Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Eashita Wazed
Chonnam National University
Choonsung Shin
Chonnam National University
Hieyong Jeong
h.jeong@jnu.ac.kr

Chonnam National University

Article

Keywords: phonocardiography, denoising, distribution shift, diagnostic evaluation, diagnostic preservation, cross-corpus benchmark

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11005871/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Eashita Wazed[1], Choonsung Shin[2], Hieyong Jeong[1,*]

[1]Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Republic of Korea

[2]Graduate School of Culture, Chonnam National University, Gwangju, Republic of Korea

[*]Correspondence to Hieyong Jeong (h.jeong@jnu.ac.kr)

Abstract

Heart-sound denoisers are often judged by reconstruction fidelity under artificial noise, leaving diagnostic preservation after transfer insufficiently characterised. We built a multi-corpus benchmark rather than proposing a new architecture. Criterion-based panel construction defined an executable panel of ten learned denoisers trained on a common source corpus, four classical filters and an identity control. We evaluated five 300-recording arms from four phonocardiogram corpora (1,333 unique recordings). Paired comparisons covered 18 artificial-corruption conditions and their no-added-noise passthroughs in three arms, plus two original-recording arms without paired clean references, using four fixed classifiers. Uncertainty used patient clustering where identifiers were available, recording clustering otherwise, and a parent-recording sensitivity analysis for PaHS. Mean AUROC change was +0.014 on the denoiser-development corpus but -0.085 and -0.081 on two external corpora. On external no-added-noise inputs, some learned denoisers reduced AUROC by up to 0.380 and attenuated approximately 91% of magnitude within detected cardiac-event regions on BMD-HS. On original BMD-HS recordings, the largest AUROC reduction was 0.387, with a 0.767 sensitivity loss at the pre-denoising threshold. Development-distribution benefit therefore did not establish diagnostic preservation after transfer. Target-corpus task effects, no-added-noise behaviour and fixed operating points should be evaluated before integration into a diagnostic pipeline.

Keywords: phonocardiography; denoising; distribution shift; diagnostic evaluation; diagnostic preservation; cross-corpus benchmark

Introduction

Cardiac auscultation is inexpensive and widely accessible, but its interpretation depends on training and experience. Automated analysis of phonocardiograms (PCGs) is consequently being developed for murmur detection and cardiovascular screening [1, 2]. Recordings acquired outside controlled laboratories contain respiratory sounds, stethoscope friction, patient motion, environmental interference and device noise. A denoiser is therefore often placed before the diagnostic model.

Most heart-sound denoisers are evaluated on artificially corrupted recordings for which a clean reference is available. Signal-to-noise improvement, waveform similarity and scale-invariant signal-to-distortion ratio (SI-SDR) then measure whether the output approaches that reference [3, 4]. These outcomes quantify reconstruction, but not necessarily diagnostic preservation. A transformation can improve an acoustic average while suppressing a short transient, relative spectral pattern or score-calibration cue used by a classifier. It can also preserve most score rankings while shifting clinically relevant cases across a previously calibrated threshold. Acoustic fidelity, ranking stability and operating-point stability are therefore related but distinct properties.

This distinction becomes more important under distribution shift. Patient groups, diseases, stethoscopes, sampling systems, recording sites and ambient conditions vary across PCG corpora.

Medical classifiers can exploit dataset-specific structure and fail after such changes [5–11]; a learned preprocessor is exposed to the same risk. Fixed denoiser weights may apply a much stronger transformation to an unfamiliar corpus or treat legitimate cardiac structure as interference. This behaviour remains hidden when validation is confined to the development corpus or to reconstruction metrics calculated from chosen artificial mixtures.

Published PCG work has established that noise affects classification and that time–frequency U-Nets, recurrent and attention-based networks, and adaptive filters can reduce particular interferences [12–16]. However, individual studies differ in their patients, labels, recording devices, noise sources, signal interfaces, objectives and downstream models. Reported reconstruction gains therefore do not reveal whether contemporary denoising strategies have reached a level at which diagnostic behaviour transfers reliably. Nor can separate leaderboards determine whether an observed gain comes from general noise suppression, source-corpus adaptation or compatibility with one classifier.

A common, multi-corpus benchmark is needed for three reasons. First, it makes the denoiser rather than the recording or diagnostic model the changing component in each paired comparison. Second, it exposes the capability boundary of evaluated design families under the same target data and outcomes. Third, it establishes which evidence is missing before a denoiser can be inserted into a clinical pipeline. This is a validation question rather than a model-development contest: the central issue is not which method reconstructs one benchmark best, but whether favourable evidence survives changes in corpus, noise, classifier and operating point. Such system-level evidence is necessary before claims of clinical readiness [17, 18]. Here, *diagnostic preservation* means preservation of a fixed downstream model's discrimination and decisions at a threshold specified before preprocessing. It is not a direct estimate of patient harm or human-reader performance.

We therefore conducted an empirical benchmark rather than proposing a new architecture. Resource construction followed an explicit four-stage sequence: candidate identification within the cited literature and archived executable resources available to the study, scientific eligibility, executable eligibility and complementary role assignment. Four public PCG corpora were assigned complementary source, heterogeneous-transfer, multi-disease-transfer and original-recording roles. The executable panel comprised three harmonised local architectural baselines adapted from published designs and seven study-developed variants trained on a common CirCor source partition, together with four classical filters and an identity control. Four frozen diagnostic architectures tested whether effect direction depended on one downstream inductive bias. Selection did not use reported performance, citation count or the results observed in this benchmark. Resource eligibility, execution feasibility, model provenance and benchmark role were recorded separately so that the ten implementations are not misrepresented as ten official published checkpoints. The panel was coverage-oriented and deliberately bounded; it was not an exhaustive census of every PCG dataset or denoiser published worldwide. The criterion-based construction and complementary coverage of the retained panel are summarised in Fig. 1.

The study asks whether development-distribution benefit survives cross-corpus transfer, whether a denoiser changes fixed-classifier discrimination or threshold decisions when no artificial noise is added, and whether failures on original recordings can be bounded by signal-level and statistical controls. We combine paired changes in AUROC and fixed-threshold decisions with reference-based fidelity, reference-free processing intensity and attenuation within detected cardiac-event regions. The contribution is a controlled account of the capability boundary of the evaluated denoisers and a minimum evaluation logic for separating acoustic enhancement from diagnostic preservation. It does not assume that denoising is intrinsically harmful or that one corpus can rank methods for universal deployment.

Results

Favourable development-distribution effects did not transfer consistently

The benchmark comprised five class-balanced 300-recording arms drawn from four public PCG corpora, representing 1,333 unique recordings because the two BMD-HS arms shared 167 records (Table 1). Ten learned denoisers, four classical filters and an identity control were paired with four diagnostic classifier architectures (Table 2). CirCor, P16 and BMD-HS each contributed 18 artificial-corruption conditions and a no-added-noise condition; BMD-HS and the PASCAL Classifying Heart Sounds Challenge (PaHS) additionally contributed original-recording arms. The primary classifier therefore yielded 590 learned-denoiser comparisons, and repetition with four classifiers yielded 2,360. The analytical conditions and the evidence used to test them are summarised in Table 3.

Across artificial-corruption conditions, mean paired ΔAUROC for the primary classifier was +0.014 on CirCor, the denoiser-development corpus, but -0.085 on P16 and -0.081 on BMD-HS (Fig. 2; Supplementary Table S1). Clustered 95% intervals lay below zero in 7/180 CirCor comparisons, 134/180 P16 comparisons and 108/180 BMD-HS comparisons; intervals lay above zero in 35, 0 and 14 comparisons, respectively. These are descriptive interval-exclusion counts, not multiplicity-adjusted discoveries. After Benjamini–Hochberg adjustment across the analysis-wide family of 2,360 comparisons, the primary-classifier artificial- corruption results included 3, 122 and 96 negative effects with $q < 0.05$ on CirCor, P16 and BMD-HS, respectively; the corresponding positive counts were 18, 0 and 11.

The reversal was not confined to one training objective. Task-aware denoisers changed from a mean of +0.023 on CirCor to -0.087 on P16 and -0.111 on BMD-HS; impulse-trained variants changed from +0.022 to -0.079 and -0.133. Reconstruction-only models had smaller mean effects on CirCor and BMD-HS (-0.006 and +0.004) but still averaged -0.084 on P16. The condition-level heat map shows that the corpus pattern extended across the six evaluated interference types (Supplementary Fig. S1). Exposure to a noise class during training did not guarantee task-preserving transfer, and training exposure alone did not account for the observed pattern.

No-added-noise and original-recording tests falsified strict preservation

Applying each denoiser without adding artificial noise isolated the consequence of inserting the fixed transformation. On CirCor, nine of ten denoisers changed AUROC by between -0.005 and +0.002, while lunet changed it by -0.016 (Fig. 3; Supplementary Table S2). On P16, all ten intervals lay below zero, with effects from -0.102 to -0.179. On BMD-HS, all seven task-aware or impulse-trained models had intervals below zero, with changes from -0.250 to -0.380; lunet also reduced AUROC by 0.112, whereas fcn and unet changed it by +0.033 and +0.064 with intervals compatible with no change. All ten P16 effects and eight of ten BMD-HS effects remained negative at $q < 0.05$ after the global multiplicity adjustment. These recordings were not necessarily acoustically clean: the test means only that no additional corruption was imposed. Failure nevertheless demonstrates that the inserted module was not task neutral on that target corpus.

The original-recording arms removed dependence on an artificial noise model. On BMD-HS, every task-aware and impulse-trained denoiser reduced AUROC (Fig. 4). The largest change was -0.387 (95% CI -0.475 to -0.299; globally adjusted $q = 0.0052$) for cunet, reducing AUROC from 0.77 to 0.38; the other task-aware and impulse-trained models ranged from -0.276 to -0.370, and lunet changed AUROC by -0.127. In contrast, PaHS effects ranged from -0.005 to +0.028 and had no consistent direction. Thus, the result is corpus-dependent rather than a universal effect of denoising.

The rank formulation gives these changes a direct interpretation. Each balanced evaluation arm contained 22,500 positive–negative pairs. The original BMD-HS effect of -0.387 therefore represents approximately 8,708 more pair-order losses than gains, and the no-added-noise effect of -0.380 represents approximately 8,550. At the threshold fixed before denoising, task-aware and impulse-trained BMD-HS models reduced sensitivity by 0.733–0.767, increased specificity by 0.287–0.307 and reduced balanced accuracy by 0.223–0.230 (Supplementary Table S3). The largest sensitivity change corresponds to approximately 115 fewer true-positive decisions among 150 positives. On PaHS, AUROC could remain nearly unchanged while fixed-threshold balanced accuracy fell by as much as 0.203. Held-out-threshold analysis preserved these qualitative shifts, whereas condition-specific re-optimisation provides only a descriptive best case (Supplementary Table S4). These observations falsify strict preservation for the affected corpus–model combinations; they do not certify combinations whose intervals span zero because no clinically justified non-inferiority margin was specified prospectively.

Transfer failure coincided with distribution-dependent and event-aligned attenuation

When an uncorrupted reference could be constructed, change in SI-SDR and ΔAUROC were positively correlated across the 540 learned-denoiser–condition pairs ($r = 0.41$–0.60 by corpus; Supplementary Fig. S2). Mean SI-SDR change and ΔAUROC were +3.5 dB and +0.014 on CirCor, -6.2 dB and -0.085 on P16, and -17.2 dB and -0.081 on BMD-HS. A sign-based cross-tabulation placed 97/180 CirCor pairs in the fidelity-improved/AUROC-improved quadrant, whereas 165/180 P16 and 154/180 BMD-HS pairs were in the fidelity-degraded/AUROC-degraded quadrant (Supplementary Table S5). Of 249 pairs whose clustered diagnostic interval lay below zero, only two improved SI-SDR. Fidelity was therefore informative in paired artificial-corruption experiments, but it was unavailable for the original recordings without paired clean references on which the largest loss occurred.

Reference-free processing intensity characterised the corpus-dependent transformation that accompanied different behaviour after transfer. Across the ten learned denoisers, four classical filters and identity (15 methods), whole-recording positive magnitude attenuation correlated negatively with no-added-noise ΔAUROC: $r = -0.69$ on CirCor, -0.63 on P16 and -0.93 on BMD-HS (Fig. 5; Supplementary Table S6). The seven task-aware and impulse-trained denoisers attenuated 2–4% of magnitude on CirCor, 42–48% on P16 and 90–94% on BMD-HS despite unchanged model weights. Because methods differed in architecture and objective, this association is a warning signal rather than a causal dose–response or a universal threshold.

Event localisation on the no-added-noise inputs showed that this severity was not confined to intervals outside cardiac events. Task-aware and impulse-trained models attenuated 1.0% of magnitude within detected event regions and 2.9% elsewhere on CirCor, 32.8% and 45.3% on P16, and 90.8% and 91.8% on BMD-HS (Fig. 6). Reconstruction-only models on BMD-HS were more selective, attenuating 4.6% within event regions and 34.3% elsewhere. A rule-selected illustrative case is shown in Supplementary Fig. S3. In this recording, the correct-class margin changed sign, corresponding to a change from correct to incorrect classification. The display localises the transformation but neither identifies the attenuated component nor establishes that the attenuation caused the classification change. Across 11 mask settings, mean event-region attenuation changed by at most 0.054; among the four examined denoisers, the ranges for two task-aware representatives and two reconstruction-only representatives remained separated (Supplementary Table S7). The transformation residual thus contained substantial energy aligned with detected cardiac-event

intervals, but localisation alone cannot identify that energy as causal disease information.

Controls and robustness analyses bounded alternative explanations

The identity transformation changed no result, whereas classical preprocessing was not universally safe. The 20–500 Hz band-pass had an interval below zero in 18/19 CirCor conditions, and the Wiener filter did so in 18/19 P16 conditions; the largest single-condition loss was 0.307 AUROC. On BMD-HS, however, three of four classical filters produced no interval below zero across the 19 conditions, unlike the severe no-added-noise losses of the task-aware and impulse-trained families. Diagnostic perturbation was therefore not exclusive to learned preprocessing, while those evaluated families showed unusually strong distribution dependence.

Denoiser-free controls did not reproduce the largest BMD-HS effect. The 2 kHz→16 kHz→2 kHz path was effectively inert across all corpora and classifiers (Supplementary Fig. S4; Supplementary Table S8). Removing content above 500 Hz through a 1 kHz round trip changed CirCor by -0.157 to -0.284 and P16 by -0.061 to -0.113, but BMD-HS by only -0.030 to +0.006. Task-aware and impulse-trained denoisers reduced BMD-HS root-mean-square amplitude by a median factor of 18, yet restoring each output to its original amplitude changed probabilities by at most 2×10^{-4} over 9,000 scores and changed no reported AUROC to three decimals. Resampling and global amplitude collapse were therefore insufficient explanations for the largest loss.

All four classifiers showed positive mean artificial-corruption effects on CirCor (+0.009 to +0.026) and negative means on P16 and BMD-HS; original BMD-HS means were also negative for all four (-0.043 to -0.234), whereas PaHS signs differed (-0.042 to +0.066; Supplementary Table S9). Effect size remained classifier-dependent, but the development-versus-external direction was not produced by one downstream architecture. Patient-clustered intervals were a median 1.62 times wider than instance-level intervals, and 276/590 primary-classifier comparisons still had intervals wholly below zero after clustering (Supplementary Fig. S4).

Finally, the repeated-cross-validation linear probe reached 0.800 ± 0.064 AUROC (mean $\pm$ standard deviation across 25 held-out folds) on original P16 recordings and declined to 0.732 after cunet and 0.700 after dccrn_attn, paired changes of -0.069 and -0.100. Changes after unet and fcn were only -0.005 and -0.006 despite declines in the frozen classifier. The probe therefore distinguishes some representation mismatch from reduced linear separability in one feature space, but it does not establish irreversible information loss. Its BMD-HS baseline was 0.521, precluding an equivalent claim there.

Discussion

This benchmark defines a capability boundary for the evaluated panel of heart-sound denoisers. The models could improve reconstruction and downstream performance on their development distribution, yet the same fixed transformations frequently failed to preserve classifier discrimination and fixed-threshold decisions after transfer. The finding is not that one architecture is defective or that denoising is intrinsically harmful. It is that source-corpus reconstruction or task performance is not portable evidence of downstream diagnostic preservation. The evaluated denoisers could be effective signal enhancers in matched settings while remaining insufficiently validated as components of a transferred diagnostic system.

The principal academic value lies in replacing disconnected model reports with paired, system-

level evidence. The recording, label, cohort and classifier weights were held fixed in each comparison, so the measured difference is attributable to inserting the preprocessor within that pipeline rather than to unrelated leaderboard populations. A common interface also allowed architectures, objectives and perturbation exposure to be examined under the same corpus shifts. Replication across four classifier inductive biases showed that the development-to-external reversal was not manufactured by one downstream architecture, although effect size and the PaHS direction remained classifier-dependent. The criterion-based resource framework further separates scientific eligibility, executable eligibility, benchmark role and provenance. It distinguishes three local architectural baselines adapted from published designs and seven study-developed variants, and prevents architectural resemblance from being misreported as evaluation of an original-author checkpoint. The result is a controlled assessment of a deliberately heterogeneous executable panel, not a census, prevalence estimate or universal ranking of all published denoisers.

The strength of the inference comes from convergence across tests with different failure modes. Development-distribution benefit established that the pipeline could detect favourable effects; external reversal tested transfer with weights fixed; no-added-noise and original-recording arms removed dependence on a chosen corruption; AUROC and pre-denoising thresholds tested ranking and decision stability; event localisation characterised where attenuation occurred; and resampling, amplitude restoration, patient-aware uncertainty and four classifiers challenged narrower technical explanations. No single observation proves a biological mechanism, but their agreement supports the narrower conclusion that development-distribution reconstruction evidence was insufficient to establish diagnostic preservation in the evaluated transfer settings.

The study consequently changes the relevant unit of validation. Once denoiser output becomes diagnostic input, diagnostic preservation is a relational property of the denoiser, target corpus, classifier and operating point. With unchanged weights, the task-aware and impulse-trained models attenuated only 2–4% of positive short-time magnitude on CirCor no-added-noise inputs but 90–94% on BMD-HS. An evaluation restricted to the source distribution would miss this interaction. The same observation explains why a binary distinction between trained and untrained noise is inadequate. Exposure to environmental, respiratory or impulse perturbations did not guarantee task-preserving behaviour on a new corpus. Noise-class coverage and cross-corpus preservation are separate generalisation problems.

No single metric captured this boundary. SI-SDR tracked diagnostic change in paired artificial mixtures and rarely improved when diagnostic performance clearly declined, but it was unavailable for original recordings without artificial corruption. The no-added-noise passthrough supplied a reference-free falsification test: failure shows that inserting the module is not task neutral, whereas passing cannot certify safety in finite, incomplete data. Processing intensity and event-aligned localisation showed where strong transformations occurred, without identifying the residual as causal disease information. AUROC and fixed-threshold results also answered different deployment questions; PaHS demonstrated that nearly stable ranking can coexist with a substantial operating-point loss after score shift. Classical filters harmed some corpus–condition pairs as well. Within the evaluated matrix, diagnostic perturbation was therefore not unique to deep learning. What distinguished the evaluated task-aware and impulse-trained families was their extreme, distribution-dependent transformation severity.

Together, these observations yield a minimum evidence hierarchy for future denoiser studies. Investigators should compare each processed waveform with its unprocessed counterpart on the intended target corpus; include a no-added-noise passthrough; combine reference-based fidelity, when available, with reference-free severity and cardiac-event-aligned localisation; evaluate both

ranking and an operating point fixed before preprocessing; preserve patient or recording groups when estimating uncertainty; and prospectively specify a clinically justified non-inferiority margin. Recalibration can then be evaluated separately, but condition-specific threshold optimisation should not be treated as evidence that the original pipeline was preserved. This hierarchy converts denoising from an isolated reconstruction claim into a falsifiable component- preservation claim and provides a common reporting target for the field.

The framework also motivates a future translational hypothesis about heavily processed heart sounds. Algorithmically denoised audio should not be equated with naturally clean auscultation: spectral suppression, envelope smoothing and transient attenuation can create an unfamiliar acoustic distribution. Clinicians accustomed to realistic background noise may therefore perceive aggressively processed sounds as unnatural or machine-like. Human listeners were not evaluated here, so this is a clinical hypothesis rather than a study finding. A blinded reader study comparing original, artificially corrupted, algorithmically denoised and naturally clean recordings should measure diagnostic accuracy, confidence, naturalness and perceived artefact severity.

The scope of the empirical claim remains limited. The coverage-oriented panel contained four public corpora and ten harmonised CirCor-trained implementations rather than every released PCG dataset, publication or official checkpoint; diffusion and newer transformer–recurrent denoisers were not executed. The documented selection stages cannot support claims of worldwide coverage because exhaustive retrieval and a complete candidate-count ledger were outside this empirical benchmark. Each arm was a class-balanced sample of 300 recordings rather than a prevalence-representative full corpus. The independently sampled BMD-HS arms shared 167 records but were not contrasted inferentially. CirCor used murmur labels, whereas the external corpora used normal/abnormal labels, so population, device, acquisition and label shifts cannot be separated. P16 and PaHS lacked patient identifiers; recording-level clustering was used for P16 and the primary PaHS analysis, and parent-recording clustering was added for PaHS. Each denoiser and classifier architecture was represented by one retained training run, so replication across classifier architectures does not quantify optimisation-seed variability. The analysis also conditions on one deterministic 300-recording subset per arm; bootstrap intervals do not capture variation across alternative eligible subsets. These constraints make the results a capability-boundary demonstration within the evaluated matrix rather than an estimate of how often all denoisers fail.

Several mechanistic and translational questions also remain. Architecture, capacity, objective and perturbation exposure varied together, precluding causal attribution to one design choice. The envelope-derived cardiac mask was not a manually validated S1/S2 or murmur annotation; its sensitivity analysis established robust localisation but not that the attenuated component caused diagnostic loss. The linear probe assessed only one representation and cannot establish irreversible information loss. Original recordings without artificial corruption lacked a paired clean reference, no prospective deployment or listener experiment was performed, and no clinical non-inferiority margin was specified before analysis. Matched full-scale retraining, selective restoration of cardiac- and non-cardiac-region residuals, expert mask validation and prospective target-specific reader studies are therefore needed to distinguish recoverable representation mismatch from information loss that persists after adaptation.

In summary, within a bounded but deliberately heterogeneous multi-corpus benchmark, several evaluated heart-sound denoisers failed to preserve fixed-classifier discrimination and threshold decisions after transfer, including when no artificial noise was added. The largest effect reduced AUROC from 0.77 to 0.38 in the original-recording BMD-HS arm. In the separate BMD-HS no-added-noise arm, task-aware and impulse-trained denoisers attenuated approximately 91% of positive

magnitude within detected cardiac-event regions. These findings do not imply that every denoiser is harmful or that every attenuated component was causal disease information. They establish that development-distribution fidelity or task benefit is insufficient evidence of diagnostic preservation on a new corpus. Target-corpus task evaluation, no-added-noise testing, fixed-operating-point analysis and group-aware uncertainty should be treated as minimum evidence before a denoiser is inserted into a heart-sound diagnostic pipeline.

Methods

Criterion-based construction and provenance of the benchmark panel

This retrospective study is an empirical multi-corpus benchmark rather than a systematic review, meta-analysis or exhaustive census of published PCG resources. Panel construction used four documented stages: candidate identification within the literature and archived experimental resources available to the study, scientific eligibility, executable eligibility and complementary role assignment. This criterion-based structure makes the evaluated panel traceable but does not imply exhaustive literature retrieval, protocol registration or dual-reviewer screening. The eligibility criteria, role-assignment logic and provenance boundaries are reported in the Supplementary Information section "Benchmark resource and training provenance" and Supplementary Table S10.

Figure 1 uses a PRISMA-inspired reporting structure to separate identification, classification, eligibility and inclusion, with explicit side branches for exclusion reasons. It is a benchmark-selection diagram rather than a systematic-review PRISMA flowchart: because a complete retrospective candidate-count ledger was not retained, counts are shown only for the verifiable executed panel and no candidate or exclusion totals were reconstructed. Selection operated at the executable-resource level. Each candidate first had to satisfy the mandatory criteria for its resource class; among eligible assets, retention then sought complementary group-level coverage. The same resulting evaluation matrix was used throughout the reported comparative analyses, and no resource was substituted or removed because of its observed diagnostic or fidelity outcome.

Four successive stages were applied. First, literature-guided identification drew on the PCG denoising and dataset reports cited in the Introduction, their cited architectural precedents and executable resources traceable in the project archive to assemble candidate human PCG datasets, waveform-output denoising approaches, recorded-noise sources, reproducible generated perturbation mechanisms and downstream diagnostic architectures. Second, scientific eligibility required human waveform-level PCGs, stable recording identifiers, a diagnostic label mappable without new clinical adjudication and at least a recording-level grouping unit. Denoisers had to return a waveform that could be inserted before a fixed diagnostic model. Reviews, classification-only methods, non-PCG modalities and resources without accessible waveforms were excluded from the experimental panel. Third, executable eligibility required a reproducible cohort, deterministic conversion to a common 2 kHz, 5 s interface, verifiable training–evaluation independence at the highest level supported by released metadata, and either frozen executable code and weights or sufficient specification for a traceable harmonised implementation. Publication venue, citation count and reported performance were not criteria, and the panel was not expanded or pruned according to the benchmark outcomes. Fourth, among eligible resources, role assignment sought non-redundant coverage of development relationship, disease scope, patient grouping, acquisition heterogeneity and the availability of original recordings without a paired clean reference.

These gates assigned CirCor to a patient-disjoint source-distribution test, P16 to a heterogeneous

external test and classifier source, BMD-HS to patient-grouped multi-disease transfer, and PaHS to an independent original-recording challenge. They were not pooled as interchangeable disease samples. The noise panel similarly covered recorded environmental, physiological and hospital interference together with generated swept-frequency, stationary narrowband and impulsive mechanisms at three severities. This role-based construction tests transfer across deliberately different conditions rather than estimating population prevalence.

For recorded interference, eligibility required a publicly traceable source and deterministic conversion to the common waveform interface; generated interference required a fully specified, seeded generator. Retention then required a distinct mechanism or training-exposure role. The four diagnostic probes had to emit label-compatible scores from the common input representation, have traceable frozen weights, and remain independent of denoised training data. The retained CNN2D, BiLSTM, CRNN and CNN–Transformer structures therefore provided convolutional, recurrent, hybrid and attention-based outcome probes without changing classifier weights between preprocessing conditions.

Model provenance was recorded separately from design family. The fcn, unet and lunet comparators were local architectural baselines adapted from published designs and trained under the common protocol; they were not original-author checkpoints. The remaining seven learned systems were study-developed variants assembled from cited complex, recurrent, attention and two-stage components. Collectively, the panel varied waveform versus complex time–frequency processing, convolutional versus recurrent or attention context, reconstruction-only versus task-consistency objectives, single- versus two-stage processing and standard versus impulse-matched exposure. Four deterministic filters and identity served as mechanistic controls. Four fixed diagnostic classifiers varied downstream inductive bias and were outcome probes, not denoisers. The resulting matrix is deliberately heterogeneous but bounded: it tests whether findings persist across the evaluated executable panel, not which model is universally best. Diffusion and newer transformer–recurrent PCG denoisers, among other resources, would require additional experiments and are outside the present inference.

Benchmark design and diagnostic-preservation estimands

The unit of analysis was the change produced by inserting preprocessing into a fixed diagnostic pipeline. For recording i, the observed waveform was represented conceptually as $x_i = s_i + n_i$, where s_i is patient-derived acoustic content and n_i extrinsic noise. Neither component is separately observed in original recordings without artificial corruption. Denoiser d produced $\widehat{x}_{id} = D_d(x_i)$, and $r_{id} = x_i - \widehat{x}_{id}$ was termed the *transformation residual*; it was not labelled removed noise because it can contain attenuated patient-derived content, attenuated noise and transformation artefacts.

With fixed classifier g, original and transformed scores were $q_i = g(x_i)$ and $\widehat{q}_{id} = g(\widehat{x}_{id})$. For corpus or condition c, the primary estimand was

$$\Delta U_{d,c} = \mathrm{AUROC}_c(\widehat{q}, y) - \mathrm{AUROC}_c(q, y). \tag{1}$$

Empirical AUROC is the mean concordance of all positive–negative score pairs. If $e_{id} = \widehat{q}_{id} - q_i$ and $\gamma_{ij} = q_i - q_j$, then

$$\Delta U_{d,c} = \frac{1}{N_+ N_-} \sum_{i:y_i=1} \sum_{j:y_j=0} \{\phi(\gamma_{ij} + e_{id} - e_{jd}) - \phi(\gamma_{ij})\}, \tag{2}$$

where $\phi(z)$ equals 1, 1/2 or 0 for positive, tied or negative differences. A negative ΔU is therefore the net fraction of pair orderings broken rather than repaired by preprocessing. At fixed threshold

t, a decision can change only when the score perturbation reaches the original threshold margin. Formal rank-preservation, threshold-preservation and residual non-identifiability statements are proved in the Supplementary Information section "Mathematical verification of the preservation conditions."

Diagnostic preservation would ideally be assessed as non-inferiority, $\Delta U_{d,c} \geq -\delta_U$, using a clinically justified margin fixed in advance. No such margin was available for this retrospective benchmark. We therefore used zero only as a strict reference for detecting degradation: an interval wholly below zero falsifies strict preservation, whereas an interval spanning zero does not establish equivalence or safety.

Data resources, cohort construction and noise conditions

Four public PCG corpora were used [19–22]. CirCor supplied multi-location paediatric recordings, patient identifiers and murmur annotations. It was the denoiser-development distribution and supported patient-disjoint murmur evaluation. Records labelled Murmur=Unknown were excluded; each site recording inherited its patient-level Present or Absent label. PhysioNet/CinC 2016 (P16) supplied a heterogeneous normal/abnormal challenge benchmark and the training source for the abnormality classifiers. Its `REFERENCE.csv` labels were mapped as -1 to normal and $+1$ to abnormal, retaining the first occurrence of duplicate record names across released listings. BMD-HS supplied multi-disease recordings with repeated locations and patient identifiers. A record was normal when the released `N` indicator equalled 1 and abnormal otherwise. The PASCAL Classifying Heart Sounds Challenge corpus (PaHS) supplied an independent original-recording challenge; its directory labels mapped `murmur` to abnormal and `normal` to normal. CirCor therefore has a narrower murmur task than the normal/abnormal external corpora, and cross-corpus comparisons were interpreted as pipeline-transfer tests rather than interchangeable estimates of one disease prevalence.

Five fixed, class-balanced arms of 300 recordings were constructed and reused within every applicable method–condition–classifier comparison (Table 1). Equal arm size and class balance standardised the number of scored recordings and positive–negative pairs under a common computational budget; they did not equalise effective sample size, and the remaining grouping differences were handled by cluster-aware uncertainty. These arms were not intended to reproduce clinical prevalence. Records were partitioned by label, shuffled with a corpus-specific deterministic seed derived from base seed 42, sampled equally by class and then reshuffled. CirCor contributed 300 records from 197 patients, P16 300 records without public patient identifiers, and the artificial-corruption BMD-HS arm 300 records from 91 patients. The original-recording BMD-HS and PaHS arms contained 300 records from 97 patients and 300 clips from 224 parent recordings, respectively. The two BMD-HS arms were drawn independently and shared 167 record identifiers; this overlap was disclosed but no direct cross-arm inferential contrast was made. Across all arms there were 1,333 unique recordings. All waveforms were resampled to 2 kHz and truncated or zero-padded to 5 s.

CirCor evaluation records were drawn from a pool of 377 patients held out from all denoiser training. The CirCor classifier used 1,818 other recordings from 525 patients, with neither recording nor patient overlap with the 300-record evaluation arm. The P16 classifier used 2,940 records and excluded the 300 evaluation identifiers; patient-level separation could not be verified because P16 does not publish patient identifiers. BMD-HS and PaHS were never used to train a denoiser or classifier and were scored zero-shot by the P16 abnormality models. The consolidated manifest and separation reports are provided in Supplementary Data 2.

Artificial corruption used three recorded and three generated interference types at 0, 5 and 10 dB

SNR, yielding 18 conditions per CirCor, P16 and BMD-HS arm. The set was chosen to cover complementary mechanisms rather than to imitate one deployment site: ESC-50 environmental sounds represented diverse non-stationary surroundings, ICBHI respiratory sounds represented physiological overlap, and the Hospital Ambient Noise Dataset represented clinical ambient interference [23–25]. Generated chirps, narrowband tones and impulses probed swept-frequency, quasi-stationary spectral and transient disturbances. Recorded clips were cropped randomly or tiled before cropping. Chirps swept linearly between start and end frequencies drawn from 50–400 Hz; narrowband interference summed one to three sinusoids drawn from 40–300 Hz with random phase and amplitude; impulse interference used 5–15 signed unit spikes at random positions. Noise was scaled by global signal and noise power, and achieved SNR was verified. Every generator was seeded from corpus, record path, noise type, SNR and base seed so all methods received the same corrupted waveform.

Environmental and respiratory recordings were also used during denoiser training. Hospital ambient, chirp and narrowband interference were unseen by all learned models. Impulses were unseen by the eight non-impulse-trained models but exposure-matched for the two impulse-trained variants. Accordingly, results were not reduced to a binary seen/unseen-noise claim. The no-added-noise condition passed the original loaded recording through each model without imposing one of these artificial corruptions. The original BMD-HS and PaHS arms were evaluated separately because no paired clean reference exists for their naturally occurring recording conditions.

Denoisers, comparators, classifiers and training

Ten learned implementations were trained on a common CirCor partition and grouped by their primary design role (Table 2). Five task-aware models comprised a complex U-Net (cunet), a deep complex convolution recurrent network (dccrn), an attention variant (dccrn_attn), and end-to-end and frozen-front-end two-stage refiners (stage2_e2e and stage2_frozen). Their objectives combined waveform reconstruction with a consistency term from a fixed encoder. Two impulse-trained variants (dccrn_b and stage2_b_frozen) used related structures plus impulse exposure. Three smaller reconstruction-only baselines were a fully convolutional autoencoder (fcn), a U-Net (unet) and a U-Net–LSTM hybrid (lunet). Architectural components were adapted from established ECG, image, PCG and speech enhancement designs [3, 26–29]; they should not be interpreted as ten independently published PCG checkpoints or a ranking of all published denoisers.

CirCor patients were partitioned 60%/40% into the denoiser pool and the 377-patient reserve, with 10% of the denoiser pool held out by patient for validation. Training mixtures added environmental or respiratory noise with SNR sampled uniformly from -5 to +15 dB, and 20% of samples had no added noise. Only the two impulse-trained variants received impulses, on half of their training samples; validation remained impulse-free. Retained checkpoint `train_args` confirmed 50 epochs, learning rate 2×10^{-4}, batch size 16, seed and mixture and split parameters for the single-stage task-aware and impulse-trained models. Matching trainer control flow specified AdamW and gradient clipping at 5.0. For the two-stage models, checkpoint arguments confirmed 40 epochs and the following refiner/front-end learning rates: $2 \times 10^{-4}/2 \times 10^{-5}$ for the end-to-end model, whereas frozen variants updated only the refiner at 2×10^{-4}. Matching trainer control flow specified Adam and minimum-total-validation-loss retention. The fcn, unet and lunet checkpoints did not retain run arguments. Their matched archived baseline trainer specified default settings of AdamW, learning rate 2×10^{-4}, batch size 16, gradient clipping at 5.0 and 50 epochs. For noise-corrupted samples, their objective was the sum of negative SI-SDR and multi-resolution short-time Fourier magnitude loss. For no-added-noise samples, it was waveform L_1 loss minus 0.1 times SI-SDR. The trainer specified no learning-rate scheduler or early stopping. Importantly, these baselines were not selected

by minimum validation loss: the checkpoint with the highest mean validation SI-SDR was retained.

The single-stage objective weights for waveform reconstruction, identity, task consistency and gating regularisation were 1.0, 1.0, 0.5 and 0.4. The end-to-end two-stage objective added frequency anchoring and refiner regularisation and used weights 1.0, 1.5, 0.5, 0.4, 0.5 and 0.05. The frozen-front-end variant disabled gating and frequency anchoring and used 1.0, 1.5, 0.5, 0.0, 0.0 and 0.01. The consistency encoder was used only during optimisation. Parameter counts ranged from 2.99 to 6.15 million for task-aware or impulse-trained models and from 0.09 to 0.20 million for the reconstruction baselines (Supplementary Table S11). Capacity, objective and architecture therefore vary together; family comparisons are robustness checks, not factorial attribution to one component.

Four deterministic classical transformations and identity provided non-learned controls: a 20–500 Hz zero-phase fourth-order Butterworth band-pass; a local Wiener filter with a 51-sample window; a median filter with kernel length 5; and spectral subtraction using the median spectrum of the quietest 10% of frames with a 5% magnitude floor. Each produced the same 2 kHz, 5 s output interface as the learned models.

Four diagnostic classifiers consumed per-recording-standardised log-mel spectrograms: a two-dimensional CNN (cnn2d, primary), a bidirectional LSTM, a convolutional–recurrent network and a CNN with transformer encoder [30, 31]. Each 5-s, 2-kHz waveform was converted to a 64-band log-mel power spectrogram using a 256-sample Fourier transform, a 64-sample hop and a 0–800-Hz mel range. Power was converted to decibels relative to the recording maximum, yielding 157 frames, and the complete time–frequency representation was standardised using its recording-specific mean and standard deviation. The four architectures contained 355,650, 594,434, 681,442 and 354,786 trainable parameters, respectively; these counts exclude non-trainable batch-normalisation statistics and positional-encoding buffers.

The matched production trainer's default configuration specified 30 epochs, Adam with a learning rate of 10^{-3}, batch size 32, cross-entropy loss and run-initialisation seed 42. Class imbalance was handled by replacement sampling with weights inversely proportional to class frequencies. The retained classifier weights were held fixed across all denoising comparisons. The source specifies no internal validation split, learning-rate scheduler, early stopping or validation-based checkpoint selection and retains the state after the final configured epoch. No denoised recording entered classifier training.

The CirCor classifiers were fitted to 1,818 recordings from 525 patients for Present-versus-Absent murmur detection after excluding the complete patient-level evaluation reserve. This task is supported by the preserved split and source-data records; the generic checkpoint class names do not independently encode murmur semantics. The P16 classifiers were fitted to 2,940 normal/abnormal recordings after excluding all evaluation identifiers and were then applied without retraining to BMD-HS and PaHS. Within every comparison, input records, labels, evaluation subsets and classifier weights were fixed; only preprocessing changed. Retained checkpoint metadata independently confirmed architecture, source corpus, input dimensions and training-set size for all eight classifier instances. Optimisation settings were recovered from the production source whose output path and checkpoint schema matched those files; because the checkpoints did not embed command-line arguments, an unrecorded command-line override cannot be independently excluded. Configuration and provenance details are given in Supplementary Table S12 and the Supplementary Information section "Benchmark resource and training provenance." The same run-level limitation applies to the source-linked fcn, unet and lunet settings that were not embedded in their checkpoints; it does not apply to the seven learned systems whose `train_args` were retained.

Evaluation outcomes, robustness and statistical analysis

The primary outcome was paired ΔAUROC in Eq. 1. Sensitivity, specificity and balanced accuracy were also calculated at the Youden-optimal threshold fitted on the undenoised reference for each corpus and classifier and then held fixed after preprocessing. A held-out sensitivity analysis fitted the threshold on one random half of the available clusters and evaluated both conditions on the other half over 200 splits. Thresholds re-optimised separately within each condition were reported only as a descriptive best case.

The no-added-noise passthrough evaluated whether a module behaved approximately as an identity transformation on target recordings without requiring a paired clean reference. For artificial-corruption arms, where the uncorrupted waveform was known, reference-based fidelity was the denoised-minus-corrupted change in SI-SDR. For original recordings, no reference-based fidelity claim was made. Reference-free processing intensity was the fraction of positive short-time magnitude attenuated over the whole recording. Its Pearson association with no-added-noise ΔAUROC was calculated across the 15 learned, classical and identity methods within each corpus and interpreted descriptively.

To localise the transformation, short-time magnitude was partitioned into detected cardiac-event regions and their complement. Short-time Fourier transforms used a 256-sample Hann window (128 ms), 192-sample overlap, 64-sample hop (32 ms), 129 frequency bins and 7.8 Hz spacing. The analytic-signal envelope of the original input was smoothed with a zero-phase second-order 20 Hz low-pass. Peaks had at least 0.15 s separation and prominence equal to the difference between the envelope's 75th percentile and median. Each interval was extended until the envelope fell below 35% of that peak's amplitude and was intersected with 20–500 Hz. The mask was derived only from the original recording.

For mask R, positive magnitude attenuation was

$$A_{id}^{R} = \frac{\sum_{t,f} R(t,f)[|X_i(t,f)| - |\widehat{X}_{id}(t,f)|]_+}{\sum_{t,f} R(t,f)|X_i(t,f)| + 10^{-12}}. \tag{3}$$

We denoted event-region attenuation by A^C, complementary-region attenuation by A^E and their descriptive selectivity contrast by $A^E - A^C$. Added magnitude was accumulated separately so suppression was not conflated with introduced content. Because the latent decomposition $x = s + n$ is non-identifiable, these outcomes localise observable attenuation but do not identify the residual as causal cardiac or disease information.

Alternative explanations and robustness were tested with the identity and classical comparators; denoiser-free 2 kHz→1 kHz→2 kHz and 2 kHz→16 kHz→2 kHz zero-phase polyphase paths; restoration of every denoised waveform to its original root-mean-square amplitude; replication across four classifiers; and 11 one-factor-at-a-time mask settings. Mask alternatives were envelope low-pass 10/40 Hz, interval boundary 0.25/0.50, peak separation 0.10/0.20 s, frequency band 20–400/20–600 Hz and temporal dilation ±50/±100 ms around the nominal setting. Mask sensitivity was recomputed for cunet, dccrn_attn, fcn and unet on 100 records per corpus.

An exploratory information probe assessed linear separability after preprocessing. Each recording's standardised log-mel representation was reduced to 40 principal components, with standardisation and PCA fitted inside each fold, and entered into regularised logistic regression. Five-fold cross-validation was repeated five times with identical folds for the original and transformed data. Folds were stratified and patient-grouped for CirCor and BMD-HS and stratified by record for P16. This low-capacity probe measures one representation, not information recoverable by every classifier.

Percentile 95% confidence intervals used 2,000 paired bootstrap resamples with random seed 0. Primary intervals resampled patients for CirCor and BMD-HS and records for P16 and PaHS. All conditions and all records belonging to a selected patient were resampled together. A PaHS sensitivity analysis resampled its 224 parent recordings; this widened intervals by a median factor of 1.12 without changing any interval-based conclusion.

Two-sided empirical bootstrap p values used a plus-one finite-resampling correction and were adjusted by Benjamini–Hochberg across the analysis-wide family of 2,360 learned-denoiser–condition–classifier comparisons. Counts of intervals wholly above or below zero were descriptive and were not called multiplicity-adjusted discoveries. Classical-filter, fidelity, processing-intensity, mask-sensitivity and probe analyses were secondary or exploratory; their correlations and interval summaries were descriptive. Because the evaluation arms were class-balanced, prevalence-sensitive quantities should not be read as deployment-prevalence estimates.

Reproducibility records include the complete 2,360-row effect table (Supplementary Data 1), a 1,500-row arm-level manifest, subset records and available patient/record separation checks (Supplementary Data 2). Reported signal-analysis constants were checked programmatically against the executing code. Public release of the executable analysis and an archival DOI is described in the Code availability section.

Only publicly available, de-identified recordings were analysed, and no new participant data were collected. The original corpus reports describe their respective ethical approval and consent procedures [19–22]. No participant was contacted and no identifiable private information was accessed.

Use of generative AI. During manuscript preparation, the authors used OpenAI ChatGPT to improve language, organisation and consistency. The authors reviewed and edited the output and take full responsibility for the manuscript. No AI tool was used to generate or alter the reported experimental observations.

Data availability

All recordings analysed in this study are publicly available from their original distributors, and no new participant data were generated. The four clinical corpora are: CirCor DigiScope version 1.0.3 (PhysioNet, https://doi.org/10.13026/tshs-mw03); PhysioNet/CinC 2016 version 1.0.0 (PhysioNet, https://physionet.org/content/challenge-2016/1.0.0/); BMD-HS (BUET mHealth Laboratory, CC BY-NC-ND 4.0, https://github.com/sani002/BMD-HS-Dataset); and the PASCAL Classifying Heart Sounds Challenge material (https://www.peterjbentley.com/heartchallenge/index.html). Recorded-noise sources are: ESC-50 (https://github.com/karolpiczak/ESC-50); the ICBHI 2017 Respiratory Sound Database (https://www.kaggle.com/datasets/vbookshelf/respiratory-sound-database); and the Hospital Ambient Noise Dataset of Ali and Shuvo (CC BY-NC-ND 4.0, https://doi.org/10.34740/KAGGLE/DSV/2173743). Neither the source hospital audio nor any derived mixture is redistributed with this article; users should obtain each corpus from the cited page under its own licence terms. Supplementary Data 1 contains all 2,360 learned-denoiser-condition-classifier paired effects. Supplementary Data 2 contains the consolidated evaluation manifest, source metadata, subset-selection records, consistency checks and the available training-evaluation separation reports.

Code availability

The source code and numerical result files required to reproduce the findings of this manuscript are publicly available in Zenodo at 10.5281/zenodo.22702225 [33].

References

[1] Chorba, J. S. et al. Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform. *J. Am. Heart Assoc.* **10**, e019905 (2021).

[2] Reyna, M. A. et al. Heart murmur detection from phonocardiogram recordings: the George B. Moody PhysioNet Challenge 2022. *PLOS Digit. Health* **2**, e0000324 (2023).

[3] Ali, S. N., Shuvo, S. B., Al-Manzo, M. I. S., Hasan, A. & Hasan, T. An end-to-end deep learning framework for real-time denoising of heart sounds for cardiac disease detection in unseen noise. *IEEE Access* **11**, 87887–87901 (2023). https://doi.org/10.1109/ACCESS.2023.3292551.

[4] Le Roux, J., Wisdom, S., Erdogan, H. & Hershey, J. R. SDR—half-baked or well done? In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)* 626–630 (2019).

[5] Finlayson, S. G. et al. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021).

[6] Zech, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLOS Med.* **15**, e1002683 (2018).

[7] Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).

[8] Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).

[9] DeGrave, A. J., Janizek, J. D. & Lee, S. I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).

[10] Koch, L. M., Baumgartner, C. F. & Berens, P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. *npj Digit. Med.* **7**, 120 (2024).

[11] Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digit. Med.* **5**, 48 (2022).

[12] Shariat Panah, D., Hines, A. & McKeever, S. Exploring the impact of noise and degradations on heart sound classification models. *Biomed. Signal Process. Control* **85**, 104932 (2023). https://doi.org/10.1016/j.bspc.2023.104932.

[13] González-Rodríguez, C., Alonso-Arévalo, M. A. & García-Canseco, E. Robust denoising of phonocardiogram signals using time–frequency analysis and U-Nets. *IEEE Access* **11**, 52466–52479 (2023).

[14] Jakubec, M., Lieskovská, E. & Počta, P. A robust deep learning based model for denoising phonocardiogram signals in clinical environments. *Eng. Appl. Artif. Intell.* **161**, 112286 (2025). https://doi.org/10.1016/j.engappai.2025.112286.

[15] Pauline, S. H., Dhanalakshmi, S., Kumar, R., Narayanamoorthi, R. & Lai, K. W. A low-cost multistage cascaded adaptive filter configuration for noise reduction in phonocardiogram signal. *J. Healthc. Eng.* **2022**, 3039624 (2022). https://doi.org/10.1155/2022/3039624.

[16] Shervegar, V. M. A robust sliding window adaptive filtering technique for phonocardiogram signal denoising. *Expert Syst.* **42**, e13361 (2025). https://doi.org/10.1111/exsy.13361.

[17] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).

[18] Sokol, K., Fackler, J. & Vogt, J. E. Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap. *npj Digit. Med.* **8**, 345 (2025). https://doi.org/10.1038/s41746-025-01725-9.

[19] Oliveira, J. et al. The CirCor DigiScope dataset: from murmur detection to murmur classification. *IEEE J. Biomed. Health Inform.* **26**, 2524–2535 (2022).

[20] Ali, S. N. et al. BUET multi-disease heart sound dataset: a comprehensive auscultation dataset for developing computer-aided diagnostic systems. *Comput. Methods Programs Biomed. Update* **9**, 100237 (2026). https://doi.org/10.1016/j.cmpbup.2026.100237.

[21] Liu, C. et al. An open access database for the evaluation of heart sound algorithms. *Physiol. Meas.* **37**, 2181–2213 (2016).

[22] Bentley, P., Nordehn, G., Coimbra, M. & Mannor, S. The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) results. http://www.peterjbentley.com/heartchallenge/index.html (2011).

[23] Piczak, K. J. ESC: dataset for environmental sound classification. In *Proc. ACM Int. Conf. Multimedia* 1015–1018 (2015).

[24] Rocha, B. M. et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiol. Meas.* **40**, 035001 (2019).

[25] Ali, S. N. & Shuvo, S. B. Hospital ambient noise dataset. *Kaggle* (2021). https://doi.org/10.34740/KAGGLE/DSV/2173743. Licensed under CC BY-NC-ND 4.0.

[26] Chiang, H. T. et al. Noise reduction in ECG signals using fully convolutional denoising autoencoders. *IEEE Access* **7**, 60806–60813 (2019).

[27] Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *MICCAI* 234–241 (2015).

[28] Hu, Y. et al. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement. In *Proc. Interspeech* 2472–2476 (2020).

[29] Luo, Y. & Mesgarani, N. Conv-TasNet: surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**, 1256–1266 (2019).

[30] Vaswani, A. et al. Attention is all you need. In *Adv. Neural Inf. Process. Syst.* 5998–6008 (2017).

[31] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).

[32] Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, e215–e220 (2000).

[33] Wazed, E., Shin, C. & Jeong, H. Analysis code for "A multi-corpus benchmark reveals limits of diagnostic preservation in heart-sound denoising". *Zenodo* https://doi.org/10.5281/zenodo.22702225 (2026).

Funding

This work was supported by the National Research Foundation of Korea grant funded by the Korean government (MSIT) (No. RS-2026-25469002), and by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026 (Project No. RS-2026-25513329).

Author contributions

E.W.: Methodology, Software, Formal analysis, Investigation, Visualization and Writing–original draft. C.S.: Conceptualization and Writing–review and editing. H.J.: Conceptualization, Supervision, Funding acquisition, Writing–original draft and Writing–review and editing. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Figures

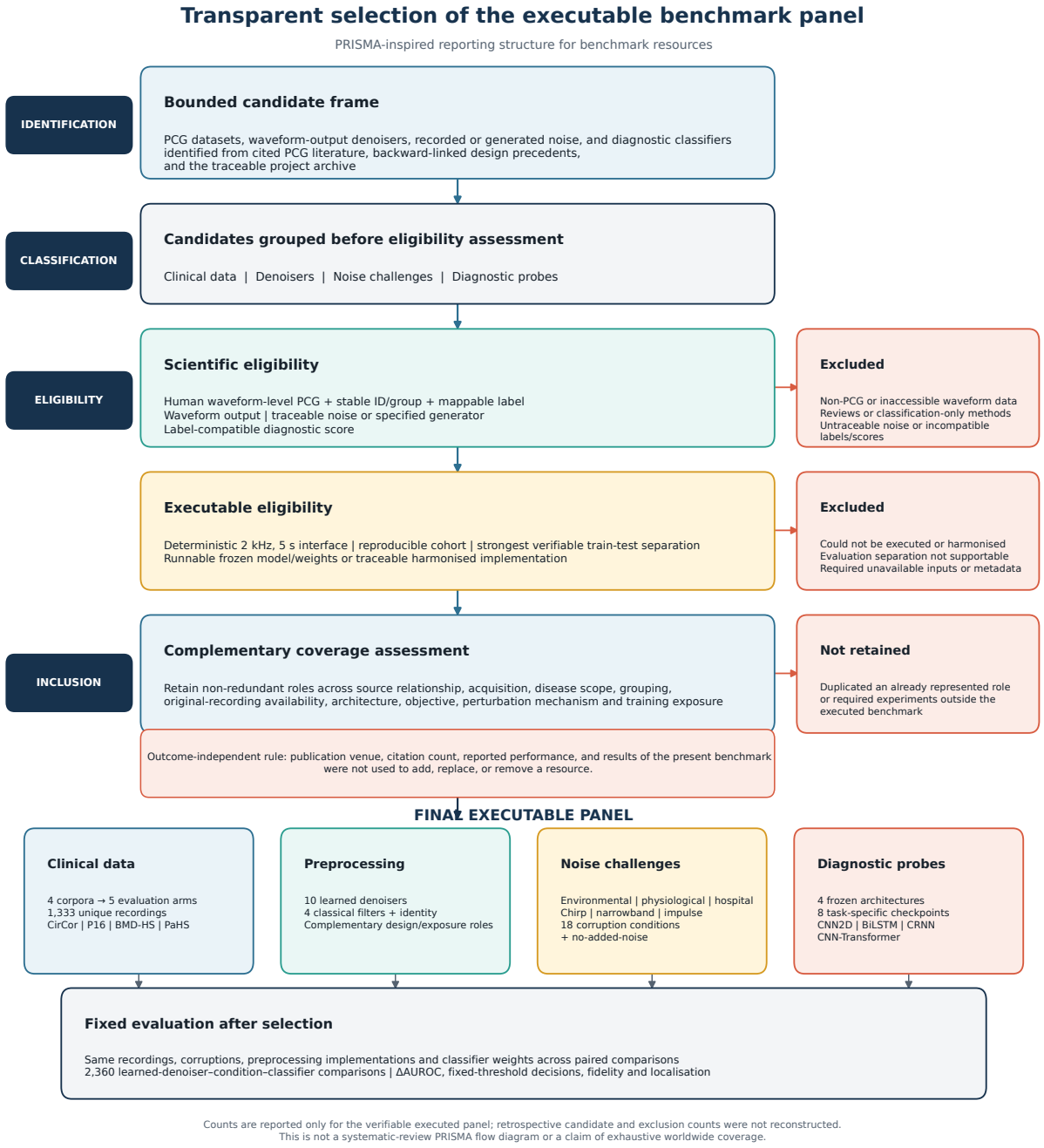

Figure 1: PRISMA-inspired selection of the executable benchmark panel. A bounded candidate frame passed resource classification, scientific and executable eligibility, and complementary-coverage assessment; side branches state exclusion reasons. Selection did not use publication venue, citation count, reported performance or present-study outcomes. Counts are limited to the verifiable executed panel because retrospective candidate and exclusion totals were not reconstructed. This is a transparent benchmark-selection diagram, not a systematic-review PRISMA flowchart or a claim of exhaustive worldwide retrieval.

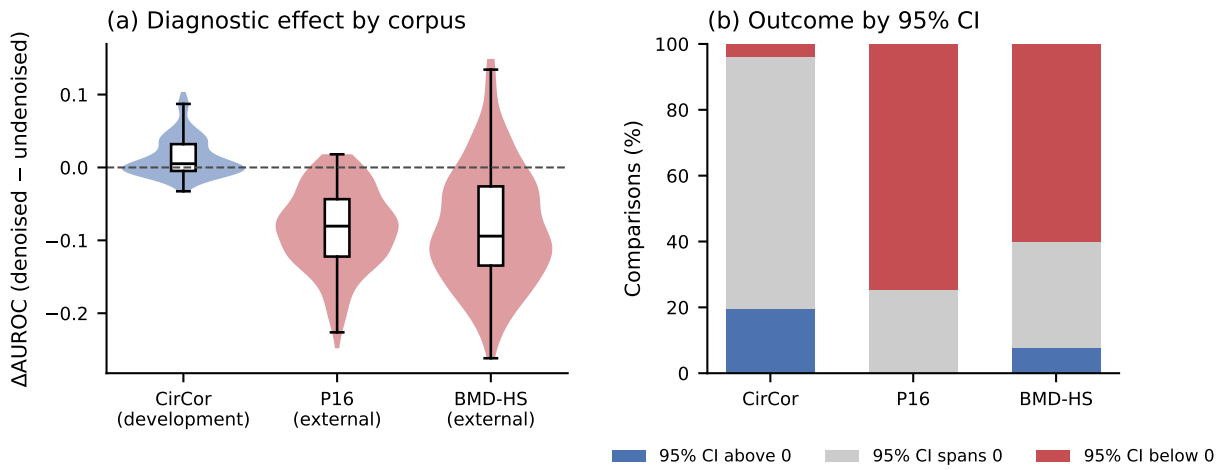

Figure 2: Diagnostic effect of learned denoising by corpus. (a) Paired ΔAUROC relative to the undenoised input across ten denoisers and 18 artificially corrupted conditions ($n = 180$ method–condition points per corpus). Violins show distributions; embedded boxes show median and interquartile range. (b) Percentages whose clustered 95% bootstrap interval lay above zero, spanned zero or lay below zero. Clusters were patients for CirCor and BMD-HS and recordings for P16. Interval categories are descriptive and are not multiplicity-adjusted discovery calls.

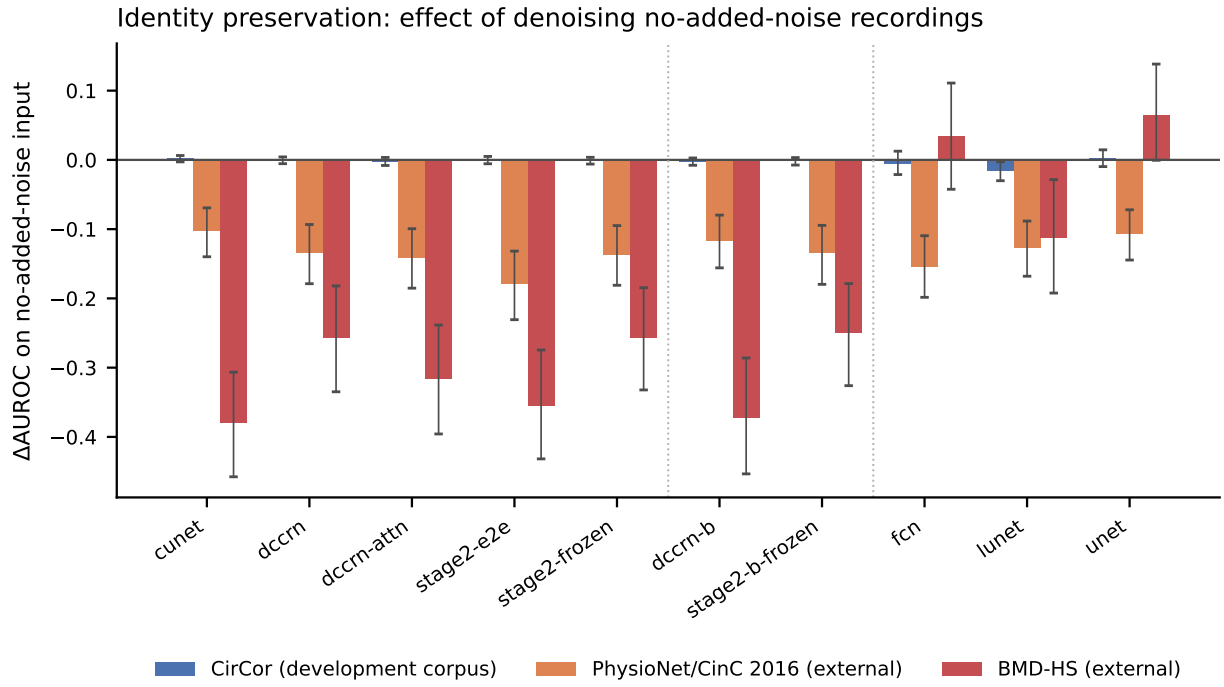

Figure 3: No-added-noise passthrough. Paired ΔAUROC when each learned denoiser was applied without adding artificial noise. Error bars are clustered percentile 95% bootstrap intervals (2,000 resamples; patient clusters for CirCor and BMD-HS and recording clusters for P16). Vertical separators identify denoiser families.

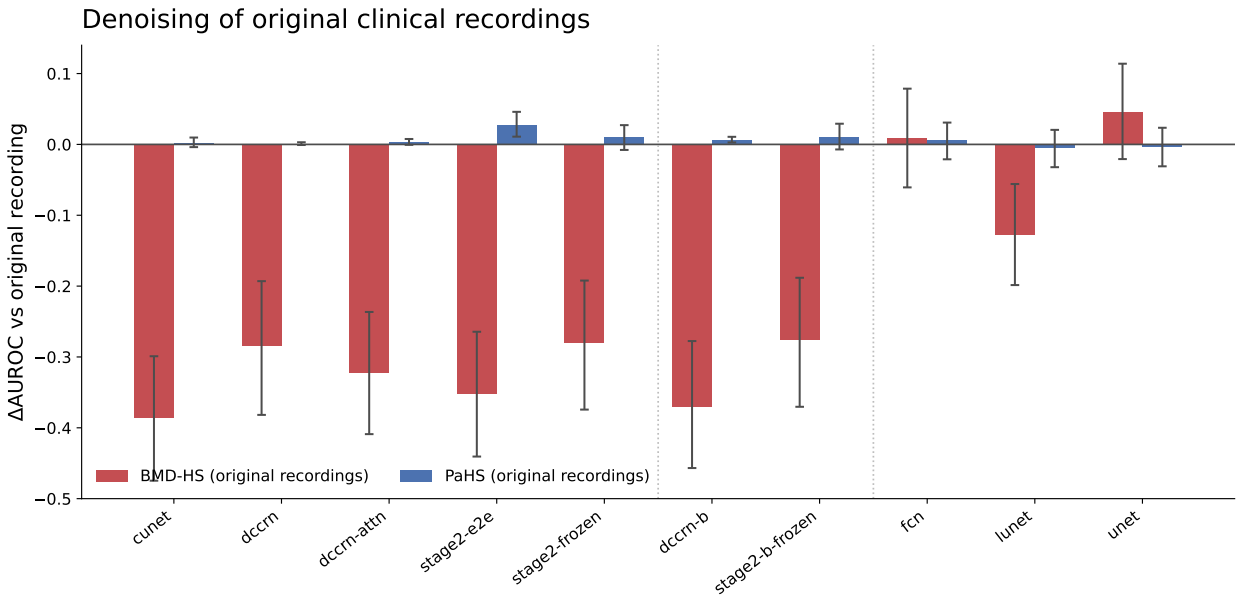

Figure 4: Original recordings without artificial corruption. Paired ΔAUROC after denoising released BMD-HS and PaHS recordings without paired clean references ($n = 300$ per arm). Error bars are clustered percentile 95% bootstrap intervals. BMD-HS used patient clusters; PaHS used recording clusters, with parent-recording clustering assessed separately.

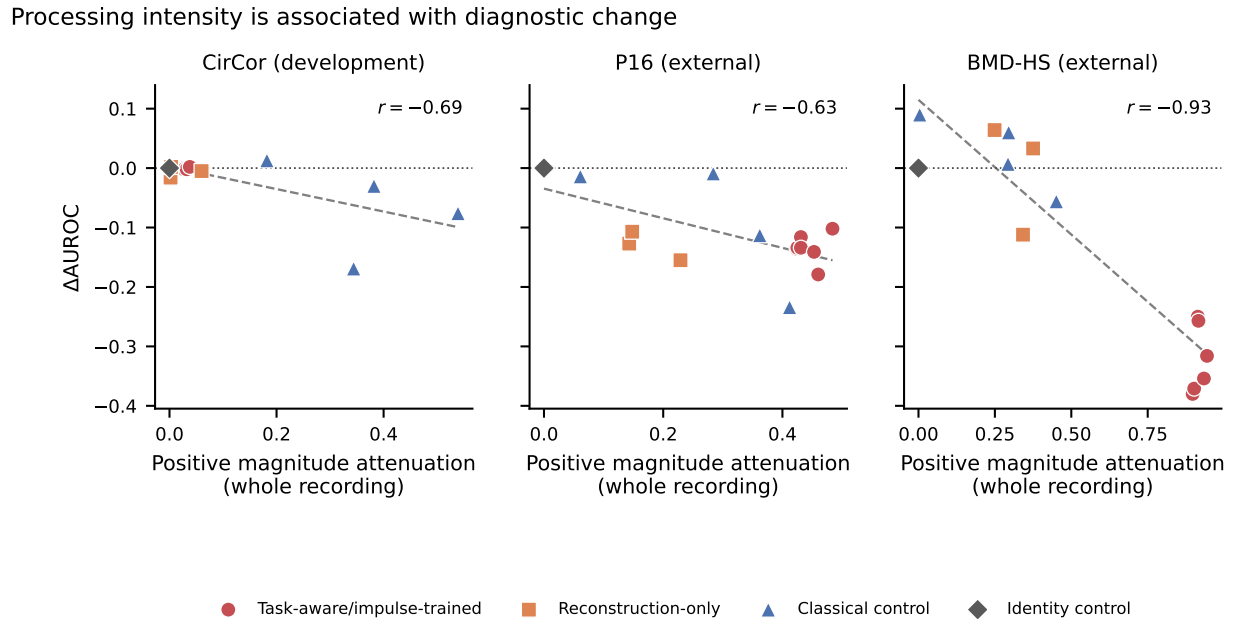

Figure 5: Processing intensity and diagnostic change. Whole-recording fraction of positive short-time magnitude attenuated versus no-added-noise ΔAUROC for ten learned denoisers, four classical filters and the identity control ($n = 15$ methods per corpus). Dashed lines are within-corpus least-squares fits and r is Pearson correlation. These heterogeneous-method correlations are descriptive, not causal or predictive models.

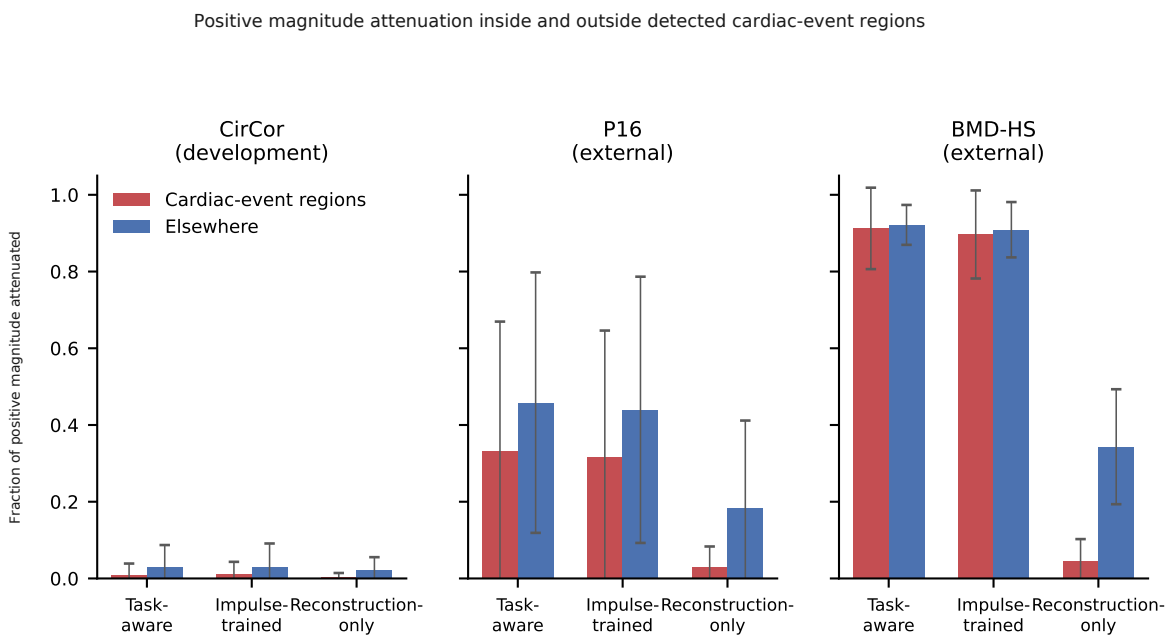

Figure 6: Localisation of positive magnitude attenuation in no-added-noise inputs. Fraction of original short-time magnitude attenuated within envelope-detected cardiac-event regions and elsewhere, stratified by denoiser family and corpus ($n = 300$ recordings per corpus). Bars show means and error bars show standard deviations. Events were detected from the original input only. The figure localises attenuation but does not establish that the attenuated component caused the diagnostic effect.

Tables

Table 1: Public clinical corpora, evaluation arms and design rationale. Positive/negative means murmur present/absent for CirCor and abnormal/normal for the other corpora. The two BMD-HS arms were drawn independently and share 167 recordings; their counts must not be summed as unique participants. "Recording" denotes that no patient identifier was released.

Corpus and arm	Role and task	Evaluation set	Why it is informative here
CirCor, artificial corruption + no-added-noise	Source; murmur detection	300 (150/150); 197 patient clusters	Multi-location paediatric PCGs with patient identifiers permit a patient-disjoint, held-out source-distribution benchmark.
P16, artificial corruption + no-added-noise	Classifier source; normal/abnormal classification	300 (150/150); 300 recording clusters	Heterogeneous challenge benchmark provides an external denoiser test and the normal/abnormal classifier used for zero-shot transfer.
BMD-HS, artificial corruption + no-added-noise	External; normal/abnormal classification	300 (150/150); 91 patient clusters	Contemporary multi-disease, repeated-location PCGs enable patient-clustered testing under substantial corpus shift.
BMD-HS, original recordings	External original-recording arm	300 (150/150); 97 patient clusters	Tests the fixed pipeline without an artificial noise model or paired clean reference.
PaHS, original recordings	External challenge; abnormality	300 (150/150); 300 recording clusters from 224 parents	Independent material tests corpus dependence; parent-recording resampling checks the available dependence structure.

Table 2: Model, provenance and preprocessing inventory. The set was chosen to vary architecture, objective and perturbation exposure while retaining identical paired inputs and fixed downstream weights. Provenance describes the evaluated implementation, not merely its architectural inspiration. Family comparisons are therefore robustness checks, not factorial tests of a single design choice.

Component	Members	Distinguishing feature	Role in the preservation assessment
Task-aware denoisers	cunet, dccrn, dccrn_attn, stage2_e2e, stage2_frozen	Study-developed variants; reconstruction plus fixed-encoder consistency	Spans complex, recurrent, attention and two-stage processing intended to preserve task structure.
Impulse-trained denoisers	dccrn_b, stage2_b_frozen	Study-developed variants with impulse exposure during training	Tests whether matching a perturbation class prevents transfer failure.
Reconstruction-only denoisers	fcn, unet, lunet	Local architectural baselines adapted from published designs; no task-consistency term	Provides smaller-capacity reconstruction-only contrasts under the same data and interface.
Classical controls	Band-pass, Wiener, median, spectral subtraction, identity	Deterministic controls; no learned weights	Checks whether diagnostic perturbation is unique to learned models and tests bandwidth, smoothing and spectral-estimation effects.
Primary classifier	cnn2d	Study-developed fixed outcome probe; local time–frequency convolutions	Primary diagnostic architecture for paired effect estimation; not a denoiser.
Sequence classifier	bilstm	Study-developed fixed outcome probe; bidirectional temporal recurrence	Tests dependence on sequential modelling; not a denoiser.
Hybrid classifier	crnn	Study-developed fixed outcome probe; convolution plus recurrence	Tests a combined local and temporal representation; not a denoiser.
Attention classifier	cnn_transformer	Study-developed fixed outcome probe; convolution plus self-attention	Tests an alternative long-range inductive bias; not a denoiser.

Table 3: Mapping from analytical preservation conditions to empirical evidence. The study can falsify strict preservation and localise attenuation, but it cannot certify a clinically acceptable non-inferiority margin or causally identify the attenuated component. AUROC pair equivalents follow main-text Eq. 2; counts are approximate because displayed AUROCs are rounded and ties contribute one half.

Condition	Empirical test	Result and inference
Diagnostic preservation, $\Delta U \geq 0$ (strict reference)	Paired artificially corrupted and original-recording AUROC	External mean effects were negative; on original BMD-HS, cunet changed AUROC by -0.387, about 8,708 net pair-order losses among 22,500 positive–negative pairs. Strict preservation was falsified.
No-added-noise preservation, $\Delta U^{(0)} \geq 0$	Insert the denoiser without artificial corruption	All 10 P16 effects and 8 of 10 BMD-HS effects (all 7 task-aware or impulse-trained models plus lunet) were negative at $q < 0.05$; the worst BMD-HS effect was -0.380 (about 8,550 net pair-order losses).
Stable fixed-threshold decisions	Score shift must not cross the reference threshold	The largest BMD-HS sensitivity change was -0.767, corresponding to about 115 fewer true-positive decisions among 150 positives.
Cardiac-event sparing	Compare A^C with A^E	Task-aware and impulse-trained BMD-HS models attenuated 0.908 within event regions and 0.918 elsewhere on average. Event regions were not spared, but residual source and causality remain unidentified.
Alternative explanations	Resampling-only and amplitude-restoration controls	The 16-kHz resampling path was inert; BMD-HS 1-kHz effects were -0.030 to $+0.006$ and amplitude restoration changed no reported AUROC to three decimals. Neither reproduced the learned-model loss.
Classifier robustness	Repeat with cnn2d, bilstm, crnn and cnn_transformer	Mean artificial-corruption effects were positive on CirCor and negative on P16 and BMD-HS for all four classifiers; effect size and PaHS direction remained classifier dependent.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ScientificReportsSupplementaryInformation.pdf
- SupplementaryData2manifests.zip