An XAI-Based Auditing Framework for Detecting *Shortcut Learning* in Visual Cardiac Murmur Screening

Juliano Costa Reis[1], **Ronaldo Martins da Costa**[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brazil

julianocosta@discente.ufg.br, ronaldocosta@ufg.br

Abstract. *We propose an interpretability auditing framework for spectrogram-based cardiac murmur screening, combining Leave-One-Group-Out cross-validation, Grad-CAM, the sanity check of Adebayo et al., Deletion/Insertion fidelity metrics and a structured random-saliency control. Across 36 model instances on PhysioNet/CinC 2016, in-distribution AUC of 0.945 falls to 0.497 at an unseen site (36 of 36 instances), with severe overconfidence. Yet explanations remain faithful: sanity holds in 76.3% of maps and Grad-CAM beats random saliency on both fidelity metrics in all configurations. We further show that the reported disagreement between Deletion and Insertion is largely driven by the model's operating point and by measurement choices. Faithful explanations do not imply a useful model.*

1. Introduction

Cardiovascular diseases are among the leading causes of premature mortality, and cardiac murmurs detectable by auscultation are early markers of valvular disease [Roth et al. 2020]. The shortage of cardiologists, especially in low- and middle-income regions, motivates automated screening systems based on the phonocardiogram (PCG) [Ma et al. 2024]. A dominant line of work converts the temporal signal into a Mel spectrogram and classifies it as an image, leveraging ImageNet-pretrained networks [Li et al. 2022, Kim et al. 2022]. In this context, compact architectures designed for edge devices become essential to enable point-of-care screening [Krones et al. 2024, Kim et al. 2024].

Despite reported in-distribution performance, two issues are rarely examined jointly. First, models trained on multi-site data may learn shortcuts: exploiting background textures associated with the stethoscope or acoustic environment instead of the murmur structure [Geirhos et al. 2020]. Such shortcuts inflate metrics under stratified validation but collapse when the model encounters an unseen site. Second, even when explainability techniques such as Grad-CAM [Selvaraju et al. 2017] are applied, it is rarely verified whether the explanations are faithful to what the model actually uses, or whether they merely appear plausible to the human eye.

This paper shifts the focus from the solution to the audit. Rather than proposing yet another classifier, we propose and apply a reproducible framework to diagnose shortcut learning and assess explanation fidelity in visual murmur screening. In applying it we find a regime of general interest to the XAI community: models whose out-of-domain performance is indistinguishable from chance, yet whose explanations mostly pass the sanity check and outperform the random-saliency control on fidelity metrics. The audit

detected exactly the phenomenon that motivated its construction, and detected it in our own models.

Contributions. (i) An auditing framework integrating LOGO-CV, Grad-CAM, the sanity check [Adebayo et al. 2018], Deletion/Insertion [Petsiuk et al. 2018] and a structured random-saliency control, as a reproducible pipeline (Algorithm 1); (ii) its application to 36 model instances (3 architectures $\times$ 6 folds $\times$ 2 regimes) on PhysioNet/CinC 2016; (iii) evidence of near-chance cross-site performance, consistent with shortcut learning although not sufficient to demonstrate it in isolation; (iv) the finding that sanity and fidelity tests are largely satisfied precisely in this regime, establishing that explanation fidelity does not imply model usefulness; (v) a demonstration that disagreement between fidelity metrics across regimes can be an artifact of the model's operating point, with a normalization that corrects it; and (vi) the finding that energy-based localization is limited where murmur and normal cardiac energy overlap.

2. Related Work

Murmur screening as computer vision. Since the PhysioNet/CinC 2016 challenge [Clifford et al. 2016], the dominant approach treats spectrograms as images, applying CNNs and, more recently, Vision Transformers [Li et al. 2022, Kim et al. 2022]. Convolutional-recurrent architectures have also been explored [Alkhodari and Fraiwan 2021], and edge-oriented work prioritizes compact models [Ma et al. 2024]. This literature predominantly reports classification metrics under stratified validation, in which recordings from the same site appear in both training and test; cross-site robustness is therefore rarely measured, and it is this gap that the LOGO-CV design addresses.

Shortcut learning and domain heterogeneity. [Geirhos et al. 2020] formalize the phenomenon: deep networks exploit spurious correlations that do not generalize beyond the training distribution. In PCG, inter-site heterogeneity — equipment, acoustic environment, protocol — imprints distinct background textures on spectrograms, natural candidates for shortcuts, and leave-one-group-out protocols expose the effect. In medical imaging the evidence is robust: [Zech et al. 2018] show a pneumonia-screening CNN losing performance at an unseen hospital while readily identifying the institution of origin, and [DeGrave et al. 2021] show radiographic COVID-19 detectors selecting shortcuts over clinical signal. Such protocols remain little used in medical audio, which makes empirical evidence scarce in that domain.

Evaluating explanations. Grad-CAM [Selvaraju et al. 2017] produces gradient-based saliency maps and became a de facto standard in PCG-as-image pipelines, partly for its low computational cost. Plausible maps, however, are not necessarily faithful: [Adebayo et al. 2018] propose sanity checks via weight randomization, and [Petsiuk et al. 2018] propose Deletion/Insertion, which measure the impact on confidence when the most salient pixels are progressively removed or inserted. Alternatives with distinct assumptions exist and fall into three families: perturbation, such as LIME [Ribeiro et al. 2016] and RISE [Petsiuk et al. 2018]; axiomatic attribution, such as Integrated Gradients [Sundararajan et al. 2017] and SHAP [Lundberg and Lee 2017]; and activation-map variants, such as Grad-CAM++ [Chattopadhay et al. 2018] and Score-CAM [Wang et al. 2020]. The choice among them is not neutral. We adopt Grad-CAM

because it is the method used in the literature we audit; the framework, however, is agnostic to the explainer. [Krishna et al. 2024] show that explanations from different methods frequently disagree and that practitioners resort to *ad hoc* heuristics.

XAI applied to phonocardiogram. Two recent works are the closest precedent. [Özcan 2024] combines Mel spectrograms, transfer learning and Occlusion Sensitivity for murmur detection and grading; [Alrabie and Barnawi 2025] compare pretrained architectures, among them MobileNet and EfficientNetV2B0, and apply Grad-CAM to verify whether models attend to clinically meaningful components of the cardiac cycle (S1, systole, S2, diastole). Both establish that *where the model looks* is a pertinent question here. Neither, however, evaluates explanation *fidelity* through perturbation metrics, nor robustness across collection sites.

Positioning. These lines have advanced in isolation, and to our knowledge no prior work jointly audits shortcut learning and explanation fidelity in spectrogram-based murmur screening under a leave-one-site-out protocol. Our novelty is not a new explanation method but the integration of these axes into a single reproducible audit, and the finding — first reported here in this domain — that explanation quality tests can be largely satisfied by models without appreciable out-of-domain discriminative capacity.

3. Methodology

3.1. Dataset

We use PhysioNet/CinC 2016 [Clifford et al. 2016]: 3,240 recordings from six sites (a–f), with heterogeneous equipment and environments. The set exhibits a class imbalance of 3.87:1 in the normal:abnormal direction (2,575 normal against 665 abnormal) and a dominant site: site *e* concentrates 66.1% of the samples, configuring a strong intra-dataset shift that drives the cross-domain challenge. Prevalence also inverts across sites: 71.4% abnormal at site *a* (117 normal, 292 abnormal) against 8.5% at site *e* (1,958 normal, 183 abnormal), so the dominant site is also the most discrepant in class distribution. The remaining sites contribute 490 (b), 31 (c), 55 (d) and 114 (f) recordings.

3.2. Visual Representation

Each recording is resampled and segmented into fixed windows; the Short-Time Fourier Transform yields the Mel spectrogram (128 bands), normalized to [0,1]. The matrix is resized to 224×224 and replicated into three channels (RGB) with ImageNet normalization, activating the low-level filters of the pretrained networks. Inference preprocessing is identical to training, a necessary condition for the saliency maps to be faithful.

3.3. Models and Training Regimes

We evaluate three compact, edge-oriented architectures — MobileNetV3-Small [Howard et al. 2019], EfficientNet-B0 [Tan and Le 2019] and MobileViT-S [Mehta and Rastegari 2022].

Rationale for selection. The architectures were chosen not for absolute accuracy but to cover complementary families of compact models under a fixed budget, since the target scenario is point-of-care screening in low-resource settings [Krones et al. 2024], where edge compatibility is a requirement [Ma et al. 2024]. MobileNetV3-Small

Table 1. Differences between the training regimes. The two differences are simultaneous and conceptually linked, so that effects observed between the protocols cannot be attributed to the balanced sampler in isolation.

Aspect	Protocol II	Protocol III
Sampling	`shuffle=True` (simple random)	`WeightedRandomSampler` with weights inversely proportional to the frequency of each (site, class) pair
Class weight	`pos_weight` = 3.873	`pos_weight` = 1.0
Optimizer	AdamW (weight decay = 0.01)	AdamW (weight decay = 0.01)

and EfficientNet-B0 represent depthwise-separable design and compound scaling; MobileViT-S introduces a hybrid convolution–attention inductive bias, following the adoption of transformer backbones for PCG [Kim et al. 2022]. Backbones from the first two families have been evaluated on heart sounds with Grad-CAM analysis [Alrabie and Barnawi 2025]. We do not claim these three dominate the PCG-as-image literature, which ranges from AlexNet [Li et al. 2022] to convolutional-recurrent models [Alkhodari and Fraiwan 2021]. The contrast matters for the audit: local receptive fields and global attention aggregate spectrographic evidence differently, and including both guards the conclusions against being an artifact of a single inductive bias.

Training regimes. We compare two regimes, here denoted Protocols II and III. They differ in two simultaneous aspects, and not only in sampling, which restricts the causal attribution of the observed effects (Table 1). Each architecture $\times$ regime pair is trained under LOGO-CV with six folds, totalling 36 trained model instances. We emphasize that this number corresponds to repeated trainings of three architectures under distinct partitions and regimes, and not to 36 different network designs: it reflects the breadth of the audit, not architectural diversity.

3.4. Training and Validation Protocol

For each fold the held-out site is reserved exclusively for testing. From the five remaining sites, 15% of the recordings — stratified by (site, class) — form an independent validation set. The scheduler, early stopping and checkpoint selection are governed exclusively by the validation AUC; the held-out site is evaluated once, at the end of training, and an automatic check ensures that no recording and no site of the test set appears in training or validation. We adopt AUC rather than F1 because it is threshold-free: since Protocol III shifts the operating point (Section 4.4), selecting by F1 would confound a threshold shift with a gain in discrimination.

3.5. Auditing Pipeline

The methodological core is Algorithm 1. Each sample is explained by the model of the fold that left it out, ensuring the audit reflects generalization rather than memorization. Two steps support the analyses of Section 4: the random-saliency control and the recording of the confidence on the intact input, required for the normalization of Section 3.7.

3.6. Metrics

Sanity [Adebayo et al. 2018]: we compare the original map to that of a model with target layer and classifier randomized, via SSIM; a low SSIM indicates dependence on the

Input: trained models $\{M_{(a,s,p)}\}$, samples D
Output: aggregated tables + paired Wilcoxon tests

```
for each sample x of site s, label y in D do
  for each architecture a, protocol p do
    M ← M_(a,s,p)                                  // held-out site = s
    cam ← GradCAM(M, x, class = y)
    // 1. fidelity
    p_full ← P_M(y | x)                             // operating point
    del ← Deletion(M, x, cam, y)
    ins ← Insertion(M, x, cam, y)
    // 2. sanity
    M' ← randomize_weights(M)
    ssim ← SSIM(cam, GradCAM(M', x))
    // 3. random-saliency control
    cam_r ← 7×7 noise upsampled to 224×224
    del_r, ins_r ← Deletion/Insertion(M, x, cam_r, y)
    // 4. localization (support)
    far ← energy(cam ∩ mask) / energy(cam)
    record(del, ins, del_r, ins_r, ssim, far, p_full)
```

Algorithm 1. XAI auditing pipeline

learned weights. **Deletion/Insertion** [Petsiuk et al. 2018]: progressively removing (Deletion; lower is better) or inserting (Insertion; higher is better) the most salient pixels, we measure the area under the confidence curve. **Random-saliency control**: we repeat both with noise maps at the same effective resolution as Grad-CAM (7×7, bilinearly interpolated), preserving spatial smoothness and making the control paired. **FAR**: the fraction of Grad-CAM energy falling on a binary foreground mask obtained by Otsu thresholding of the Mel spectrogram, an exploratory localization indicator. **Calibration**: mean confidence, accuracy and Expected Calibration Error (ECE, 15 bins) [Guo et al. 2017] over the complete held-out set.

3.7. Sampling and Statistical Treatment

Sample selection. Fidelity metrics were computed over a balanced sample of eight recordings per class at each site, drawn with a fixed seed (42), totalling 95 samples — site c has only seven normal recordings, which explains the difference with respect to 96. The same 95 samples were used across all architectures and regimes, which makes comparisons paired on the same input.

Normalization by the operating point. Deletion and Insertion are areas under confidence curves whose ceiling is the model's confidence on the intact input. To make them comparable across regimes we define $Del^{\text{norm}} = Del/p^{\text{ref}}$ and $Ins^{\text{norm}} = Ins/p^{\text{ref}}$, where p^{ref} is the mean of $p_{\text{full}} = P(y \mid x)$ over the samples of that architecture–regime–fold cell. We use a per-model scalar rather than sample-wise division because p_{full} is near zero in part of the samples (32 to 70 of 570 per configuration), which destabilizes the individual quotient. Normalized values above 1 merely indicate that the area exceeds that model's reference confidence.

Statistical treatment. The 95 samples are nested within six sites, and each cell corre-

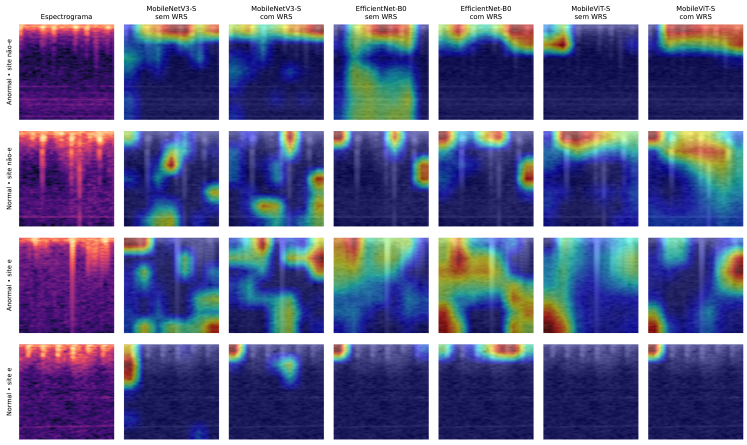


Figure 1. Grad-CAM maps per architecture and regime (without and with the balanced sampler, WRS), in class and site strata; column 1 is the input Mel spectrogram. Each sample is explained by the model of the fold that left it out, for its true class. Illustrative only; quantitative evaluation is in Tables 3 and 4.

sponds to a single trained model, so tests treating them as independent may incur pseudoreplication. We report three layers: paired Wilcoxon per sample; the count of folds in which the effect occurs in the indicated direction; and a 95% confidence interval from hierarchical bootstrap (10,000 resamplings of sites and, within them, of samples). Where multiple comparisons occur we also report the Holm-adjusted p-value.

4. Results

Figure 1 qualitatively illustrates how the highlighted regions vary across architectures, regimes and sites in the trained models, and serves as a visual reference for the quantitative analyses that follow. The examples do not constitute evidence of fidelity — that assessment is in Tables 3 and 4 — but they help interpret the heterogeneity quantified there. Two patterns are visible. First, the maps tend to concentrate on the low-frequency bands, at the top of the spectrogram, where, according to the per-band energy analysis, murmur energy and normal heart-sound energy overlap. Second, the contrast between the rows for site *e* and the remaining ones evidences the distinct behaviour when the dominant site is held out for testing.

4.1. In-Distribution and Out-of-Domain Performance

Table 2 reports performance over the complete held-out sets. Under LOGO-CV the 36 instances barely discriminate murmurs at the unseen site: mean AUC 0.497, with 18 of 36 below 0.50 and only one above 0.70. The models are also severely overconfident — mean confidence 0.872 against accuracy 0.448, with ECE up to 0.497.

The contrast between the two first numeric columns is the central result. The same models that reach AUC 0.945 on unseen recordings from the training sites fall to 0.497 — chance — at a site they have never seen. The mean drop is 0.448 and occurs in 36 of 36 instances (paired Wilcoxon, $p = 2.9 \times 10^{-11}$), ranging from 0.389 to 0.516 across sites (Figure 2).

Table 2. Performance and calibration, mean of the six folds. The validation column refers to unseen recordings from the same training sites; the test column, to the entirely held-out site.

Architecture	Reg.	AUC val.	AUC test	Acc.	Conf.	ECE	$P(\text{abn.})$
MobileNetV3-S	II	0.954	0.481	0.403	0.857	0.459	0.641
MobileNetV3-S	III	0.950	0.528	0.458	0.930	0.494	0.625
EfficientNet-B0	II	0.938	0.421	0.420	0.887	0.497	0.597
EfficientNet-B0	III	0.934	0.488	0.450	0.917	0.492	0.468
MobileViT-S	II	0.956	0.529	0.458	0.784	0.336	0.642
MobileViT-S	III	0.939	**0.535**	**0.501**	0.855	**0.371**	0.518
Mean		**0.945**	**0.497**	**0.448**	**0.872**	**0.442**	**0.582**

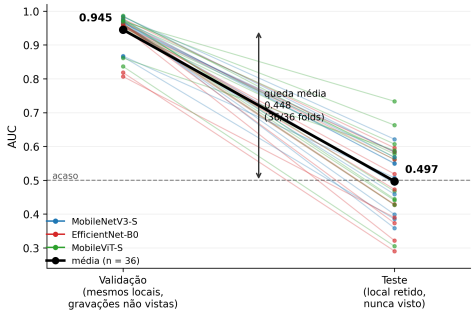


Figure 2. AUC on validation (unseen recordings from the same sites) and on the held-out site, for the 36 instances. Thin lines are models; the thick line is the mean. The validation AUC is an upper bound, being the epoch selection criterion, and the split was by recording rather than patient; neither affects the out-of-domain estimate.

4.2. Sanity of the Explanations

Grad-CAM maps mostly pass the sanity check: the mean SSIM between the original map and that of the randomized model is 0.341, and 76.3% of the 570 measurements fall below the 0.5 threshold, indicating dependence on the learned weights. No architecture–regime combination has a mean above the threshold, although MobileViT-S is consistently closest (0.407 under Protocol II and 0.441 under III, against 0.290–0.304 for the others). The measurement used the complete balanced sample, explaining the true class of each recording.

4.3. Fidelity: Grad-CAM versus Random Saliency

Table 3 presents the control we consider essential to any audit of this kind: do fidelity metrics distinguish the explanation from structured noise? The answer is affirmative and without reservation. In the six architecture–regime combinations, Grad-CAM obtains lower Deletion and higher Insertion than random saliency, with significance in all twelve comparisons after Holm adjustment and, more demandingly, with hierarchical confidence intervals by site excluding zero in every one of them.

4.4. The Disagreement Between Deletion and Insertion and Its Origin

Comparing the regimes under the measurement convention inherited from the literature, Deletion and Insertion point in opposite directions. We conclude that this is produced

Table 3. Grad-CAM against structured random saliency, $n = 95$ per cell. Δ is Grad-CAM minus random (Deletion: lower is better, so negative favours Grad-CAM; Insertion: higher is better). All twelve comparisons are significant after Holm adjustment ($p \leq 0.0024$) under paired Wilcoxon on the same sample, and all hierarchical 95% intervals exclude zero. The *f* column gives the number of the six sites in which the mean effect has the expected sign.

		Deletion			Insertion		
Architecture	Reg.	Δ	95% CI	f	Δ	95% CI	f
MobileNetV3-S	II	-0.139	$[-0.209; -0.058]$	5/6	$+0.131$	$[+0.041; +0.207]$	5/6
MobileNetV3-S	III	-0.174	$[-0.246; -0.097]$	6/6	$+0.173$	$[+0.104; +0.240]$	6/6
EfficientNet-B0	II	-0.066	$[-0.113; -0.017]$	5/6	$+0.114$	$[+0.036; +0.190]$	5/6
EfficientNet-B0	III	-0.078	$[-0.140; -0.009]$	5/6	$+0.139$	$[+0.074; +0.202]$	6/6
MobileViT-S	II	-0.065	$[-0.107; -0.029]$	6/6	$+0.071$	$[+0.031; +0.114]$	6/6
MobileViT-S	III	-0.069	$[-0.110; -0.028]$	6/6	$+0.067$	$[+0.010; +0.132]$	5/6

Table 4. Ablation of the measurement choices. Paired Wilcoxon per sample, Holm-adjusted within each block. $n = 47$ under the original measurement (normal samples only) and $n = 95$ otherwise. In bold, the two architectures in which *both* metrics decrease under the original measurement — the disagreement; and the only effect surviving corrected measurement, in the coherent direction.

		Deletion			Insertion		
Architecture	Measurement	II	III	Δ	II	III	Δ
MobileNetV3-S	original	0.677	0.388	-0.283	0.803	0.779	-0.023
	corrected	0.339	0.334	-0.005	0.604	0.685	$+$**0.079**
EfficientNet-B0	original	0.481	0.279	$-$**0.204**	0.599	0.470	$-$**0.128**
	corrected	0.460	0.429	-0.028	0.614	0.637	$+0.025$
MobileViT-S	original	0.446	0.350	$-$**0.097**	0.682	0.464	$-$**0.215**
	corrected	0.430	0.437	$+0.007$	0.568	0.587	$+0.018$

by two measurement choices combined with the model's operating point, and does not primarily follow from a property of the explanations. The first choice is which class to explain: the original measurement tracked the confidence always for the *abnormal* class, regardless of the sample's label. The second is the composition of the sample, restricted to the normal class. The combination measures the removal of evidence of abnormality in normal hearts, a regime in which P(abnormal) is low and the curves operate in a narrow range.

Table 4 isolates these factors. Under the original measurement the disagreement is clear and robust: in EfficientNet-B0 and MobileViT-S *both* metrics decrease with significance after Holm adjustment and with hierarchical intervals excluding zero. Under corrected measurement the pattern disappears: of six comparisons only the Insertion of MobileNetV3-S survives, in the *coherent* direction ($\Delta = +0.079$; Holm $= 0.022$; CI $[+0.006; +0.152]$).

Normalization by the operating point (Section 3.7) confirms the diagnosis by another route. Applied to the original measurement, it inverts the sign of Insertion in

MobileNetV3-S (from -0.023 to $+0.067$; $p = 0.014$) and in EfficientNet-B0 (from -0.128 to $+0.095$), while Deletion falls further (-0.464 and -0.264), making the two metrics coherent and both indicating Protocol III as better. Much of what appeared to be disagreement was the ceiling of the curves descending along with the model's confidence.

A third result delimits the role of the operating point. If Deletion (lower is better) and Insertion (higher is better) measured fidelity coherently, they should correlate negatively across samples. We observe the opposite: under corrected measurement ($n = 95$) Spearman's rho is positive in all six cells, between 0.283 and 0.656. Controlling for p_{full} by partial correlation, the association decreases but does not vanish — 0.377 and 0.228 for MobileNetV3-S (II, III), 0.359 and 0.328 for EfficientNet-B0, 0.504 and 0.085 for MobileViT-S — remaining significant in five of the six cells. The operating point therefore explains a substantial part of the association, but not all of it; additional factors, such as the redundancy of spectrographic evidence, likely contribute.

4.5. Localization (FAR)

FAR increases significantly for MobileNetV3-S ($\Delta = +0.058$, $p = 0.0013$) and MobileViT-S ($\Delta = +0.060$, $p = 0.0010$) and is neutral for EfficientNet-B0 ($\Delta = -0.003$, $p = 0.122$). Because of the mask limitation of Section 5, this reflects concentration on the energy region, not necessarily on the murmur structure; the effect is heterogeneous across folds and absent in fold *e*.

5. Discussion

Faithful explanations of a useless model. The models learn the task in distribution (AUC 0.945), do not discriminate murmurs out of domain (AUC 0.497) and are severely overconfident (confidence 0.87 against accuracy 0.45). Even so, their explanations mostly pass the sanity check and outperform structured random saliency on both fidelity metrics in every configuration. There is no contradiction: fidelity measures the relationship between the explanation and the model, not between the explanation and clinical reality. The map is faithful — to the wrong model. None of the XAI evaluation instruments currently used as quality seals would have flagged these models as inadequate for screening.

The drop between distributions is consistent with shortcut learning, and the contrast from 0.945 to 0.497 discards three competing explanations. Training instability and intrinsic task difficulty are excluded, since the validation AUC shows that training converges and the task is learnable from these spectrograms. Prevalence shift across sites — 8.5% abnormal at site *e* against 71.4% at site *a* — does not explain it either, because AUC is invariant to the class proportion: if only prevalence changed, keeping the conditional score distributions fixed, AUC would not move. Its collapse indicates that the representations themselves do not transfer. Two explanations remain that the current design cannot separate: general acoustic heterogeneity, and the small size of some test sets. Against the latter, site *e*, with 2,141 recordings, shows AUC 0.411 — below chance with a large sample. A direct demonstration would require a classifier of the site of origin, which we leave as future work. The protocol exposes a fragility that stratified validation would hide behind an AUC of 0.945.

Practical guidance under metric disagreement. Three operational rules follow. First, never select an explanation method or a training regime from a single fidelity metric: under the original measurement both metrics decrease significantly in two architectures, so a

single-metric report would support opposite conclusions depending on the metric chosen. Second, report the operating point: since both metrics are bounded by the confidence on the intact input, comparisons across regimes should normalize by it or report it alongside (Table 4). Third, treat disagreement as a diagnostic signal: divergence persisting after normalization would indicate a decision resting on distributed or redundant evidence, a regime in which compact saliency maps are poor summaries.

The qualitative examples of Figure 1 cohere with this picture: maps concentrate on the highest-energy bands and vary visibly across regimes, but the variation does not translate into a gain in discrimination (Table 2).

Limitation of energy-based localization. The per-band energy analysis showed strong concentration in a few bands, where murmur energy and normal heart-sound (S1/S2) energy overlap. The threshold-based mask therefore does not isolate the pathological structure, and FAR measures energy concentration rather than murmur localization. This is itself a contribution: seemingly objective localization metrics may be inadequate depending on the representation and should be validated before use. A promising alternative is to anchor the mask in cardiac-cycle components, as done by [Alrabie and Barnawi 2025], rather than in raw spectrogram energy.

Threats to validity. Six. (i) The validation AUC of Table 2 is an upper bound on in-distribution performance, since that set was the epoch selection criterion; the difference with respect to the test set is nonetheless too large to be produced by that bias. (ii) The train/validation split was made by recording, not by patient, as PhysioNet/CinC 2016 provides no subject identifier; part of the in-distribution performance may therefore reflect patient memorization. Neither affects the out-of-domain estimate, in which an entire site is held out. (iii) Regimes II and III still differ in two simultaneous aspects (Table 1), so effects cannot be attributed to the balanced sampler in isolation. (iv) We use a single public dataset and a single explainer; the restriction is deliberate, as PhysioNet/CinC 2016 is, to our knowledge, the only widely used PCG corpus with explicit site-provenance labels. (v) The Grad-CAM target layer differs across architectures, which may influence direct magnitude comparisons. (vi) The 95 samples are nested within only six sites and each cell is a single trained model; despite hierarchical bootstrap and multiplicity correction, the number of units at the level where variability occurs is small, and comparisons between regimes should be read as exploratory.

6. Conclusion

We presented and applied an interpretability auditing framework for visual cardiac murmur screening, combining LOGO-CV, Grad-CAM, the sanity check, Deletion/Insertion and a random-saliency control. Across 36 model instances the audit reveals a regime of general interest: models reaching AUC 0.945 in distribution fall to 0.497 at an unseen site, with severe overconfidence, while the usual explanation quality tests remain satisfied — sanity in 76.3% of maps and superiority over random saliency in all twelve fidelity comparisons. Explanation fidelity does not imply model usefulness. We further show that the disagreement between fidelity metrics across training regimes is produced by measurement choices combined with the operating point, disappearing under corrected measurement, and that energy-based localization is limited in this representation. The result argues for multi-metric, domain-stratified audits, conditioned above all on the prior

reporting of out-of-domain performance. Future work should extend the protocol to an external set such as CirCor DigiScope [Oliveira et al. 2022] and to other explainer families, and should address the design limitations exposed here: separating the factors that still distinguish the regimes, adopting patient-level partitioning where possible, and directly demonstrating shortcut learning through a classifier of the site of origin.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9525–9536, Montreal, QC, Canada.

Alkhodari, M. and Fraiwan, L. (2021). Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings. *Computer Methods and Programs in Biomedicine*, 200:105940.

Alrabie, S. and Barnawi, A. (2025). Are artificial intelligence models listening like cardiologists? Bridging the gap between artificial intelligence and clinical reasoning in heart-sound classification using explainable artificial intelligence. *Bioengineering*, 12(6):558.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, Lake Tahoe, NV, USA.

Clifford, G. D. et al. (2016). Classification of normal/abnormal heart sound recordings: The PhysioNet/CinC challenge 2016. In *2016 Computing in Cardiology Conference (CinC)*, pages 609–612, Vancouver, BC, Canada.

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619.

Geirhos, R. et al. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70*, pages 1321–1330, Sydney, Australia.

Howard, A. et al. (2019). Searching for MobileNetV3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, Seoul, Korea (South).

Kim, J., Park, G., and Suh, B. (2022). Classification of phonocardiogram recordings using vision transformer architecture. In *2022 Computing in Cardiology (CinC)*, volume 49, pages 1–4, Tampere, Finland.

Kim, S. et al. (2024). Artificial intelligence (AI) applications for point of care ultrasound (POCUS) in low-resource settings: A scoping review. *Diagnostics*, 14(15):1669.

Krishna, S. et al. (2024). The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*.

Krones, F. et al. (2024). From theoretical models to practical deployment: A perspective and case study of opportunities and challenges in AI-driven cardiac auscultation research for low-income settings. *PLOS Digital Health*, 3(12):e0000437.

Li, X., Ng, G. A., and Schlindwein, F. S. (2022). Transfer learning in heart sound classification using Mel spectrogram. In *2022 Computing in Cardiology (CinC)*, volume 49, pages 1–4, Tampere, Finland.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4765–4774, Long Beach, CA, USA.

Ma, S., Chen, J., and Ho, J. W. K. (2024). An edge-device-compatible algorithm for valvular heart diseases screening using phonocardiogram signals with a lightweight convolutional neural network and self-supervised learning. *Computer Methods and Programs in Biomedicine*, 243:107906.

Mehta, S. and Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In *10th International Conference on Learning Representations (ICLR)*.

Oliveira, J. et al. (2022). The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2524–2535.

Özcan, F. (2024). Rapid detection and interpretation of heart murmurs using phonocardiograms, transfer learning and explainable artificial intelligence. *Health Information Science and Systems*, 12(1):43.

Petsiuk, V., Das, A., and Saenko, K. (2018). RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the 29th British Machine Vision Conference (BMVC), paper 151*, Newcastle, UK.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, San Francisco, CA, USA.

Roth, G. A. et al. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study. *Journal of the American College of Cardiology*, 76(25):2982–3021.

Selvaraju, R. R. et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Venice, Italy.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70*, pages 3319–3328, Sydney, Australia.

Tan, M. and Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97*, pages 6105–6114, Long Beach, CA, USA.

Wang, H. et al. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 24–25, Seattle, WA, USA.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683.