, −0.099]	−0.316	[−0.396, −0.238]
	stage2_e2e	+0.000	[−0.006, +0.005]	−0.179	[−0.231, −0.132]	−0.354	[−0.432, −0.274]
	stage2_frozen	−0.001	[−0.006, +0.004]	−0.136	[−0.181, −0.095]	−0.256	[−0.332, −0.185]
Impulse-trained	dccrn_b	−0.002	[−0.008, +0.003]	−0.116	[−0.156, −0.080]	−0.371	[−0.453, −0.286]
	stage2_b_frozen	−0.002	[−0.007, +0.003]	−0.134	[−0.180, −0.095]	−0.250	[−0.326, −0.179]
Reconstruction-only	fcn	−0.005	[−0.021, +0.013]	−0.155	[−0.198, −0.110]	+0.033	[−0.042, +0.111]
	lunet	−0.016	[−0.030, −0.003]	−0.127	[−0.168, −0.088]	−0.112	[−0.192, −0.029]
	unet	+0.002	[−0.010, +0.015]	−0.107	[−0.145, −0.072]	+0.064	[−0.001, +0.138]

Table S3: Operating-point change on original recordings without artificial corruption with the decision threshold held fixed. The threshold is the Youden-optimal point of the undenoised reference for each corpus and classifier, applied unchanged to the denoised condition, as it would be when a denoiser is inserted into an existing pipeline; the final column repeats the analysis with the threshold fitted on a held-out half of the clusters and averaged over random splits. Values are for the primary classifier.

		Fixed reference threshold			Held-out threshold
Corpus	Denoiser	ΔSens.	ΔSpec.	ΔBal. acc.	ΔBal. acc.
BMD-HS	cunet	−0.753	+0.300	−0.227	−0.209
	dccrn	−0.753	+0.293	−0.230	−0.212
	dccrn_attn	−0.767	+0.307	−0.230	−0.215
	stage2_e2e	−0.767	+0.307	−0.230	−0.215
	stage2_frozen	−0.753	+0.307	−0.223	−0.207
	dccrn_b	−0.733	+0.287	−0.223	−0.207
	stage2_b_frozen	−0.753	+0.307	−0.223	−0.207
	fcn	+0.000	−0.067	−0.033	−0.013
	lunet	−0.567	+0.260	−0.153	−0.138
	unet	−0.007	+0.020	+0.007	+0.034
PaHS	cunet	+0.007	+0.000	+0.003	+0.010
	dccrn	−0.007	+0.000	−0.003	+0.001
	dccrn_attn	+0.007	+0.000	+0.003	+0.004
	stage2_e2e	+0.067	−0.127	−0.030	−0.019
	stage2_frozen	+0.060	−0.200	−0.070	−0.052
	dccrn_b	+0.013	+0.000	+0.007	+0.002
	stage2_b_frozen	+0.127	−0.313	−0.093	−0.085
	fcn	+0.187	−0.580	−0.197	−0.177
	lunet	+0.180	−0.587	−0.203	−0.184
	unet	+0.133	−0.347	−0.107	−0.096

Table S4: Operating-point changes after denoising of the original recordings. Sensitivity and specificity are evaluated at a Youden-optimal threshold re-estimated within each condition; these values are descriptive best-case summaries rather than fixed-threshold deployment estimates. AUPRC intervals use patients as the clustering unit for BMD-HS and recordings for PaHS.

Denoiser	Corpus	Classifier	ΔSens.	ΔSpec.	ΔBal. acc.	ΔAUPRC	95% CI
cunet	BMD-HS	cnn2d	−0.753	+0.307	−0.223	−0.345	[−0.420, −0.244]
cunet	PaHS	cnn2d	+0.013	+0.000	+0.007	+0.004	[−0.002, +0.012]
dccrn	BMD-HS	cnn2d	+0.107	−0.533	−0.213	−0.286	[−0.369, −0.182]
dccrn	PaHS	cnn2d	+0.013	+0.000	+0.007	+0.001	[−0.001, +0.004]
dccrn_attn	BMD-HS	cnn2d	−0.727	+0.287	−0.220	−0.305	[−0.384, −0.202]
dccrn_attn	PaHS	cnn2d	+0.013	+0.000	+0.007	+0.004	[−0.000, +0.010]
dccrn_b	BMD-HS	cnn2d	−0.740	+0.307	−0.217	−0.320	[−0.407, −0.214]
dccrn_b	PaHS	cnn2d	+0.013	+0.000	+0.007	+0.007	[+0.003, +0.012]
fcn	BMD-HS	cnn2d	−0.093	+0.060	−0.017	+0.007	[−0.061, +0.076]
fcn	PaHS	cnn2d	−0.033	+0.007	−0.013	+0.012	[−0.016, +0.040]
lunet	BMD-HS	cnn2d	−0.253	+0.053	−0.100	−0.110	[−0.183, −0.038]
lunet	PaHS	cnn2d	+0.013	−0.027	−0.007	+0.001	[−0.026, +0.027]
stage2_b_frozen	BMD-HS	cnn2d	−0.560	+0.140	−0.210	−0.273	[−0.358, −0.172]
stage2_b_frozen	PaHS	cnn2d	−0.040	+0.047	+0.003	+0.022	[+0.002, +0.043]
stage2_e2e	BMD-HS	cnn2d	−0.753	+0.307	−0.223	−0.317	[−0.398, −0.216]
stage2_e2e	PaHS	cnn2d	+0.013	+0.000	+0.007	+0.034	[+0.015, +0.055]
stage2_frozen	BMD-HS	cnn2d	+0.000	−0.420	−0.210	−0.277	[−0.361, −0.175]
stage2_frozen	PaHS	cnn2d	−0.047	+0.053	+0.003	+0.020	[+0.002, +0.041]
unet	BMD-HS	cnn2d	−0.127	+0.187	+0.030	+0.047	[−0.010, +0.114]
unet	PaHS	cnn2d	−0.040	+0.007	−0.017	+0.006	[−0.023, +0.033]

Table S5: Direction of reference-based fidelity and diagnostic changes in artificially corrupted experiments. Each learned-denoiser–condition pair is classified by the sign of denoised-minus-noisy SI-SDR relative to the clean reference and by the sign of paired ΔAUROC. Counts are descriptive and do not encode confidence-interval or multiplicity thresholds.

Fidelity / classifier AUROC	CirCor	P16	BMD-HS
fidelity improved / AUROC improved	97	2	0
fidelity improved / AUROC degraded	39	5	0
fidelity degraded / AUROC improved	11	8	26
fidelity degraded / AUROC degraded	33	165	154

Table S6: Whole-recording processing intensity and corresponding no-added-noise diagnostic effect. Positive attenuation is the fraction of positive short-time magnitude attenuated from the whole recording. The table includes ten learned denoisers, four classical filters and the identity control, all applied to the same recordings and scored by the same primary classifier.

Corpus	Method	Family	Positive attenuation	ΔAUROC
CirCor	identity	Identity control	0.000	+0.000
CirCor	lunet	Reconstruction-only	0.002	−0.016
CirCor	unet	Reconstruction-only	0.003	+0.002
CirCor	stage2_frozen	Task-aware/impulse-trained	0.023	−0.001
CirCor	dccrn	Task-aware/impulse-trained	0.026	−0.000
CirCor	dccrn_attn	Task-aware/impulse-trained	0.026	−0.002
CirCor	stage2_e2e	Task-aware/impulse-trained	0.027	+0.000
CirCor	stage2_b_frozen	Task-aware/impulse-trained	0.027	−0.002
CirCor	dccrn_b	Task-aware/impulse-trained	0.032	−0.002
CirCor	cunet	Task-aware/impulse-trained	0.038	+0.002
CirCor	fcn	Reconstruction-only	0.060	−0.005
CirCor	spectral_subtraction	Classical control	0.182	+0.013
CirCor	bandpass	Classical control	0.344	−0.169
CirCor	median	Classical control	0.382	−0.030
CirCor	wiener	Classical control	0.539	−0.076
P16	identity	Identity control	0.000	+0.000
P16	median	Classical control	0.061	−0.014
P16	lunet	Reconstruction-only	0.143	−0.127
P16	unet	Reconstruction-only	0.148	−0.107
P16	fcn	Reconstruction-only	0.229	−0.155
P16	spectral_subtraction	Classical control	0.284	−0.009
P16	bandpass	Classical control	0.362	−0.113
P16	wiener	Classical control	0.412	−0.234
P16	stage2_frozen	Task-aware/impulse-trained	0.424	−0.136
P16	dccrn	Task-aware/impulse-trained	0.425	−0.134
P16	dccrn_b	Task-aware/impulse-trained	0.431	−0.116
P16	stage2_b_frozen	Task-aware/impulse-trained	0.431	−0.134
P16	dccrn_attn	Task-aware/impulse-trained	0.453	−0.141
P16	stage2_e2e	Task-aware/impulse-trained	0.460	−0.179
P16	cunet	Task-aware/impulse-trained	0.484	−0.102
BMD-HS	identity	Identity control	0.000	+0.000
BMD-HS	median	Classical control	0.004	+0.090
BMD-HS	unet	Reconstruction-only	0.249	+0.064
BMD-HS	wiener	Classical control	0.293	+0.007
BMD-HS	spectral_subtraction	Classical control	0.295	+0.060
BMD-HS	lunet	Reconstruction-only	0.342	−0.112
BMD-HS	fcn	Reconstruction-only	0.375	+0.033
BMD-HS	bandpass	Classical control	0.451	−0.056
BMD-HS	cunet	Task-aware/impulse-trained	0.897	−0.380
BMD-HS	dccrn_b	Task-aware/impulse-trained	0.902	−0.371
BMD-HS	stage2_frozen	Task-aware/impulse-trained	0.914	−0.256
BMD-HS	stage2_b_frozen	Task-aware/impulse-trained	0.914	−0.250
BMD-HS	dccrn	Task-aware/impulse-trained	0.916	−0.257
BMD-HS	stage2_e2e	Task-aware/impulse-trained	0.934	−0.354
BMD-HS	dccrn_attn	Task-aware/impulse-trained	0.944	−0.316

Table S7: Sensitivity of BMD-HS attenuation estimates to the cardiac-event mask. The nominal setting plus two alternatives for each of five parameters gives 11 one-factor-at-a-time settings. Entries are ranges of mean attenuation over 100 recordings for two evaluated task-aware representatives (cunet and dccrn_attn) and two reconstruction-only representatives (fcn and unet).

Denoiser	Cardiac-event regions	Elsewhere
cunet	0.860–0.914	0.898–0.942
dccrn_attn	0.924–0.954	0.943–0.967
fcn	0.062–0.097	0.364–0.569
unet	0.010–0.021	0.236–0.400

Table S8: Resampling and bandwidth-sensitivity control. No-added-noise recordings were passed through each resample-out/resample-back path without denoising. Confidence intervals use patients as the clustering unit for CirCor and BMD-HS and recordings for P16. Spectral correlation and energy ratio compare the resampled waveform with the input.

Pipeline	Corpus	Classifier	Spectral corr.	Energy ratio	ΔAUROC	95% CI
2000→1000→2000	BMD-HS	bilstm	1.0000	1.0019	−0.030	[−0.059, −0.004]
2000→1000→2000	CirCor	bilstm	0.8922	0.8000	−0.157	[−0.258, −0.055]
2000→1000→2000	P16	bilstm	0.9977	0.9972	−0.088	[−0.126, −0.054]
2000→1000→2000	BMD-HS	cnn2d	1.0000	1.0019	+0.006	[−0.001, +0.013]
2000→1000→2000	CirCor	cnn2d	0.8922	0.8000	−0.182	[−0.251, −0.115]
2000→1000→2000	P16	cnn2d	0.9977	0.9972	−0.113	[−0.153, −0.078]
2000→1000→2000	BMD-HS	cnn_transformer	1.0000	1.0019	−0.003	[−0.014, +0.009]
2000→1000→2000	CirCor	cnn_transformer	0.8922	0.8000	−0.284	[−0.359, −0.215]
2000→1000→2000	P16	cnn_transformer	0.9977	0.9972	−0.061	[−0.092, −0.033]
2000→1000→2000	BMD-HS	crnn	1.0000	1.0019	−0.000	[−0.016, +0.015]
2000→1000→2000	CirCor	crnn	0.8922	0.8000	−0.174	[−0.238, −0.106]
2000→1000→2000	P16	crnn	0.9977	0.9972	−0.063	[−0.094, −0.034]
2000→16000→2000	BMD-HS	bilstm	1.0000	1.0010	+0.000	[−0.001, +0.001]
2000→16000→2000	CirCor	bilstm	0.9995	0.9958	−0.001	[−0.003, +0.002]
2000→16000→2000	P16	bilstm	1.0000	1.0015	+0.000	[−0.001, +0.001]
2000→16000→2000	BMD-HS	cnn2d	1.0000	1.0010	+0.000	[−0.000, +0.000]
2000→16000→2000	CirCor	cnn2d	0.9995	0.9958	−0.000	[−0.001, +0.001]
2000→16000→2000	P16	cnn2d	1.0000	1.0015	+0.000	[+0.000, +0.000]
2000→16000→2000	BMD-HS	cnn_transformer	1.0000	1.0010	−0.000	[−0.001, +0.000]
2000→16000→2000	CirCor	cnn_transformer	0.9995	0.9958	+0.000	[−0.000, +0.001]
2000→16000→2000	P16	cnn_transformer	1.0000	1.0015	−0.000	[−0.000, +0.000]
2000→16000→2000	BMD-HS	crnn	1.0000	1.0010	−0.000	[−0.001, +0.000]
2000→16000→2000	CirCor	crnn	0.9995	0.9958	+0.000	[−0.001, +0.001]
2000→16000→2000	P16	crnn	1.0000	1.0015	+0.000	[−0.001, +0.001]

Table S9: Robustness of the corpus pattern across classifier architectures. For the artificially corrupted conditions, values are the mean change in AUROC with, in parentheses, the number of denoiser–condition comparisons out of 180 whose clustered 95% confidence interval lies entirely below zero. Clustering uses patients for CirCor and BMD-HS and recordings for P16. The last two columns give the mean effect on the original recordings.

	Artificial corruption: mean ΔAUROC (degraded/180)			Original recordings: mean ΔAUROC	
Classifier	CirCor	P16	BMD-HS	BMD-HS	PaHS
cnn2d	+0.014 (7)	−0.085 (134)	−0.081 (108)	−0.234	+0.006
crnn	+0.014 (1)	−0.048 (99)	−0.041 (64)	−0.079	+0.066
cnn_transformer	+0.009 (10)	−0.016 (61)	−0.059 (83)	−0.074	−0.042
bilstm	+0.026 (9)	−0.067 (111)	−0.031 (42)	−0.043	−0.016

Table S10: Criterion-based construction of the executable benchmark panel. The stages record the decision rules applied to the cited literature and archived experimental resources. They are not a PRISMA flow diagram or an exhaustive literature search: a complete search log and candidate-count ledger were not retained, so no identification or exclusion counts are claimed. Performance in the present benchmark was not used to add or remove a resource.

Stage	Required evidence	Excluded at this stage	Function in the retained panel
Candidate identification	A human-PCG dataset, waveform-output denoising approach, recorded-noise source, specified generated perturbation or diagnostic architecture described in the cited literature or traceable project archive	Resources outside the bounded literature-and-archive frame	Defines the bounded candidate frame used for this empirical benchmark.
Scientific eligibility	Human PCG waveforms, stable identifiers, mappable labels and grouping for datasets; waveform output for denoisers; a traceable source or specified generator for noise; label-compatible scores for classifiers	Reviews, classification-only denoisers, non-PCG modalities, inaccessible waveforms, untraceable perturbations or labels requiring new clinical adjudication	Ensures that every retained resource can answer the same paired diagnostic-preservation question.
Executable eligibility	Deterministic conversion or generation at the 2-kHz, 5-s interface; reproducible cohorts; strongest supported training–evaluation separation; runnable frozen code and weights or a traceable harmonised implementation	Assets that could not be executed, harmonised or separated from evaluation with the retained evidence	Prevents comparison of paper-reported headline values obtained from incompatible populations and pipelines.
Complementary role assignment	Complementary group-level coverage of source relationship, acquisition heterogeneity, disease scope, grouping information, original-recording availability, architecture, objective and perturbation exposure	Eligible resources that added no intended coverage, or required a new experiment outside the executed benchmark	Assigns CirCor, P16, BMD-HS and PaHS distinct data roles and spans the evaluated denoiser and classifier design axes without claiming universal coverage.

Table S11: Training configuration of the ten evaluated learned denoisers. All learned denoisers used the same patient-disjoint development partition, environmental and respiratory noise at −5 to +15 dB, 20% no-added-noise samples, 5 s clips at 2 kHz, batch size 16 and gradient clipping at 5.0. Task-aware and impulse-trained systems retained the checkpoint with the lowest specified total validation objective. In contrast, the fcn, unet and lunet baselines used AdamW for 50 epochs at 2×10^{-4} and retained the checkpoint with the highest mean validation SI-SDR. For the seven task-aware and impulse-trained systems, epochs, batch size, learning rates, seed, mixture range and split fractions were read from checkpoint-embedded `train_args`; optimiser, clipping, objectives and retention rules were verified from matching trainer control flow. The fcn, unet and lunet values are matched archived-trainer defaults; their checkpoints did not store run-level command-line arguments, so an unrecorded override cannot be independently excluded for those three baselines.

Model	Family	Total parameters (M)	Epochs	Learning rate
cunet	Task-aware	2.99	50	2×10^{-4}
dccrn	Task-aware	5.49	50	2×10^{-4}
dccrn_attn	Task-aware	6.15	50	2×10^{-4}
stage2_e2e	Task-aware	5.64	40	2×10^{-4} refiner / 2×10^{-5} front end
stage2_frozen	Task-aware	5.64	40	2×10^{-4} refiner / front end frozen
dccrn_b	Impulse-trained	5.49	50	2×10^{-4}
stage2_b_frozen	Impulse-trained	5.64	40	2×10^{-4} refiner / front end frozen
fcn	Adapted reconstruction baseline	0.09	50	2×10^{-4}
unet	Adapted reconstruction baseline	0.10	50	2×10^{-4}
lunet	Adapted reconstruction baseline	0.20	50	2×10^{-4}

Table S12: Diagnostic-classifier configuration and provenance. Checkpoint-verified entries were read from retained model or training-split artefacts. Source-linked entries were recovered from the production trainer whose output path, filename pattern and seven-field checkpoint schema matched all eight retained checkpoints. The checkpoint schema did not redundantly store the optimisation hyperparameters or run-level command-line arguments. Source-linked values are therefore reported as production-source defaults rather than independently checkpoint-verified run settings.

Item	Setting	Evidence
Architectures and trainable parameters	cnn2d: 355,650; bilstm: 594,434; crnn: 681,442; cnn_transformer: 354,786	Executable definitions and checkpoint-compatible state topology
Input representation	5 s at 2 kHz; 256-sample Fourier transform; 64-sample hop; 64 mel bands over 0–800 Hz; 157 frames; recording-wise standardisation	Production source; 64 mel bands and 157 frames also checkpoint-verified
Training cohorts	CirCor: 1,818 recordings (343 positive, 1,475 negative) from 525 patients; P16: 2,940 recordings (515 abnormal, 2,425 normal)	Checkpoint metadata, training-split records and separation report
Optimisation	30 epochs; batch size 32; Adam; learning rate 10^{-3}; cross-entropy; run-initialisation seed 42	Matched production-source defaults; run-level command arguments not retained
Class balancing	Inverse-class-frequency weighted sampling with replacement	Production-source control flow
Validation and retention	No internal validation split, scheduler or early stopping; state after the final configured epoch retained	Production-source control flow
Evaluation isolation	CirCor evaluation patients and P16 evaluation identifiers excluded before training; no denoised input used for training	Split and separation artefacts plus production source

Supplementary Data 1

The accompanying `Supplementary_Data_1_all_effects.csv` contains 2,360 paired comparisons. For each classifier, CirCor and P16 each contribute 190 comparisons (ten denoisers times 18 artificially corrupted conditions plus the no-added-noise condition), BMD-HS contributes 200 (the same 190 plus ten original-recording comparisons), and PaHS contributes ten original-recording comparisons, for 590 per classifier and 2,360 across four classifiers. Each row reports corpus, condition, denoiser, classifier, paired ΔAUROC, clustered 95% interval, finite-corrected bootstrap p value, Benjamini–Hochberg-adjusted q value, and numbers of clusters and recordings.

Supplementary Data 2

The separate `Supplementary_Data_2_manifests.zip` contains the consolidated 1,500-row arm-level manifest, source metadata, five subset-selection records, consistency checks, available recording- and patient-separation reports, the Hospital Ambient Noise provenance record and a programmatic check of reported signal-analysis parameters. It documents 1,333 unique recordings and the intentional 167-record overlap between the two BMD-HS arms without redistributing source or derived audio.