AGRI-FIDELITY: EVALUATING THE RELIABILITY OF LISTENABLE EXPLANATIONS FOR POULTRY DISEASE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing XAI metrics measure faithfulness for a single model, ignoring model multiplicity where near-optimal classifiers rely on different or spurious acoustic cues. In noisy farm environments, stationary artifacts such as ventilation noise can produce explanations that are faithful yet unreliable, as masking-based metrics fail to penalize redundant shortcuts. We propose `AGRI-Fidelity`, a reliability-oriented XAI metric for listenable explanations in poultry disease detection without spatial ground truth. The method combines cross-model consensus with cyclic temporal permutation to construct null distributions and compute a False Discovery Rate (FDR), suppressing stationary artifacts. Across real and controlled datasets, `AGRI-Fidelity` complements masking-based metrics by distinguishing prediction sensitivity from cross-model reliability at the sample level.

1 INTRODUCTION

Animal disease detection is critical for livestock health monitoring and timely intervention. With the increasing availability of poultry audio from commercial farms, deep learning models have proven effective at identifying disease-related vocalization patterns (Yajie et al., 2023; Vidic et al., 2017). However, despite strong performance, these models remain black boxes, limiting adoption in agricultural settings where expert validation and trust are essential (Handelman et al., 2019).

Explainable AI (XAI) identifies acoustic features driving individual predictions. Listenable explanations let stakeholders hear important time-frequency regions, improving audio interpretability (Paissan et al., 2024; Wullenweber et al., 2022; Parekh et al., 2022). However, in noisy farms, faithful explanations may remain unreliable: persistent ventilation or machinery noise can serve as predictive shortcuts, leading XAI to highlight domain-irrelevant signals and create false interpretability. Near-optimal models may also use different acoustic cues, producing inconsistent explanations for the same input and weakening trust.

Why are existing XAI metrics insufficient for reliable evaluation? Existing XAI evaluation metrics (Fresz et al., 2024; Banerjee & Barnwal, 2022), including faithfulness (Li et al., 2023a) and fidelity-based variants (Miró-Nicolau et al., 2025), primarily assess whether an explanation aligns with a model's internal decision process. However, they do not quantify whether the reliance is task-relevant or driven by spurious correlations. Consequently, these metrics can assign high scores to explanations that faithfully reflect model behavior yet provide little assurance of reliability or domain validity. This gap is particularly problematic in safety-critical applications, where stakeholders require principled guarantees about explanation trustworthiness. Bioacoustic animal disease detection models may achieve high accuracy through confounders such as persistent background noises and microphone placement rather than physiologically meaningful vocalizations. Existing metrics assess internal consistency, not statistical validity, potentially rewarding explanations highlighting domain-irrelevant artifacts (Agrawal et al., 2026). Explanations may also differ across near-optimal models, revealing instability missed by single-model metrics (Marx et al., 2023).

Compounding this issue, obtaining ground-truth feature-importance annotations is challenging in real-world farm environments. Distance- or overlap-based metrics that assume precise spatial or temporal masks are therefore impractical for evaluating explanations from noisy, unconstrained bioacoustic data (Stodt et al., 2023; Kadir et al., 2023). These limitations underscore the need for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

an evaluation framework that moves beyond single-model faithfulness toward statistically grounded cross-model reliability assessment.

Our approach. We propose `AGRI-Fidelity`, a semantic reliance metric evaluating time-frequency evidence in poultry audio listenable explanations. It extends fidelity with reliability scoring through cross-model consensus, separating meaningful bioacoustic signals from stationary artifacts while integrating causal impact and statistical stability.

Key contributions. **(1)** We introduce a **reliability-centered evaluation paradigm** for bioacoustic disease XAI, distinguishing model-centric faithfulness from task-level trustworthiness and reframing single-model fidelity as a statistically validated cross-model property without spatial ground truth. **(2)** We propose `AGRI-Fidelity`, integrating **cross-model consensus** with fidelity-based causal validation. Stratifying agreement across architecturally diverse near-optimal models quantifies explanation stability within consensus beyond individual models. **(3)** We develop a **tier-specific permutation-based null construction** using cyclic temporal shifts to estimate empirical False Discovery Rates (FDR), statistically suppressing continuous stationary artifacts while preserving sparse, time-localized physiological signals without retraining or external annotations. **(4)** We provide **theoretical guarantees**: stationary artifacts asymptotically yield FDR $\to 1$ (complete suppression), whereas sparse time-locked signals yield FDR $\to 0$, formally establishing the formulation's robustness. **(5)** Extensive experiments on real-world and controlled poultry vocalization datasets demonstrate `AGRI-Fidelity`'s **reliable differentiation of genuine bioacoustic markers from faithful but domain-irrelevant artifacts**, outperforming standard masking-based metrics with stable committee-level behavior across random initializations. `AGRI-Fidelity` is, to our knowledge, the first framework to combine cross-model consensus with empirical FDR estimation specifically to identify spurious attributions and augment existing fidelity metrics with an explicit reliability score.

2 Related work

2.1 XAI evaluation metrics and their limitations

Single-Model fidelity metrics. In XAI, fidelity measures the causal relationship between identified features and model predictions. Traditional fidelity metrics mask features deemed important by an XAI method and quantify the resulting confidence change. Core variants include Average Increase (AI) and Average Drop (AD) (Chattopadhay et al., 2018). AI computes the fraction of samples for which target-class confidence increases when only the masked-in region is retained, whereas AD measures the average relative confidence decrease when the input is masked. Similarly, Average Gain (AG) (Zhang et al., 2024) evaluates the relative confidence increase after masking. To reduce local structure disruption, the Iterative Removal of Features (IROF) metric (Rieger & Hansen, 2020) groups inputs into meaningful segments and replaces flagged regions with a baseline value for efficient segment-level evaluation. However, because these metrics assess causal alignment within a single model, they do not account for spurious correlations or cross-model instability, motivating a reliability-oriented evaluation framework such as `AGRI-Fidelity`.

Explanation methods for bioacoustic and listenable interpretability. These fidelity metrics were largely developed and validated on vision or tabular tasks. Listenable methods such as Cough-LIME (Wullenweber et al., 2022) and other audio attribution techniques (Paissan et al., 2024; Parekh et al., 2022) adapt attribution to time-frequency spectrograms, but are typically evaluated using the same single-model fidelity metrics, inheriting their blindness to cross-model instability and stationary acoustic artifacts common in real-world farm audio.

In-Distribution fidelity evaluation. Standard masking generates out-of-distribution (OOD) inputs that can induce model shock and distort fidelity measurements. Remove and Retrain (ROAR) (Hooker et al., 2019) removes important features and retrains models to reassess feature importance at substantial computational cost. Remove and Debias (ROAD) (Rong et al., 2022) avoids retraining while mitigating mask-structure information leakage. F-Fidelity (Zheng et al., 2024) also corrects OOD-induced evaluation artifacts. Alternatively, Luss et al. (2021) generate contrastive explanations through latent-space projection, identifying in-distribution modifications that flip predictions. These approaches mitigate OOD bias but focus on single-model behavior, overlooking cross-model instability and statistically persistent spurious correlations.

Formalizing XAI evaluation robustness. Beyond empirical masking, several works formalize explanation evaluation mathematically. Infidelity (Yeh et al., 2019) grounds faithfulness in a Taylor expansion, requiring the explanation vector to approximate a local linear predictor of model behavior. Structural robustness metrics instead focus on explanation volatility. Sensitivity (Yeh et al., 2019) uses adversarial optimization to measure the maximum divergence between explanation maps for nearly identical inputs, while Evaluation Stability (Alvarez-Melis & Jaakkola, 2018) quantifies consistency via a local Lipschitz constant. Chalasani et al. (2020) further analyze faithfulness violations, showing that model smoothness through adversarial training is necessary for stable feature attribution. However, these formulations primarily assess local smoothness and perturbation stability within a single model and do not address cross-model disagreement or statistically persistent spurious correlations in real-world noisy environments. Beyond individual explanations, Tomsett et al. (2020) propose sanity checks for evaluation metrics, showing that common fidelity and faithfulness metrics can disagree or fail basic consistency tests. This meta-evaluation motivates `AGRI-Fidelity`: rather than assuming metric trustworthiness, we test whether observed cross-model consensus exceeds a statistically grounded null.

Distance and aggregation-based metrics. Approaches improve explanations through aggregation or controlled benchmarking. Bhatt et al. (2020) aggregate XAI methods for one model, producing smoother saliency maps. CLEVR-XAI (Arras et al., 2022) uses synthetic datasets with ground-truth masks for direct overlap-based evaluation. However, aggregation retains model-specific bias, while synthetic benchmarks with perfect masks are scarce in real-world bioacoustics, limiting distance-based evaluation in noisy environments.

2.2 Cross-Model consensus and its limitations

To address limitations of single-model evaluation, work on model multiplicity and the Rashomon set (Breiman, 2001; Fisher et al., 2019; Schwarzschild et al., 2023) argues that task-relevant features should generalize across near-optimal models. CAMO (Yu et al., 2024) measures inter-model agreement via prediction and logic consensus using approximate Shapley values, while Li et al. (2023b) aggregate committee explanations and compare them to an aggregate baseline. Conformal prediction has been used to construct explanation uncertainty sets with distribution-free variability bounds (Marx et al., 2023), and cross-explainer methods (Dietz et al., 2024; Thinn & Saeed, 2026) analyze agreement across explanation methods for a single model. However, these approaches measure overlap or bounded variability rather than statistical validity. Stationary background artifacts can yield high agreement across models or explainers and falsely appear robust, since similarity aggregation does not test whether consensus exceeds chance. This gap motivates `AGRI-Fidelity`, which uses explicit null modeling for statistically grounded consensus evaluation.

Overall, existing XAI metrics exhibit three core limitations in real-world bioacoustic settings. **(1) Spurious faithfulness.** Standard fidelity and in-distribution metrics ignore persistent spurious correlations (Fresz et al., 2024; Banerjee & Barnwal, 2022). When models exploit stationary artifacts, masking-based metrics validate this reliance rather than flagging it. ROAR (Hooker et al., 2019) mitigates masking bias but requires costly retraining. In contrast, `AGRI-Fidelity` uses cross-model consensus with permutation-based False Discovery Rate (FDR) to suppress artifact-driven agreement without retraining. **(2) Dependence on ground-truth masks.** Distance- and overlap-based metrics assume precise spatial or temporal annotations, which are rarely available in noisy farm audio. `AGRI-Fidelity` removes this requirement by constructing an empirical null via cyclic temporal permutation, enabling reliability assessment without ground-truth feature maps. **(3) Agreement without statistical validation.** Consensus-based methods quantify inter-model similarity but do not test whether agreement exceeds chance alignment. Persistent artifacts may therefore appear robust due to high overlap alone. `AGRI-Fidelity` integrates fidelity with tier-specific FDR analysis, ensuring consensus reflects statistically validated causal structure rather than coincidental similarity. Unlike conformal and bootstrap-based methods that quantify predictive uncertainty (Marx et al., 2023; Li et al., 2023b), `AGRI-Fidelity` evaluates statistical validity directly in attribution space, distinguishing artifact-driven agreement from statistically meaningful consensus. Our use of false-discovery terminology follows the multiple-testing literature (Benjamini & Hochberg, 1995), but the proposed quantity is not the Benjamini–Hochberg step-up procedure: it is a tier-wise empirical estimate, defined as the expected number of consensus bins under cyclic-shift permutations divided by the observed number of consensus bins.

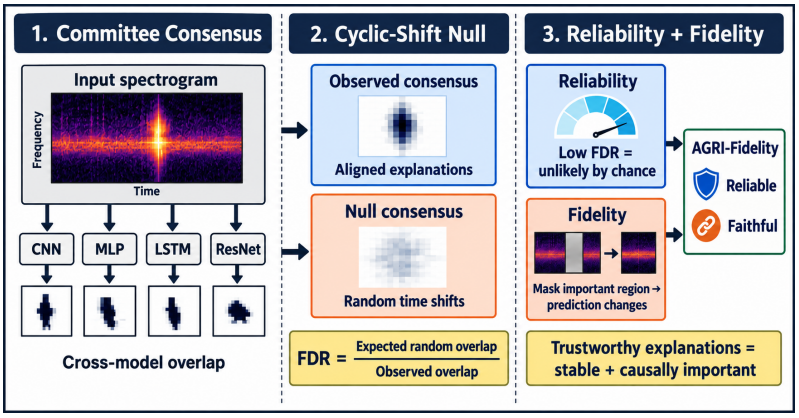


Figure 1: **AGRI-Fidelity framework overview.** The panels show cross-model consensus, the cyclic-shift null for reliability, and integration with masking-based fidelity, identifying explanations stable across models and influential to predictions.

3 Problem Statement

The primary challenge in evaluating feature attribution for bioacoustic disease detection is the presence of continuous stationary artifacts such as ventilation systems, that co-exist with the target signal. Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset of poultry vocalization recordings x_i with corresponding health labels $y_i \in \{\text{Healthy}, \text{Unhealthy}\}$ collected in real-world farm environments. Let $\mathcal{M} = \{M_1, \ldots, M_K\}$ denote a committee of trained classification models with comparable predictive performance on $\mathcal{D}$, and let $M^* \in \mathcal{M}$ be a designated target model within the committee whose explanation we wish to evaluate. Given an explanation method $\mathcal{E}$, we obtain an explanation map for an input x by identifying time-frequency regions $P^*(x) = \mathcal{E}(M^*, x)$ deemed important for the prediction. As formalized in Section 4, $\mathcal{E}$ is applied not only to M^* but to every model in $\mathcal{M}$, so that $P^*(x)$ can be assessed against the resulting committee-wide consensus.

The goal of our evaluation metric is to **determine whether $P^*(x)$ provides a reliable interpretation of the model's decision.** Existing XAI metrics primarily assess fidelity, whether masking $P^*(x)$ changes the prediction of M^*. However, *fidelity alone does not ensure reliability* for two reasons. **(1) Model multiplicity.** Near-optimal models $M_k \in \mathcal{M}$ may rely on different acoustic cues, yielding explanations $E(M_k, x)$ that vary substantially. An explanation faithful to M^* may therefore capture model-specific behavior rather than shared task-relevant structure. **(2) Spurious correlations in noisy environments.** Farm audio is noisy, enabling models to exploit persistent acoustic artifacts that are predictive but physiologically irrelevant. Explanations may faithfully highlight such features while reflecting **artifact-driven shortcuts** instead of genuine bioacoustic signals.

Therefore, a reliable explanation for audio-based poultry disease detection must satisfy three criteria: **(1) Fidelity.** It must identify time-frequency regions that causally influence the prediction. **(2) Stability.** It should be consistent across near-optimal models in the committee $\mathcal{M}$, reflecting task-relevant structure rather than continuous artifacts. **(3) Domain alignment.** It must emphasize physiologically meaningful bioacoustic markers over background noise, enabling expert validation.

Accordingly, this work seeks to develop an explanation evaluation framework that jointly quantifies causal fidelity, cross-model stability, and domain alignment, enabling systematic assessment of explanation reliability in noisy, real-world bioacoustic disease detection settings.

4 AGRI-Fidelity: Reliability-Oriented Evaluation Framework

Tiered cross-model consensus. Let $\mathcal{M} = \{M_1, \ldots, M_K\}$ denote a committee of K near-optimal classification models, where each M_k is a trained classifier operating on spectrogram inputs (in our setting, $K = 4$). Let $X \in \mathbb{R}^{T \times F}$ denote the log-magnitude spectrogram of an audio recording, where T represents the number of time frames and F the number of frequency bins.

Given X, an explainer $\mathcal{E}$ produces a continuous attribution map $\mathcal{E}(M_k, X) \in \mathbb{R}^{T \times F}$ for each model M_k. To isolate the most salient features, we retain a top-k fraction of the positive attribution values (default $k = 0.05$) and binarize the result, yielding a binary explanation mask $E_k \in \{0, 1\}^{T \times F}$

for each M_k, where $E_k(t,f) = 1$ denotes bin (t,f) as important and $E_k(t,f) = 0$ otherwise. We compute the pixel-wise observed consensus score

$$S_{\text{obs}}(t,f) = \frac{1}{K}\sum_{k=1}^{K} E_k(t,f), \tag{1}$$

which represents the proportion of models in the committee that identify time–frequency bin (t,f) as salient. Since each $E_k(t,f) \in \{0,1\}$, $S_{\text{obs}}(t,f)$ takes discrete values in $\{0, \frac{1}{K}, \ldots, 1\}$, where $S_{\text{obs}}(t,f) = 1$ indicates full agreement and lower values reflect partial or model-specific attribution. Because each E_k is binary, $S_{\text{obs}}(t,f)$ assumes discrete agreement levels corresponding to the exact number of models selecting bin (t,f). Let $\Lambda = \left\{\frac{1}{K}, \frac{2}{K}, \ldots, 1\right\}$ denote the set of non-zero consensus tiers, where $\lambda = \frac{m}{K}$ indicates agreement among m models. For $K = 4$, $\Lambda = \{0.25, 0.5, 0.75, 1.0\}$. To retain this agreement structure, we stratify S_{obs} into level-specific binary masks

$$C_\lambda(t,f) = \begin{cases} 1, & S_{\text{obs}}(t,f) \geq \lambda, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

for each $\lambda \in \Lambda$. This tiered construction preserves the distinction between weak and strong inter-model agreement, preventing high-consensus regions from merging with low-consensus, model-specific attributions, thereby enabling reliability evaluation at different confidence thresholds.

Permutation-based statistical validation. To determine whether cross-model consensus reflects meaningful structure rather than coincidental alignment, we construct an empirical null distribution using cyclic temporal permutation. Recall that T denotes the number of time frames in the spectrogram. For each of B permutation iterations, we independently shift each binary explanation mask $E_k \in \{0,1\}^{T\times F}$ along the time axis by a random offset $\Delta_k^{(b)} \sim \mathcal{U}\{0, \ldots, T-1\}$:

$$E_k^{(b)}(t,f) = E_k((t + \Delta_k^{(b)}) \bmod T, f), \tag{3}$$

where $b = 1, \ldots, B$. Temporal shifts preserve spectral structure and marginal statistics while destroying inter-model temporal alignment, which is the carrier of bioacoustic signals. For each permutation b, we compute the corresponding null consensus $S_{\text{null}}^{(b)}(t,f) = \frac{1}{K}\sum_{k=1}^{K} E_k^{(b)}(t,f)$, which represents agreement under random temporal alignment. Applying the same tiered stratification defined previously yields $C_\lambda^{\text{null},(b)}$ for each $\lambda \in \Lambda$. We then compute the Empirical False Discovery Rate (FDR) for each consensus tier:

$$\widehat{\text{FDR}}(\lambda) = \min\left\{1, \frac{\frac{1}{B}\sum_{b=1}^{B}\sum_{t,f} C_\lambda^{\text{null},(b)}(t,f)}{\sum_{t,f} C_\lambda(t,f)}\right\}. \tag{4}$$

Here, the numerator estimates the expected bins reaching agreement level λ under the null, and the denominator counts observed agreement; clipping at one keeps the estimate and $R = 1 - \widehat{\text{FDR}}$ in $[0,1]$. Stationary artifacts stay invariant under shifts, yielding $\widehat{\text{FDR}}(\lambda) \approx 1$, whereas sparse signals lose alignment under permutation, giving $\widehat{\text{FDR}}(\lambda) \approx 0$. Reliability is defined as $R(X) = 1 - \widehat{\text{FDR}}(\lambda_{\max})$, where $\lambda_{\max} = \max\{\lambda \in \Lambda : C_\lambda(X) \neq \varnothing\}$ is the highest populated observed-consensus tier. If all K models agree somewhere, $\lambda_{\max} = 1$; otherwise it falls back to the next-highest populated tier, keeping R well-defined.

Reliability integration via fidelity. Consensus alone does not guarantee predictive influence. A region may have low $\widehat{\text{FDR}}(\lambda_{\max})$ from stable cross-model alignment yet remain irrelevant to the prediction. We also incorporate fidelity. Let $X \setminus C_\lambda$ denote the input after removing bins in tier λ; fidelity measures whether masking these regions degrades model M_k's confidence, indicating model dependence on the selected regions under this perturbation. An explanation is reliable only when it combines low empirical FDR (agreement beyond chance) with high fidelity (prediction sensitivity under masking). The same protocol evaluates the explanation map $P^*(X) = \mathcal{E}(M^*, X)$ from any target listenable explainer $\mathcal{E}$ for a designated model $M^* \in \mathcal{M}$: we apply $\mathcal{E}$ to each committee model, yielding $\{E_k(X) = \mathcal{E}(M_k, X)\}_{k=1}^{K}$ (including $E_{M^*}(X) = P^*(X)$), form the tiered consensus C_λ and empirical FDR as above, and compute each M_k's fidelity under the consensus mask.

Joint reporting of fidelity and reliability. We report reliability and fidelity separately because they capture different properties. The implementation also stores their product as an optional summary, but both components should accompany it: fidelity measures masking sensitivity, whereas reliability measures cross-model alignment beyond the null. We report $R(X) = 1 - \widehat{\mathrm{FDR}}(\lambda_{\max})$ alongside fidelity. Figure 1 summarizes the framework; Appendix J provides a numerical example.

5 THEORETICAL GUARANTEES OF AGRI-FIDELITY

We use the term *guarantees* here in a limited sense. The theoretical results are not intended as the primary novelty of our work; rather, they serve as sanity checks that formally characterize the behavior of the cyclic-shift null in the two extreme cases targeted by the construction: a perfectly stationary mask and a perfectly aligned sparse signal. Under the assumptions that the cyclic shifts are approximately independent and sparsity ratio $\rho \ll 1$ (validated empirically in Section 7), the following results formalize the desirable properties of the proposed `AGRI-Fidelity` metric:

Theorem 1 (**Safety against stationary artifacts).** *If the committee models converge on a feature that is strictly stationary (time-invariant), the Consensus algorithm will asymptotically assign a False Discovery Rate (FDR) of* 1.0 *at full consensus, resulting in a Reliability Score of* 0.

Theorem 2 (**Sensitivity to sparse signals).** *If the committee models converge on a sparse, time-localized signal, the expected FDR at full consensus (*$\lambda = 1$*) approaches* ρ^{K-1} *(where* $\rho \ll 1$ *is the sparsity ratio), allowing the feature to be retained with high confidence.*

The two guarantees follow from comparing the number of observed consensus pixels with the number of consensus pixels expected under the cyclic-shift null model. For strictly stationary artifacts, cyclic permutations do not destroy inter-model overlap. The null consensus count equals the observed consensus count at every threshold λ. Substituting these equal counts into the empirical FDR definition gives $\mathrm{FDR}(\lambda) = 1$, and hence the reliability score $1-\mathrm{FDR}(\lambda)$ becomes zero. In contrast, for a time-localized signal occupying a fraction $\rho = \tau/T$ of the temporal axis, observed masks overlap on the true signal window, yielding a consensus count proportional to $F\tau$. Under independent cyclic shifts, the probability that all K shifted masks overlap at the same time-frequency location is $\approx \rho^K$, so the expected null count scales as $FT\rho^K$. Thus, $\mathrm{FDR} \approx \rho^{K-1}$ at full consensus, which vanishes for $\rho \ll 1$ and $K \geq 2$. Full derivations, including the finite-sample $+1$ smoothing term omitted from this asymptotic sketch, are given in Appendix A.

6 EXPERIMENTAL DESIGN & SETUP

Datasets. We evaluate on multiple datasets: the **Poultry dataset** (Aworinde et al., 2023), consisting of 96 kHz, 24-bit audio collected over 65 days and resampled to 22–24 kHz for our experiments; a **Poultry dataset with a spurious feature**, constructed by injecting high-frequency noise into all unhealthy samples of the original dataset (Aworinde et al., 2023). According to our analysis, we successfully injected the spurious feature into $\approx 66\%$ of our unhealthy samples and $\approx 12\%$ of our healthy samples already contained a similar feature; **SmartEars poultry** (Qiao et al., 2026), a 6,000-sample subset of five-second clips labeled Healthy, Unhealthy, or None, where we retain only Healthy and Unhealthy samples; **SmartEars poultry denoised**, the same dataset (Qiao et al., 2026) processed using Python's `noisereduce` and Harmonic-Percussive Source Separation (HPSS); and **SwineCough** (Ministry of Agriculture, Food and Rural Affairs, 2024), a clean dataset of dry and abdominal swine cough audio used to evaluate behavior under minimal background noise.

Baseline metrics. We compare `AGRI-Fidelity` with standard attribution-based metrics. **Faithfulness** (Parekh et al., 2022) evaluates whether explanation-selected features alone preserve the model's prediction, while **Fidelity** (Tomsett et al., 2020) measures the drop in prediction confidence when the explanation region is removed, capturing model reliance on highlighted features. **Average Increase (AI)**, **Average Drop (AD)**, and **Average Gain (AG)** (Chattopadhay et al., 2018; Zhang et al., 2024) quantify changes in model confidence under masking, reflecting whether explanations amplify or suppress predictive confidence across samples. **Sparseness** (Chalasani et al., 2020) measures whether predictions depend on a small set of salient features, whereas **Complexity** (Bhatt et al., 2020) computes the entropy of attribution scores to assess localization.

Table 1: Quantitative comparison on the Chicken dataset with a spurious feature using 52 samples.

Baseline	AI↑	AD↓	AG↑	Faithfulness↑
LMAC–TD	69.2308 ± 46.6041	5.9754 ± 19.4286	**50.9550 ± 40.4056**	0.0939 ± 0.2316
L2I	25.0000 ± 43.7237	9.3818 ± 14.2014	8.1596 ± 17.5258	0.3899 ± 0.3538
CoughLIME	**73.0769 ± 44.7888**	**1.6002 ± 5.0503**	46.0910 ± 37.1815	0.3966 ± 0.4516
AudioLIME	15.3846 ± 36.4321	8.2388 ± 11.0209	9.4607 ± 24.3363	**0.6832 ± 0.3995**

Baseline	Sparse↑	Comp.↓	AGRI-Fidelity	
			Fidelity↑	Reliability↑
LMAC–TD	**0.8284 ± 0.1362**	7.9551 ± 2.4117	**0.1018 ± 0.0420**	**0.7445 ± 0.2999**
L2I	0.0422 ± 0.0337	12.2492 ± 0.0518	0.1009 ± 0.1145	0.4859 ± 0.4007
CoughLIME	0.3727 ± 0.1651	12.6604 ± 0.4552	0.0425 ± 0.0450	0.5154 ± 0.3865
AudioLIME	0.2639 ± 0.0812	**2.1870 ± 0.1322**	0.0765 ± 0.0929	0.5263 ± 0.3982

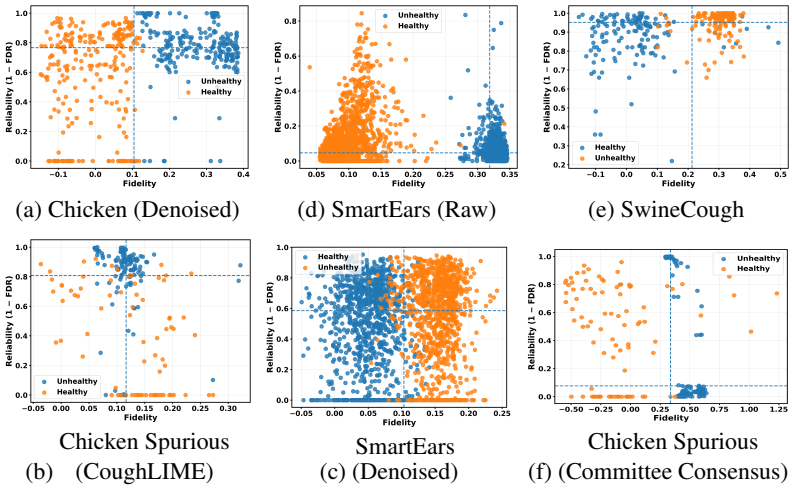
Baseline explainers. To evaluate AGRI-Fidelity, we consider four state-of-the-art listenable explanation methods. **(1) Listen to Interpret (L2I) (Parekh et al., 2022).** L2I generates sparse, class-discriminative time–frequency masks using non-negative matrix factorization to create explanations. **(2) Listenable Maps for Audio Classifier in Time Domain (LMAC-TD) (Mancini et al., 2025).** LMAC-TD is a post-hoc method that trains a decoder to produce explanations directly in the time domain. Both L2I and LMAC-TD were implemented using the SpeechBrain Toolkit (Ravanelli et al., 2021). **(3) CoughLIME (Wullenweber et al., 2022).** CoughLIME adapts LIME to audio by perturbing spectrogram regions and fitting a local surrogate model to explain cough-based predictions. **(4) AudioLIME (Haunschmid et al., 2020).** AudioLIME extends this LIME-based approach to general audio, producing localized explanations via surrogate modeling.

7 Experimental results & analyses

This section presents baseline explainer comparisons, consensus model performance across datasets, and analyses of FDR behavior and stability. **Interpretability with reliability scores.** Table 1 compares LMAC–TD, L2I, CoughLIME, and AudioLIME under standard masking metrics and AGRI-Fidelity. Under masking-based metrics, CoughLIME appears strongest in terms of Average Increase and Average Drop, achieving the highest AI (73.08) and lowest AD (1.60), while LMAC–TD achieves the highest AG (50.96). These results might suggest that CoughLIME and LMAC–TD provide the strongest explanations. However, masking metrics measure confidence changes within a single model and do not assess cross-model statistical validity. CoughLIME has the lowest fidelity (0.0425) and only moderate reliability (0.5154), despite its strong AI and AD. In contrast, LMAC–TD achieves the highest reliability (0.7445). This discrepancy shows that a large confidence response under masking does not necessarily imply that the highlighted regions are statistically stable across models. AGRI-Fidelity therefore provides information about explanation reliability that is not captured by masking-based metrics alone.

Decoupling causal influence from statistical reliability. AGRI-Fidelity's two components further distinguish causal influence from cross-model reliability. LMAC–TD and L2I obtain similar fidelity values (0.1018 and 0.1009, respectively), yet their reliability scores differ substantially (0.7445 and 0.4859). Similarly, CoughLIME achieves strong AI and AD but has the lowest fidelity (0.0425) and a reliability of only 0.5154. These differences demonstrate that fidelity and reliability capture complementary properties: fidelity measures the effect of the selected region on prediction confidence, whereas reliability measures whether the highlighted structure is consistently supported across the model committee beyond chance alignment.

Structurally inconsistent attributions analysis. AGRI-Fidelity is sensitive not only to the concentration of an attribution map but also to its structural alignment across the model committee. L2I achieves fidelity comparable to LMAC–TD (0.1009 versus 0.1018), but it obtains the lowest reliability (0.4859). CoughLIME produces greater sparseness than L2I and AudioLIME (0.3727), yet its reliability remains moderate (0.5154). In contrast, LMAC–TD combines the greatest sparseness (0.8284) with the highest reliability (0.7445). These results show that neither causal influence nor localization alone guarantees reliable consensus. Explanations whose salient regions are fragmented,



(a) Chicken (Denoised) (d) SmartEars (Raw) (e) SwineCough

(b) Chicken Spurious (CoughLIME) (c) SmartEars (Denoised) (f) Chicken Spurious (Committee Consensus)

Figure 2: **Fidelity–Reliability quadrant distributions across datasets.** Panels (a)–(e) show per-sample `AGRI-Fidelity` components using CoughLIME, while panel (f) uses the committee-consensus explanation. Dashed lines mark median fidelity and reliability, dividing each panel into four interpretative quadrants.

Table 2: **FDR trend counts by class.**

Class	Total	Downward	Near-1
Healthy	139	120	19
Unhealthy	121	40	81

Table 3: **Mean fidelity by class and FDR trend.**

Class	Trend	CNN	LSTM	MLP	ResNet
Healthy	Downward	0.1589	0.0206	0.0457	0.0159
	Near-One	0.121	0.0319	0.0137	0.0175
Unhealthy	Downward	0.5161	0.2194	0.1833	0.4263
	Near-One	0.7994	0.4519	0.7040	0.8542

method-specific, or poorly aligned with the committee are reproduced more frequently under the cyclic-shift null model and consequently receive lower reliability. `AGRI-Fidelity` therefore penalizes structurally inconsistent attribution patterns even when they produce competitive fidelity or conventional masking scores, ensuring that reliability reflects statistically validated committee-level agreement rather than isolated saliency responses.

Understanding the high standard deviations. (Table 1) The relatively large standard deviations primarily reflect heterogeneity across test recordings rather than uncertainty in the estimated means. For AI, the large dispersion follows directly from its binary sample-level definition. Each sample contributes either 0 or 100, and a method with success probability p has an expected standard deviation of approximately $100\sqrt{p(1-p)}$. For example, CoughLIME achieves AI on 38 of the 52 samples ($p = 0.7308$), yielding an expected standard deviation of approximately 44.4, consistent with the observed value of 44.79. Statistical analyses for all datasets is given in Appendix H.2.

7.1 Fidelity–Reliability distributions across datasets

Limitations of one-dimensional fidelity evaluation. Comparing Chicken Denoised and Chicken Spurious reveals a limitation of fidelity-only evaluation: spurious Unhealthy artifacts attain fidelity scores nearly indistinguishable from genuine Denoised signals. By adding Reliability as an orthogonal dimension, `AGRI-Fidelity` resolves this ambiguity, shifting artifact-driven explanations from the high-quality quadrant to the low-reliability regime.

Class-Specific decoupling under spurious injection. In Figure 2(b), Unhealthy samples show low fidelity but high reliability. Since the injected spurious cue is continuous and stationary, CoughLIME's local perturbations evaluate limited temporal neighborhoods rather than the full background, causing recordings to emphasize sparse acoustic events instead of the global artifact. Near-zero FDR values dominate consensus, producing high reliability. Figure 2(f) isolates consensus-model reliability: Unhealthy samples show higher fidelity but mostly low reliability, consistent with Table 2.

Diagnosing shortcut learning in real-world farm audio. In SmartEars (Denoised), explanations span high- and low-reliability regions with structured separation. In SmartEars (Raw), they collapse toward low reliability despite moderate-to-high fidelity, indicating shortcut learning from persistent barn noise rather than disease-specific events. `AGRI-Fidelity` exposes this behavior, whereas masking metrics alone suggest acceptable performance. Moreover, models often achieve high accuracy by finding a feature that exists only (or primarily) in one class. If the metric gave the same score to all classes, it would be validating a spurious correlation.

Preservation of sparse, time-localized biological signals. On SwineCough, which contains isolated transient cough events with minimal background noise, the model committee converges on time-localized regions, yielding reliability scores between 0.9 and 1.0. This shows that `AGRI-Fidelity` preserves sparse, biologically meaningful signals without penalization.

7.2 Consensus stability and FDR trend analysis

Defining FDR trend types. For each sample, we classify $\mathrm{FDR}(\lambda)$ across populated tiers $\lambda \in \Lambda$. A curve is *Near-1* if $\mathrm{FDR}(\lambda_{\max}) \geq 0.8$: null and observed consensus remain similar at the strictest tier, indicating a stationary artifact (Theorem 1). Otherwise, it is *Downward*, with $\mathrm{FDR}(\lambda_{\max}) < 0.8$, indicating that FDR falls as the consensus threshold tightens and is consistent with a time-localized signal whose cyclic-shift overlap vanishes rapidly (Theorem 2). Thus, the labels summarize behavior at the highest populated consensus tier rather than requiring strict monotonicity at every intermediate tier. We apply the same 0.8 cutoff uniformly across all datasets and models reported in Tables 2 and 3. Consensus-model performance appears in Appendix H.1.

Table 2 summarizes FDR trend distributions across classes. Using the Near-1 trend to identify spurious behavior (Unhealthy as ground truth) yields 81 TP, 19 FP, 40 FN, and 120 TN, corresponding to 81.0% precision, 66.2% recall, 86.3% specificity, and 77.3% accuracy. Based on the success rates of our spurious feature injection, these results show that the Near-1 FDR trend detects artifact-dominated explanations with high precision, specificity, and recall. Table 3 reports mean fidelity by class and FDR trend. For Unhealthy samples, fidelity remains higher than for Healthy samples across models and trends, indicating committee agreement on the injected continuous artifacts.

Implications of high fidelity under Near-One consensus. For Unhealthy samples, Near-One samples have *higher* fidelity than Downward samples (Table 3; 0.7994 vs. 0.5161 for CNN and 0.8542 vs. 0.4263 for ResNet). Yet high FDR indicates reliance on the injected artifact: masking it sharply reduces confidence. Thus, fidelity alone cannot distinguish genuine signals from shortcuts. In deployment, high-fidelity, high-FDR explanations warrant scrutiny rather than trust. Additional Audio Spectrogram Transformer (AST) (Gong et al., 2021) analysis appears in Appendix I.

Computational cost. `AGRI-Fidelity` operates post hoc on pre-trained models. Computing fidelity requires two forward passes per consensus model, while the permutation-based null construction has complexity $\mathcal{O}(B \cdot K \cdot T \cdot F)$. These operations are vectorized and efficient in practice. Although the overall runtime exceeds the lightweight masking metrics (AI, AD, AG), it remains substantially lower than ROAR, which retrained models for multiple masking thresholds ($\geq K/2$). We have given details about our design rationale for this work in Appendix B.

8 Conclusions and future work

We introduce `AGRI-Fidelity`, a reliability-aware framework for bioacoustic disease explanations. Its cyclic-shift null models chance agreement, separating time-localized physiological cues from stationary artifacts. Across architectures, `AGRI-Fidelity` highlights distinct, consensus-supported indicators. Although single-model fidelity varies across seeds, committee-level directionality remains stable, demonstrating robust consensus evaluation.

Future work should evaluate the framework across additional species, environments, and model families and use expert-annotated bioacoustic events to assess whether consensus-supported regions are clinically meaningful. It should also investigate in-distribution masking and null constructions for nonstationary background noise.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, including improving clarity, correcting grammatical errors, and summarizing or condensing portions of the manuscript to satisfy page limitations. They were also used to obtain general explanations of selected XAI and statistical concepts during development. AI-generated responses were not treated as scholarly sources, and the tools were not relied upon to conduct the literature review, identify the research gap, formulate the central research questions, or independently design the methodology.

Generative AI tools additionally assisted with limited code adaptation and workflow automation. This included creating reproducible data-split utilities, adapting existing implementations to use common train, validation, and test partitions, automating previously implemented experimental procedures, compiling the results reported in Table 1, and producing selected supplementary analyses. The tools did not independently generate or fabricate experimental data. All experiments were executed using the authors' implementations and computing environments. The authors inspected the generated code, verified the data partitions and sample alignment, reviewed the statistical procedures, and checked the resulting outputs against the underlying per-sample data.

All AI-assisted text, code, analyses, and outputs were critically reviewed and revised by the authors. The authors take full responsibility for the methodology, experimental results, interpretations, claims, citations, and final content of this work, including any material produced with the assistance of generative AI.

ETHICS STATEMENT

This work aims to improve the reliability and transparency of explainable AI systems for audio-based animal-health monitoring. The proposed `AGRI-Fidelity` framework is intended for legitimate applications such as identifying unreliable explanations, detecting reliance on spurious acoustic artifacts, and supporting expert assessment of bioacoustic disease-detection systems. Although our evaluation introduces and analyzes spurious features, these artifacts are used solely to examine the limitations of existing explanation metrics and validate the proposed reliability framework. We do not advocate the deliberate introduction of such artifacts into deployed systems.

`AGRI-Fidelity` should not be considered a replacement for veterinary expertise, clinical testing, animal-welfare assessment, or careful validation of the underlying classifier. Explanation reliability does not guarantee diagnostic correctness, physiological relevance, or suitability for deployment. False predictions or misleading explanations could result in delayed treatment, unnecessary intervention, or inappropriate animal-management decisions. Responsible deployment should therefore combine `AGRI-Fidelity` with expert review, representative data collection, external validation, uncertainty assessment, and continuous monitoring for dataset shift and environmental confounders. Where farm audio may capture workers or other identifiable information, data collection and storage should additionally follow applicable consent, privacy, access-control, and data-governance requirements. The framework is intended to support, rather than replace, accountable human decision-making in animal-health applications.

REPRODUCIBILITY STATEMENT

To support independent verification and reproduction, an anonymized repository containing all materials required to reproduce this work has been provided with the submission. The repository includes the complete implementation of `AGRI-Fidelity`, all consensus-model and baseline explainer implementations, data-preprocessing and spectrogram-generation scripts, fixed train–validation–test split manifests, split-generation utilities, random seeds, training configurations, and evaluation scripts.

We additionally provide the trained consensus and baseline models, model checkpoints, training histories, learning curves, classification accuracies, generated explanation maps, per-sample metric files, reliability and FDR outputs, alignment audits, paired statistical-test results, and the scripts used to reproduce all tables and figures in the main paper and supplementary material. The supplementary

material documents the model architectures, hyperparameters, preprocessing procedures, evaluation settings, software dependencies, file structures, and implementation details.

The anonymized repository also contains a step-by-step reproduction guide with the commands required to execute each stage of the pipeline, including data preparation, split generation, model training, explanation generation, consensus analysis, baseline evaluation, statistical testing, and result aggregation. Information identifying the public datasets, including their sources, versions, and applicable licenses, is also provided. Dataset files are included or linked according to the redistribution terms of their respective licenses.

Upon acceptance, we will remove the repository anonymization and make the complete materials publicly available in a versioned archival repository. The public release will preserve the code, configurations, trained models, data partitions, seeds, intermediate outputs, and documentation supplied with the anonymized submission, enabling the reported results to be reproduced from the original public datasets.

REFERENCES

Arun Agrawal, Srinivasan Nagaraj, S Sagar Imambi, Kalyani Tushar Borse, Pratik Gulab Dhange, Veer Sudheer Goud, and Deepak Gupta. Barriers and limitations in artificial intelligence for decoding non-human signals. In *Harnessing AI for Human-Animal Communication*, pp. 85–116. IGI Global Scientific Publishing, 2026.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Leila Arras, Ahmed Osman, and Wojciech Samek. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022.

H Aworinde, S Adebayo, A Akinwunmi, O Alabi, A Ayandiji, O Oke, A Oyebamiji, A Adeyemo, A Sakpere, and K Echetama. Poultry vocalization signal dataset for early disease detection. *Mendeley Data*, 1, 2023.

Puja Banerjee and Rajesh P Barnwal. Methods and metrics for explaining artificial intelligence models: A review. *Explainable AI: Foundations, methodologies and applications*, pp. 61–88, 2022.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pp. 1383–1391. PMLR, 2020.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Katharina Dietz, Mehrdad Hajizadeh, Johannes Schleicher, Nikolas Wehner, Stefan Geißler, Pedro Casas, Michael Seufert, and Tobias Hoßfeld. Agree to disagree: Exploring consensus of xai methods for ml-based nids. In *2024 20th International Conference on Network and Service Management (CNSM)*, pp. 1–7. IEEE, 2024.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Benjamin Fresz, Lena Lörcher, and Marco Huber. Classification metrics for image explanations: towards building reliable xai-evaluations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1–19, 2024.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Guy S Handelman, Hong Kuan Kok, Ronil V Chandra, Amir H Razavi, Shiwei Huang, Mark Brooks, Michael J Lee, and Hamed Asadi. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology*, 212(1): 38–43, 2019.

Verena Haunschmid, Ethan Manilow, and Gerhard Widmer. audiolime: Listenable explanations using source separation. *arXiv preprint arXiv:2008.00582*, 2020.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. Evaluation metrics for xai: A review, taxonomy, and practical applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pp. 000111–000124. IEEE, 2023.

Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. M4: A unified xai benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. *Advances in Neural Information Processing Systems*, 36:1630–1643, 2023a.

Xuhong Li, Haoyi Xiong, Siyu Huang, Shilei Ji, and Dejing Dou. Cross-model consensus of explanations and beyond for image classification models: an empirical study. *Machine Learning*, 112 (5):1627–1662, 2023b.

Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Yunfeng Zhang, Karthikeyan Shanmugam, and Chun-Chen Tu. Leveraging latent features for local explanations. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 1139–1149, 2021.

Eleonora Mancini, Francesco Paissan, Mirco Ravanelli, and Cem Subakan. Lmac-td: Producing time domain explanations for audio classifiers. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Charles Marx, Youngsuk Park, Hilaf Hasson, Yuyang Wang, Stefano Ermon, and Luke Huan. But are you sure? an uncertainty-aware perspective on explainable ai. In *International conference on artificial intelligence and statistics*, pp. 7375–7391. PMLR, 2023.

Ministry of Agriculture, Food and Rural Affairs. Pig cough audio dataset. `https://www.smartfarmkorea.net/unstructuredDataset/pigCoughSound.do?menuId=M11040204`, 2024. Smart Farm Korea. Accessed: 2026-02-22.

Miquel Miró-Nicolau, Antoni Jaume-i Capó, and Gabriel Moyà-Alcover. A comprehensive study on fidelity metrics for xai. *Information Processing & Management*, 62(1):103900, 2025.

Francesco Paissan, Luca Della Libera, Mirco Ravanelli, and Cem Subakan. Listenable maps for zero-shot audio classifiers. *Advances in Neural Information Processing Systems*, 37:65361–65385, 2024.

Jayneel Parekh, Sanjeel Parekh, Pavlo Mozharovskyi, Florence d'Alché Buc, and Gaël Richard. Listen to interpret: Post-hoc interpretability for audio networks with nmf. *Advances in Neural Information Processing Systems*, 35:35270–35283, 2022.

Huaxin Qiao, Zixin Chen, Chengming Ji, Jianfeng Gao, Xiaoyang Xue, Shouchang Zhou, Huanliang Xu, and Junxian Huang. Smartears: A practical framework for poultry respiratory monitoring via spectrogram-based audio classification and ai-assisted labeling. *Computers and Electronics in Agriculture*, 241:111241, 2026.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

Laura Rieger and Lars Kai Hansen. Irof: a low resource evaluation metric for explanation methods. In *Workshop AI for Affordable Healthcare at ICLR 2020*, 2020.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. Evaluating feature attribution: An information-theoretic perspective. *CoRR*, 2022.

Avi Schwarzschild, Max Cembalest, Karthik Rao, Keegan Hines, and John Dickerson. Reckoning with the disagreement problem: Explanation consensus as a training objective. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 662–678, 2023.

Jan Stodt, Christoph Reich, and Nathan Clarke. A novel metric for xai evaluation incorporating pixel analysis and distance measurement. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1–9. IEEE, 2023.

Kay Thari Thinn and Waddah Saeed. Consensus and divergence in explainable ai (xai): Evaluating global feature-ranking consistency with empirical evidence from solar energy forecasting. *Mathematics*, 14(2):297, 2026.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6021–6029, 2020.

Jasmina Vidic, Marisa Manzano, Chung-Ming Chang, and Nicole Jaffrezic-Renault. Advanced biosensors for detection of pathogens related to livestock and poultry. *Veterinary research*, 48(1): 11, 2017.

Anne Wullenweber, Alican Akman, and Björn W Schuller. Coughlime: Sonified explanations for the predictions of covid-19 cough classifiers. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1342–1345. IEEE, 2022.

Liu Yajie, Md Gapar Md Johar, and Asif Iqbal Hajamydeen. Poultry disease early detection methods using deep learning technology. *Indonesian Journal of Electrical Engineering and Computer Science*, 32(3):1712–1723, 2023.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

Andy Yu, Parke Godfrey, Lukasz Golab, Divesh Srivastava, and Jaroslaw Szlichta. Camo: Explaining consensus across models. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 5493–5496. IEEE, 2024.

Hanwei Zhang, Felipe Torres, Ronan Sicre, Yannis Avrithis, and Stephane Ayache. Opti-cam: Optimizing saliency maps for interpretability. *Computer Vision and Image Understanding*, 248: 104101, 2024.

Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. F-fidelity: A robust framework for faithfulness evaluation of explainable ai. *arXiv preprint arXiv:2410.02970*, 2024.

A THEORETICAL GUARANTEES OF AGRI-FIDELITY

In this section, we formally justify the Consensus-based metric. We prove that the cyclic-shift null model (i) provably suppresses strictly stationary artifacts by yielding maximal False Discovery Rate (FDR), and (ii) remains highly sensitive to sparse, time-localized bioacoustic signals.

A.1 CONSENSUS FORMULATION AND CYCLIC-SHIFT NULL MODEL

Let $X \in \mathbb{R}^{T\times F}$ denote the input spectrogram with T temporal frames and F frequency bins. For a committee of K diverse models, let $E_k \in \{0,1\}^{T\times F}$ represent the binary importance mask produced by the k-th model (consistent with the notation of Section 4). The observed committee consensus at pixel (t, f) is defined as $S_{\text{obs}}(t, f) = \frac{1}{K}\sum_{k=1}^{K} E_k(t, f)$. To quantify accidental agreement among models, we construct a permutation-based null distribution using independent cyclic time shifts. Specifically, for each model k and permutation b, we sample $\Delta_k^{(b)} \sim \mathcal{U}\{0, \dots, T-1\}$ (a discrete uniform offset over the T time frames, matching Eq. 3 in Section 4) and apply the shift along the temporal axis to obtain $E_k^{(b)}(t, f) = E_k((t + \Delta_k^{(b)}) \bmod T, f)$. The corresponding null consensus is $S_{\text{null}}^{(b)}(t, f) = \frac{1}{K}\sum_{k=1}^{K} E_k^{(b)}(t, f)$.

A.2 GUARANTEE I: SAFETY AGAINST STATIONARY ARTIFACTS (PROOF OF THEOREM 1)

We prove Theorem 1 (Section 5), restated here for reference:

Theorem 1. If the committee models converge on a feature that is strictly stationary (time-invariant), the Consensus algorithm asymptotically assigns FDR $= 1.0$ at full consensus, giving Reliability Score 0.

Proof. **Assumption of stationarity.** Suppose the detected feature is strictly stationary, meaning the importance mask E_k does not vary with time t. Formally, for all models k, we assume $E_k(t, f) = \phi_k(f)$, where $\phi_k(f) \in \{0, 1\}$ depends only on the frequency bin f and is constant across all temporal indices.

Observed count (N_{obs}). Under this assumption, the consensus score at each frequency bin is identical across time. For any consensus threshold $\lambda \in (0, 1]$, the total number of observed positive pixels across the spectrogram is therefore:

$$\begin{aligned} N_{obs}(\lambda) &= \sum_{t=1}^{T}\sum_{f=1}^{F} \mathbf{1}\left(\frac{1}{K}\sum_{k=1}^{K}\phi_k(f) \geq \lambda\right) \\ &= T\sum_{f=1}^{F} \mathbf{1}\left(\bar{\phi}(f) \geq \lambda\right), \end{aligned} \tag{5}$$

where $\bar{\phi}(f) = \frac{1}{K}\sum_{k=1}^{K}\phi_k(f)$ denotes the average committee agreement at frequency f. The factor T arises because each qualifying frequency bin contributes equally across time.

Null count (N_{null}). Consider the temporally shifted mask $E_k((t + \Delta_k) \bmod T, f)$. Because the mask is time-invariant, we have $E_k((t+\Delta_k) \bmod T, f) = \phi_k(f)$ for all shifts Δ_k. Thus, the cyclic permutation does not change the mask structure. Consequently, the shifted consensus equals the observed consensus:

$$S_{\Delta,obs}(t, f) = \frac{1}{K}\sum_{k=1}^{K}\phi_k(f) = S(t, f). \tag{6}$$

It follows that the expected null count matches the observed count for every permutation sample:

$$\mathbb{E}[N_{null}(\lambda)] = N_{obs}(\lambda). \tag{7}$$

FDR and reliability calculation. The final FDR now becomes

$$\text{FDR}(\lambda) = \frac{\mathbb{E}[N_{null}(\lambda)]}{N_{obs}(\lambda)} = 1. \tag{8}$$

Therefore, the Reliability Score evaluates to $R = \max(0, 1 - \text{FDR}) = 0$, meaning the stationary artifact is completely suppressed by the consensus procedure. □

A.3 GUARANTEE II: SENSITIVITY TO SPARSE SIGNALS (PROOF OF THEOREM 2)

We prove Theorem 2 (Section 5), restated here for reference:

Theorem 2. If the committee models converge on a sparse, time-localized signal, the expected FDR at full consensus ($\lambda = 1$) approaches ρ^{K-1} (where $\rho \ll 1$ is the sparsity ratio), allowing the feature to be retained with high confidence.

Proof. **Assumption of sparsity and alignment.** Let the true biological signal exist strictly within a brief time window of duration τ, such that $\tau \ll T$. We define the sparsity ratio as $\rho = \tau/T$. Assume perfect committee alignment: all K models correctly identify the feature, setting $E_k(t, f) = 1$ inside the window and 0 elsewhere.

Observed count (N_{obs}). Given perfect agreement inside the temporal window, the observed count for maximum consensus ($\lambda = 1.0$) is:

$$N_{obs}(1.0) = F \cdot \tau. \tag{9}$$

Null count (N_{null}). Under the null distribution, independent random shifts $(t + \Delta_k)$ are applied. For a pixel (t, f) to achieve full consensus ($\lambda = 1.0$) in the shifted state, all K independent masks must stochastically align at that pixel. The marginal probability that model k's shifted mask is active at time t is exactly its sparsity ratio ρ. Assuming the shifts are independent and identically distributed, the joint probability of overlap is: $\mathbb{P}(S_\Delta(t, f) = 1.0) = \rho^K$. The expected number of false discoveries over the entire spectrogram is therefore: $\mathbb{E}[N_{null}(1.0)] = F \cdot T \cdot \rho^K$.

FDR calculation. The empirical FDR used throughout this paper (Eq. 4, Section 4) is a ratio of finite-sample counts, $\mathrm{FDR}(1.0) = \mathbb{E}[N_{null}(1.0)]/N_{obs}(1.0)$, with no smoothing term; the following derivation is exact for this quantity and does not require a $+1$ correction. (A $+1$-smoothed variant, sometimes used to keep FDR estimates finite when $N_{obs} = 0$, would instead satisfy $\mathrm{FDR}(1.0) = (\mathbb{E}[N_{null}(1.0)] + 1)/(N_{obs}(1.0) + 1)$; because $\mathbb{E}[N_{null}(1.0)] = FT\rho^K \ll 1$ under sparsity, that alternative's $+1$ term would dominate the numerator and the bound below would not apply to it directly.) Substituting the expected counts yields:

$$\mathrm{FDR}(1.0) \approx \frac{\mathbb{E}[N_{null}(1.0)]}{N_{obs}(1.0)} = \frac{F \cdot T \cdot \rho^K}{F \cdot \tau} = \frac{T \cdot \rho^K}{\tau}. \tag{10}$$

Recognizing that $\tau/T = \rho$, this simplifies to:

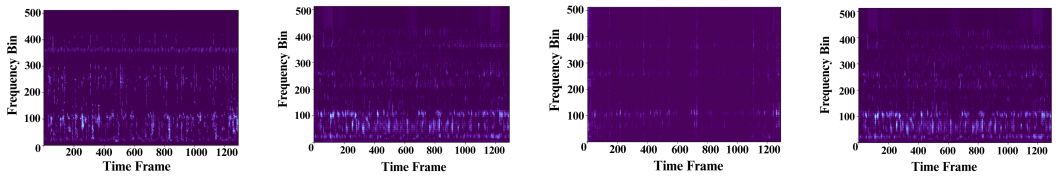
$$\mathrm{FDR}(1.0) \approx \frac{\rho^K}{\rho} = \rho^{K-1}. \tag{11}$$

Because the signal is sparse ($\rho \ll 1$) and $K \geq 2$, the term ρ^{K-1} becomes negligible. Consequently, the Reliability Score converges to $R \approx 1 - \rho^{K-1} \approx 1$, preserving the time-locked biological marker. This bound is derived for the full-consensus tier $\lambda = 1.0$; at a fractional tier $\lambda = m/K$ with $m < K$, the corresponding null-alignment probability scales instead with ρ^{m-1} under the same independence assumption, which we do not evaluate further here. □

B METHODOLOGICAL CONSIDERATIONS AND DESIGN RATIONALE

We outline the following six methodological considerations that clarify the statistical foundation and design principles of `AGRI-Fidelity`.

(1) Choice of auditing explainer. As formalized in Section 4, the explainer $\mathcal{E}$ being evaluated (e.g., LMAC-TD, L2I, CoughLIME, or AudioLIME) is itself applied to every committee model to form its own tiered consensus, FDR, and fidelity, this is the procedure underlying Tables 1 in the main paper and 9, 10, 11 in the Appendix sections H.2.1, H.2.3, H.2.4, and H.2.2. Integrated Gradients (IG) is used for a separate, complementary purpose: as the default explainer in the reliability-generation pipeline (Appendix E.6) and to visualize raw cross-model attribution agreement within the committee itself (Appendix C, Figure 3), independent of which target explainer is being audited. We favor IG for this internal, explainer-agnostic role because, unlike perturbation-based methods (e.g.,



(a) CNN (b) MLP (c) LSTM (d) ResNet

Figure 3: **Cross-Model Attribution Consistency:** Integrated gradients for a healthy sample in the denoised poultry dataset across CNN, MLP, LSTM, and ResNet consensus models.

LIME), which are sensitive to sampling variance and surrogate instability, IG satisfies completeness and implementation invariance, so residual committee disagreement in this internal diagnostic reflects differences in learned representations rather than explainer-induced noise.

(2) Explainer-agnostic evaluation. Because `AGRI-Fidelity` evaluates the final $T \times F$ attribution mask independently of how it is generated, the framework audits the statistical reliability of the resulting representation whether it was derived from gradients, decoding, or surrogate fitting, this is what makes the per-explainer comparisons in Tables 1 in the main paper and 9, 10, 11 an apples-to-apples comparison across LMAC-TD, L2I, CoughLIME, and AudioLIME.

(3) FDR vs. overlap metrics. We employ False Discovery Rate (FDR) rather than IoU, Precision, or Recall. Overlap-based measures ignore the null distribution and may reward agreement on stationary artifacts. By constructing a cyclic temporal permutation null, FDR quantifies whether consensus exceeds chance alignment, providing a statistically validated reliability criterion.

(4) Sampling-rate heterogeneity. Farm sensors operate at heterogeneous sampling rates, making separate training pipelines impractical. We standardize recordings via resampling while recognizing that aggressive downsampling may suppress high-frequency disease signatures.

(5) Absence of spatial ground truth. Bioacoustic datasets lack pixel-level annotations of disease-relevant regions in time–frequency space. In this setting, cross-model consensus under a permutation-based null serves as a conservative proxy for reliability without requiring explicit ground-truth masks.

(6) Continuous artifacts and OOD masking. Naïve zero-masking in fidelity evaluation can in principle introduce out-of-distribution artifacts that confound the resulting confidence change. In this work, fidelity masking (Section E) replaces salient bins with a constant low-energy value (-80 dB), matching the padding value already used elsewhere in preprocessing (Section E) so that the masked regions remain within the amplitude range the models were trained on. We view richer in-distribution imputation strategies (e.g., diffusing salient bins to neighboring time–frequency content rather than a constant baseline) as a promising direction for further reducing residual OOD distortion in fidelity evaluation, and leave a systematic comparison to future work (Section 8).

C Example integrated gradients

Figure 3 shows representative Integrated Gradients (IG) explanations from the four consensus models, illustrating how different architectures emphasize distinct time–frequency regions for the same audio inputs.

Figure 4 shows examples of stationary artifacts for both Healthy and Unhealthy classes across datasets.

D Baseline metrics

We elaborate on the baseline metrics we used in our work below.



(a) Healthy (b) Unhealthy

Figure 4: **Stationary Artifact Across Classes:** Integrated gradients highlighting a persistent frequency band in both healthy and unhealthy samples from the controlled poultry dataset with an injected spurious feature.

(1) Faithfulness (Parekh et al., 2022). Measures whether explanation-selected features alone preserve the model's prediction given by:

$$\mathrm{F}_n = p_c(X_n^{\text{int}}). \tag{12}$$

Here, X_n^{int} retains only explanation-relevant features for sample n, and $p_c(\cdot)$ is the predicted probability for class c. Larger values indicate stronger preservation.

(2) Fidelity (Tomsett et al., 2020). Quantifies the confidence drop after removing the explanation region:

$$\mathrm{FF}_n = p_c(X_n) - p_c\left(X_n \odot (1 - M_n)\right). \tag{13}$$

M_n denotes the explanation mask; larger positive values indicate greater model reliance on the highlighted region.

(3) Average Increase (AI) (Chattopadhay et al., 2018). Percentage of samples where confidence increases when explanation region is retained:

$$\mathrm{AI} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left[p_c(X_n \odot M_n) > p_c(X_n)\right] \times 100. \tag{14}$$

(4) Average Drop (AD) (Chattopadhay et al., 2018). Relative confidence decrease under masking with smaller values being better.

$$\mathrm{AD} = \frac{1}{N}\sum_{n=1}^{N} \frac{\max(0,\, p_c(X_n) - p_c(X_n \odot M_n))}{p_c(X_n)} \times 100. \tag{15}$$

(5) Average Gain (AG) (Zhang et al., 2024). Relative confidence increase after masking with larger values indicating stronger confidence gain.

$$\mathrm{AG} = \frac{1}{N}\sum_{n=1}^{N} \frac{\max(0,\, p_c(X_n \odot M_n) - p_c(X_n))}{1 - p_c(X_n)} \times 100. \tag{16}$$

(6) Sparseness (Chalasani et al., 2020). Measures if predictions depend on a small set of salient features; larger values indicate better explanations.

(7) Complexity (Bhatt et al., 2020). Entropy of attribution scores; smaller values indicate tighter localization. Anonymized code available here: `https://anonymous.4open.science/r/AGRI-Fidelity-AFF3/README.md`.

E Implementation details for reproducibility

E.1 CoughLIME implementation on our datasets

We adapted CoughLIME into a fully spectrogram-native pipeline for a multi-dataset setup by eliminating dependence on raw `.wav` inputs, STFT generation, and sample-rate coupling. The workflow

no longer needed audio reconstruction or any sample-rate-specific preprocessing. Instead, each sample is loaded directly from precomputed spectrograms with sampling rate of 22050Hz. Explanations are generated using the spectral decomposition (8 components) in CoughLIME, which perturbs frequency-band components directly in spectrogram space.

For evaluation, metrics were extended and integrated to the main pipeline to compute fidelity, faithfulness, average increase, average drop, average decrease, average gain, complexity (COMP), and sparseness using model predictions (p_{original}, p_{masked}) and component (3 components) perturbations.

E.2 Time-Frequency representation and preprocessing

For all AGRI-Fidelity evaluations, we fix the audio sampling rate to $f_s = 22{,}050$ Hz across all recordings. Spectrogram construction parameters are selected based on the characteristic temporal scale of the acoustic events of interest.

Poultry vocalizations in our dataset consist mainly of syllable-like calls lasting tens of milliseconds, along with longer sustained calls. To balance temporal localization and frequency resolution, we use $n_{\text{fft}} = 1024$, corresponding to a window duration of approximately 46 ms at $22{,}050$ Hz. To enable smooth temporal tracking of spectral dynamics while preventing missed transitions, we use a 75% overlap between adjacent frames, implemented via a hop length of $\texttt{hop_length} = n_{\text{fft}}/4 = 256$. Unless otherwise specified, all experiments use the fixed configuration $(f_s, n_{\text{fft}}, \texttt{window length}, \texttt{hop length}) = (22{,}050, 1024, 1024, 256)$.

As preprocessing for multi-model training, spectrograms are normalized to a fixed duration of 15 seconds to ensure uniform input dimensions across models. For the Poultry dataset variants, we extract the central 15-second segment. Padding uses a constant low-energy value (-80 dB), representing the absence of an acoustic signal. This blank signal is also used in fidelity evaluations across datasets.

E.3 Controlled artifact construction

For the controlled spurious Poultry variant, we injected a stationary high-frequency sinusoidal artifact into the *Unhealthy* class. Specifically, for every denoised unhealthy audio file ending in `__hpss.wav`, we added a 10.2 kHz sine tone with amplitude 0.005:

$$y(t) = x(t) + 0.005 \sin(2\pi \cdot 10200 \cdot t).$$

The input files were required to have sampling rate 22.05 kHz, so the artifact was injected after resampling and denoising/HPSS preprocessing. Stereo files, if present, were converted to mono by channel averaging before injection. The perturbed signal was clipped to the range $[-1, 1]$. The artifact was added to all unhealthy samples and not to healthy samples. No explicit SNR calibration was used; the artifact strength was controlled by the fixed sinusoidal amplitude 0.005.

E.4 Consensus model training details

We trained four consensus models for each dataset: CNN, MLP, LSTM, and ResNet18. All models operate on precomputed spectrograms stored as `.npy` files, with Healthy labeled as 0 and Unhealthy labeled as 1. Inputs are loaded as single-channel spectrograms with shape $[1, F, T]$; multi-channel inputs, if present, are averaged to one channel. All models use binary cross-entropy with logits as the training objective and Adam optimization with weight decay 10^{-4}. Unless otherwise stated, inputs are normalized per sample. Each dataset is split into 80% training and 20% validation. Model selection is performed by saving the checkpoint with the highest validation accuracy. No early stopping is used; each model is trained for the full number of epochs specified in Table 5. The reported evaluation metrics are computed using the best-validation checkpoint.

Table 4 defines the four architectures used in the consensus committee: CNN, MLP, LSTM, and ResNet18. The models provide complementary inductive biases over spectrogram inputs, including convolutional feature extraction, fully connected frequency-pooled representations, temporal sequence modeling, and residual image-based representation learning.

Table 4: **Consensus model architectures.** All models output one binary logit and are trained with binary cross-entropy with logits.

Model	Architecture details
CNN	Three convolutional blocks: Conv2D$(1, 32)$ with kernel 5, stride $(1, 2)$, padding 2; Conv2D$(32, 64)$ with kernel 5, stride $(1, 2)$, padding 2; Conv2D$(64, 128)$ with kernel 3, stride $(1, 2)$, padding 1. Each block uses BatchNorm and ReLU, followed by adaptive average pooling, Linear$(128, 256)$, ReLU, Dropout(0.3), and Linear$(256, 1)$.
MLP	Optional adaptive frequency pooling to 64 frequency bins, flattening, and fully connected layers with hidden dimensions $(1024, 512, 128)$. Each hidden layer uses LayerNorm, ReLU, and Dropout(0.3), followed by a Linear$(128, 1)$ output head.
LSTM	Spectrogram is treated as a temporal sequence over time frames. The model uses a 2-layer bidirectional LSTM with hidden size 128 and Dropout(0.3). The output is mean-pooled over time and passed through LayerNorm, ReLU, Dropout(0.3), and Linear$(256, 1)$. Frequency pooling is disabled in the reported configuration.
ResNet18	Torchvision ResNet18 initialized without pretrained weights. Single-channel spectrograms are repeated to three channels before input to ResNet18, and the final fully connected layer is replaced with Linear$(512, 1)$.

Table 5: **Dataset-specific training hyperparameters for consensus models.** Each entry reports epochs / learning rate / batch size. All models use Adam with weight decay 10^{-4} and best-validation-accuracy checkpoint selection. **We followed a train-test split of 80-20 for all training regimes.**

Dataset	CNN	MLP	LSTM	ResNet18
Chicken Denoised	30 / 0.01 / 32	25 / 0.01 / 32	10 / 0.01 / 32	25 / 0.01 / 32
Chicken Spurious	25 / 0.1 / 32	5 / 0.01 / 32	10 / 0.1 / 32	20 / 0.1 / 32
SmartEars Raw	50 / 0.001 / 32	40 / 0.001 / 32	50 / 0.01 / 32	10 / 0.001 / 32
SmartEars Denoised	60 / 0.001 / 32	20 / 0.001 / 32	40 / 0.0009 / 128	60 / 0.001 / 32
SwineCough	10 / 0.01 / 128	15 / 0.01 / 128	20 / 0.01 / 128	35 / 0.0001 / 128

Table 5 reports the dataset-specific training settings for each consensus model, including epochs, learning rate, and batch size. All models are trained with Adam, weight decay 10^{-4}, and checkpoint selection based on best validation accuracy.

E.5 Explanation generation settings

We generate explanations using Integrated Gradients (IG) for each trained consensus model: CNN, MLP, LSTM, and ResNet18. Explanations are generated for all Healthy and Unhealthy spectrograms. Each input spectrogram is loaded as a single-channel tensor with shape $[1, F, T]$; multi-channel inputs, if present, are averaged into one channel. In the reported configuration, we use per-sample normalization unless the model checkpoint specifies otherwise. IG is computed with 64 integration steps and a zero baseline. For each sample and model, the attribution map is saved as both a NumPy array and a visualization. The output directory is organized by dataset, class label, and model name. All details of IG generation are provided in Table 6.

E.6 Reliability generation parameters

Reliability is computed from Integrated Gradients maps produced by the four consensus models: CNN, LSTM, MLP, and ResNet18. For each dataset and class, we load the corresponding IG maps and convert each attribution map into a binary salient-region mask. Saliency masks are constructed using the top positive attribution values, with top-$k = 0.05$ of time-frequency bins retained by default. Consensus is computed by averaging the four binary masks at each time-frequency bin, producing an observed consensus map $S_{\text{obs}} \in [0, 1]$.

Table 6: Design parameters for explanation generation.

Parameter	Setting
Explainer	Integrated Gradients using Captum
Models explained	CNN, MLP, LSTM, ResNet18 consensus models
Input samples	All spectrograms from Healthy and Unhealthy classes
Input shape	Single-channel spectrogram tensor $[1, F, T]$
Multi-channel handling	Channels are averaged to obtain one channel
Normalization	Checkpoint-specific normalization; default is per-sample normalization
Baseline	Zero baseline
IG steps	64 integration steps
Target output	Binary model logit after squeezing the model output

Table 7: Design parameters for reliability generation.

Parameter	Setting
Input explanations	Integrated Gradients maps from CNN, LSTM, MLP, and ResNet18
Input classes	Healthy and Unhealthy
Mask construction	Binary masks from top positive attribution values
Top-k fraction	0.05 of time-frequency bins by default
Consensus committee size	Four models
Observed consensus	Mean of the four binary masks at each time-frequency bin
Consensus thresholds	Unique nonzero values of observed consensus S_{obs}
Null model	Cyclic temporal shifts of saliency masks along the time axis
Number of null samples	$B = 50$ by default
Default shift range	Random shifts from 10 to $T - 100$ frames
SwineCough shift range	Random shifts from 1 to $\max(1, T/2)$ frames
FDR threshold	$\alpha = 0.1$
FDR estimate	Mean discoveries divided by observed discoveries at each λ, clipped at 1
Reliability score	$1 - \text{FDR}$ at the highest observed consensus level

To estimate chance-level agreement, we generate a null distribution using cyclic temporal shifts of each model's saliency mask. By default, we use $B = 50$ null samples and draw random shifts between `min_shift`$= 10$ and `max_shift`$= T - 100$, where T is the number of time frames. For SwineCough, we use `min_shift`$= 1$ and `max_shift`=$\max(1, T/2)$ because the clips contain short transient cough events and number of time frames was much less than our other datasets. For each consensus threshold λ appearing in S_{obs}, we compute empirical FDR as the ratio between the average number of null bins satisfying $S_{\text{null}} \geq \lambda$ and the observed number of bins satisfying $S_{\text{obs}} \geq \lambda$, clipped at 1. We use $\alpha = 0.1$ to identify statistically valid consensus thresholds.

Reliability is computed at the highest observed consensus level as $1 - \text{FDR}$ at $\lambda_{\max}$. Fidelity is computed for each model by masking that model's selected salient bins with the normalized zero baseline and measuring the signed change in predicted probability before and after masking, following Eq. 13 (larger positive values indicate greater model reliance on the masked region; we do not take an absolute value, since doing so would reward explanations that suppress the correct class). The final reported `AGRI-Fidelity` score is the product of mean fidelity across the four models and the reliability score. The output file stores, for each sample, the file name, class label, mean fidelity, mean reliability, and final `AGRI-Fidelity` score. All design choices for metric generation are provided in Table 7.

E.7 Computational resources

We used an internal university HPC cluster with GPU compute nodes. Each GPU node provides 4 NVIDIA GPUs with 24 GB GPU memory per GPU, 64 CPU cores, 512 GB system RAM, and 1.7 TB local NVMe storage. Shared high-performance storage was used for project files and scratch I/O.

F EFFECT OF COMMITTEE SIZE (K) ON RELIABILITY

1. The exponentially higher confidence. According to Theorem 2 (Sensitivity to Sparse Signals), the False Discovery Rate (FDR) at full consensus ($\lambda = 1.0$) for a time-localized signal is bounded by the sparsity ratio ρ raised to the power of $K - 1$:

$$FDR(1.0) \approx \rho^{K-1} \tag{17}$$

Because a genuine bioacoustic signal (such as a cough) is sparse, the sparsity ratio ρ is strictly less than 1 ($\rho \ll 1$). When the committee size K increases (e.g., $K = 8$ or $K = 10$), this small fraction is raised to a much higher power. As $K \to \infty$, the term ρ^{K-1} collapses toward zero at an exponential rate. Consequently, the Reliability Score ($R \approx 1 - \rho^{K-1}$) asymptotically approaches 1.0.

Under the cyclic-shift null model, FDR measures the probability that models align perfectly by random chance. The mathematical odds of 4 independent random shifts aligning perfectly on a brief 46 ms audio blip are low; the odds of 10 independent models randomly aligning on that exact same blip are virtually impossible. Therefore, if a larger committee *does* achieve full consensus, `AGRI-Fidelity` is mathematically much more confident that the feature is a true biological marker, yielding a strictly higher reliability score.

2. The Impact on stationary artifacts. For continuous, stationary background noise (e.g., ventilation hum), increasing K has *no effect* on the FDR. According to Theorem 1, because the continuous artifact spans the entire temporal axis, cyclic permutations do not disrupt the inter-model overlap. The expected null count always matches the observed count:

$$\mathcal{E}[N_{null}(\lambda)] = N_{obs}(\lambda) \tag{18}$$

Therefore, $FDR = 1.0$ and the Reliability score remains strictly 0.0, regardless of whether the evaluation is performed with 4 models or 40 models.

3. The practical trade-off. While a higher K increases the *theoretical* reliability, it makes *empirical* evaluation much stricter. Due to model multiplicity and differing inductive biases across architectures (e.g., CNNs vs. LSTMs vs. Transformers), requiring a massive committee to perfectly overlap on the exact same pixels ($\lambda = 1.0$) introduces a severe bottleneck.

- **With $K = 4$:** We can easily evaluate at maximum strictness ($\lambda = 1.0$, where all 4 models agree) or high strictness ($\lambda = 0.75$, where 3 out of 4 agree).
- **With $K = 10$:** The observed consensus S_{obs} at the highest tier ($\lambda = 1.0$) will frequently drop to zero, as at least one model is highly likely to disagree on the precise time-frequency boundary of the feature. To find overlapping features, stakeholders would be forced to evaluate at lower fractional consensus tiers (e.g., $\lambda = 0.8$ or $\lambda = 0.6$).

G EFFECTS OF REVERBERATION ON `AGRI-FIDELITY` RELIABILITY

Under reverberant conditions, the impact on `AGRI-Fidelity` reliability depends on whether the resulting Room Impulse Response (RIR) tail is perceived as a stationary background effect or a localized signal extension. If reverberation manifests as a continuous, time-invariant acoustic artifact across the recording, Theorem 1 guarantees that it will asymptotically yield a False Discovery Rate (FDR) ≈ 1.0, effectively assigning it a Reliability Score ≈ 0. For sparse bioacoustic events such as poultry coughs, a short room impulse response (RIR) tail that merely extends the event duration τ is handled by Theorem 2; as long as the signal remains relatively sparse ($\tau \ll T$), the expected FDR remains negligible and the feature is retained with high confidence. However, if severe reverberation smears the signal across a large portion of the time axis, the increasing sparsity ratio ρ will mathematically reduce the reliability score ($R \approx 1 - \rho^{K-1}$), reflecting the loss of precise temporal localization required for expert validation. Furthermore, because reverberation can cause different model architectures to disagree on feature boundaries, any resulting drop in cross-model consensus will naturally lead to a higher empirical FDR and a lower reliability assessment.

Table 8: **Performance of consensus architectures across datasets.** For each dataset, we report the performance of CNN, MLP, LSTM, and ResNet consensus models. Columns report Accuracy (Acc↑), Precision (Prec↑), Recall (Rec↑), F1↑, Specificity (Spec↑), and False Positive Rate (FPR↓).

Dataset	Model	Acc↑	Prec↑	Rec↑	F1↑	Spec↑	FPR↓
Chicken Denoised	CNN	0.8654	0.9306	0.7667	0.8231	0.9360	0.0640
	MLP	0.9096	0.9189	0.8781	0.8954	0.9322	0.0678
	LSTM	**0.9615**	**0.9987**	0.8889	**0.9412**	**0.9998**†	**0.0002**
	ResNet	0.9385	0.9321	**0.9436**	0.9366	0.9313	0.0687
Chicken Spurious	CNN	**0.9865**	0.9805	**0.9913**	**0.9855**	0.9823	0.0177
	MLP	0.9365	0.9526	0.9036	0.9229	0.9618	0.0382
	LSTM	0.9423	0.9412	0.8889	0.9143	0.9706	0.0294
	ResNet	0.9500	**0.9960**	0.8962	0.9338	**0.9964**	**0.0036**
SmartEars Clean	CNN	0.7399	0.7170	0.7886	0.7498	0.6901	0.3099
	MLP	0.6710	0.6741	0.6661	0.6651	0.6756	0.3244
	LSTM	**0.8413**	**0.8797**	0.7575	**0.8141**	**0.9122**	**0.0878**
	ResNet	0.8065	0.7940	**0.8299**	0.8098	0.7827	0.2173
SmartEars Noisy	CNN	0.8311	0.8506	0.9236	0.8853	0.6068	0.3932
	MLP	**0.8967**	0.8973	0.9788	**0.9363**	0.6135	0.3865
	LSTM	0.8235	0.8250	**0.9848**	0.8976	0.2391	0.7609
	ResNet	0.8816	**0.9072**	0.9460	0.9256	**0.6532**	**0.3468**
SwineCough	CNN	0.9613	0.9553	0.9575	0.9558	0.9637	0.0363
	MLP	0.9694	0.9648	0.9654	0.9650	0.9724	0.0276
	LSTM	0.9593	0.9516	0.9555	0.9535	0.9621	0.0379
	ResNet	**0.9817**	**0.9819**	**0.9761**	**0.9790**	**0.9860**	**0.0140**

H Additional results: comparative analysis of explanation metrics

H.1 Consensus model performance across datasets.

Table 8 verifies that the consensus committee provides sufficiently accurate but architecturally diverse predictors across datasets. Under noisy SmartEars conditions, recall remains high while specificity declines, indicating that background noise can bias predictions and motivating reliability-aware explanation evaluation.

H.2 Statistical Analyses of Fidelity-Reliability Results Across Datasets

We perform a detailed statistical analysis of the Chicken Denoised, SmartEars, SmartEars Denoised, and Chicken Spurious datasets with 34 consensus models used in the main paper. For all of our evaluated datasets, the statistical analysis compared explanation methods separately within each evaluation metric, and the metric values themselves were not compared directly across metrics. For each metric, the explainers LMAC-TD, L2I, CoughLIME, and AudioLIME were evaluated on the same test samples, yielding paired observations at the sample level.

Pairwise differences in the non-binary sample-level metrics, AD, AG, faithfulness, sparseness, complexity, fidelity, and reliability, were assessed using two-sided Wilcoxon signed-rank tests. This nonparametric paired test was selected because the same recordings were evaluated by every explanation method and the resulting metric distributions were not assumed to be normally distributed.

AI was represented as a binary outcome for each sample, indicating whether the explanation increased the target-class confidence. Consequently, pairwise differences in AI were assessed using the exact McNemar test for matched binary observations.

For all four explanation methods, each metric involved six pairwise method comparisons. Holm's procedure was applied separately within each metric to control the family-wise error rate across these six comparisons. Statistical significance was defined as $p_{\mathrm{Holm}} < 0.05$. For Wilcoxon comparisons, rank-biserial correlation r_{rb} was reported to quantify the magnitude and direction of the

Table 9: Quantitative comparison on the Chicken Denoised dataset. Values are mean ± standard deviation across 52 test samples. Within each metric, methods that share a superscript letter were not significantly different after Holm correction. Methods with no letter in common differed at $p_{\text{Holm}} < 0.05$. AI was compared using the exact McNemar test; all other metrics were compared using two-sided Wilcoxon signed-rank tests. Superscript letters are metric-specific and should not be compared across columns.

Baseline	AI↑	AD↓	AG↑	Faithfulness↑
LMAC–TD	**73.0769 ± 44.7888**[a]	5.0801 ± 18.1076[a]	**49.7342 ± 38.1372**[a]	0.0942 ± 0.2440[b]
L2I	23.0769 ± 42.5436[c]	8.7578 ± 13.3723[b]	8.4091 ± 18.3646[c]	0.4086 ± 0.3419[a]
CoughLIME	48.0769 ± 50.4505[ab]	**1.6080 ± 2.8251**[a]	27.4042 ± 37.4515[b]	**0.5228 ± 0.4124**[a]
AudioLIME	46.1538 ± 50.3382[bc]	7.0320 ± 13.1816[b]	25.9908 ± 35.4451[b]	0.5023 ± 0.3182[a]

Baseline	Sparse↑	Comp.↓	AGRI-Fidelity	
			Fidelity↑	Reliability↑
LMAC–TD	**0.8055 ± 0.1032**[a]	8.8475 ± 1.2242[b]	0.0983 ± 0.1292[ab]	0.6038 ± 0.3135[a]
L2I	0.0401 ± 0.0326[c]	12.2529 ± 0.0518[c]	0.0963 ± 0.1064[ab]	**0.6362 ± 0.2862**[a]
CoughLIME	0.3174 ± 0.1163[b]	12.8489 ± 0.2681[d]	0.0578 ± 0.0847[b]	0.5789 ± 0.3138[a]
AudioLIME	0.3288 ± 0.1524[b]	**2.1374 ± 0.2014**[a]	**0.1253 ± 0.1681**[a]	0.6283 ± 0.3194[a]

paired difference. For McNemar comparisons, the numbers of discordant sample pairs favoring each method were reported. Sections H.2.1,H.2.2, H.2.3 and H.2.4 describe these analyses in detail.

H.2.1 Statistical comparison on the Chicken Denoised dataset

Table 9 presents the statistical comparison of the four explanation methods on the Chicken Denoised dataset using the same 52 test samples. Below, we summarize the pairwise statistical results separately for each evaluation metric.

Average Increase (AI). LMAC-TD achieved the highest mean AI. Exact McNemar tests showed that LMAC-TD outperformed L2I ($p_{\text{Holm}} = 3.70 \times 10^{-5}$; 30 discordant samples favored LMAC-TD and 4 favored L2I) and AudioLIME ($p_{\text{Holm}} = 0.0374$; 20 versus 6 discordant successes). CoughLIME also outperformed L2I ($p_{\text{Holm}} = 0.0360$; 17 versus 4 discordant successes). The difference between LMAC–TD and CoughLIME was not significant after correction ($p_{\text{Holm}} = 0.0575$). CoughLIME and AudioLIME were also statistically indistinguishable.

Average Decrease (AD). Lower AD indicates better performance. LMAC-TD had significantly lower AD than L2I ($p_{\text{Holm}} = 0.00288$, $r_{\text{rb}} = -0.596$) and AudioLIME ($p_{\text{Holm}} = 0.0302$, $r_{\text{rb}} = -0.506$). CoughLIME had significantly lower AD than L2I ($p_{\text{Holm}} = 0.00124$, $|r_{\text{rb}}| = 0.642$) and AudioLIME ($p_{\text{Holm}} = 0.0263$, $|r_{\text{rb}}| = 0.520$). LMAC-TD and CoughLIME did not differ significantly. Thus, the statistical results support LMAC-TD and CoughLIME as the strongest AD group, rather than establishing CoughLIME as uniquely best.

Average Gain (AG). LMAC–TD significantly outperformed every alternative in AG. Its advantage was significant against L2I ($p_{\text{Holm}} = 8.61 \times 10^{-6}$, $r_{\text{rb}} = 0.854$), CoughLIME ($p_{\text{Holm}} = 0.0197$, $r_{\text{rb}} = 0.465$), and AudioLIME ($p_{\text{Holm}} = 0.0197$, $r_{\text{rb}} = 0.499$). CoughLIME and AudioLIME both significantly outperformed L2I ($p_{\text{Holm}} = 0.0197$), but did not differ from each other. Consequently, the supported ordering was LMAC–TD first, CoughLIME and AudioLIME jointly second, and L2I last.

Faithfulness. Although CoughLIME had the largest mean faithfulness, it did not significantly outperform AudioLIME or L2I. These three methods formed one statistically indistinguishable group. LMAC–TD was significantly less faithful than L2I, CoughLIME, and AudioLIME ($p_{\text{Holm}} \leq 2.20 \times 10^{-5}$), with large absolute rank-biserial effects ($|r_{\text{rb}}| = 0.724$–0.807). The evidence therefore supports a three-method leading group, not a unique CoughLIME advantage.

Table 10: Quantitative comparison on the SmartEars dataset. Values are mean ± standard deviation across 271 paired test samples. Within each metric, methods sharing a superscript letter were not significantly different after Holm correction. Methods with no letter in common differed at $p_{\mathrm{Holm}} < 0.05$. AI was assessed using the exact McNemar test; all other metrics were assessed using two-sided Wilcoxon signed-rank tests. Superscript letters are metric-specific.

Baseline	AI↑	AD↓	AG↑	Faithfulness↑
LMAC–TD	51.2915 ± 50.0758^{c}	3.5894 ± 12.1133^{c}	37.6760 ± 41.9414^{c}	$\mathbf{0.8438 \pm 0.3143}^{a}$
L2I	36.9004 ± 48.3428^{d}	20.0765 ± 28.6492^{d}	26.4102 ± 38.9857^{d}	0.2722 ± 0.2958^{c}
CoughLIME	$\mathbf{100.0000 \pm 0.0000}^{a}$	$\mathbf{0.0000 \pm 0.0000}^{a}$	$\mathbf{97.5147 \pm 2.8108}^{a}$	0.2173 ± 0.1655^{d}
AudioLIME	73.4317 ± 44.2513^{b}	0.7455 ± 3.2800^{b}	46.9807 ± 35.0174^{b}	0.3975 ± 0.3569^{b}

Baseline	Sparse↑	Comp.↓	`AGRI-Fidelity`	
			Fidelity↑	Reliability↑
LMAC–TD	$\mathbf{0.6231 \pm 0.1310}^{a}$	10.3108 ± 0.6969^{b}	0.0681 ± 0.0975^{d}	0.4026 ± 0.1582^{b}
L2I	0.0224 ± 0.0187^{d}	12.2805 ± 0.0242^{c}	0.1957 ± 0.2396^{c}	0.3144 ± 0.2753^{c}
CoughLIME	0.1675 ± 0.0440^{c}	13.1569 ± 0.0739^{d}	$\mathbf{0.3476 \pm 0.1996}^{a}$	0.3738 ± 0.2567^{b}
AudioLIME	0.4075 ± 0.1141^{b}	$\mathbf{2.0002 \pm 0.1809}^{a}$	0.2369 ± 0.2275^{b}	$\mathbf{0.5116 \pm 0.2996}^{a}$

Sparseness. LMAC–TD was significantly sparser than every other method ($p_{\mathrm{Holm}} = 2.10 \times 10^{-9}$ for each comparison), with near-maximal rank-biserial effects ($r_{\mathrm{rb}} = 0.999$–1.000). CoughLIME and AudioLIME were both significantly sparser than L2I at the same adjusted significance level. CoughLIME and AudioLIME did not significantly differ from each other. The supported ordering was therefore LMAC–TD first, CoughLIME and AudioLIME jointly second, and L2I last.

Complexity. Lower complexity indicates better performance. Every pairwise complexity comparison remained significant after Holm correction ($p_{\mathrm{Holm}} = 2.10 \times 10^{-9}$), with absolute rank-biserial effects between 0.988 and 1.000. The complete statistically supported ordering was

$$\text{AudioLIME} < \text{LMAC–TD} < \text{L2I} < \text{CoughLIME}.$$

AudioLIME therefore achieved significantly lower complexity than every other method.

`AGRI-Fidelity` fidelity. AudioLIME had the largest mean fidelity, but its advantage was significant only relative to CoughLIME ($p_{\mathrm{Holm}} = 0.00135$, $|r_{\mathrm{rb}}| = 0.588$). AudioLIME did not significantly differ from LMAC–TD or L2I. Similarly, the numerical differences among LMAC–TD, L2I, and CoughLIME did not survive Holm correction. AudioLIME should therefore be described as numerically best, with a statistically supported advantage over CoughLIME only.

`AGRI-Fidelity` reliability. No pairwise reliability comparison was statistically significant after Holm correction. The smallest adjusted reliability p-value was 0.689. Although L2I had the largest mean reliability, the data do not support a claim that any method was more reliable than the others on the Chicken Denoised test set.

Overall, the statistical results establish metric-specific strengths rather than a single dominant explanation method. LMAC–TD was strongest for AG and sparseness and belonged to the leading AI and AD groups. AudioLIME achieved the lowest complexity and had higher fidelity than CoughLIME, but did not significantly exceed LMAC–TD or L2I in fidelity. CoughLIME belonged to the leading AD and faithfulness groups. Most importantly, the four methods did not significantly differ in consensus reliability despite their numerical differences.

H.2.2 Statistical comparison on the SmartEars dataset

Table 10 presents the statistical comparison of the four explanation methods on the SmartEars dataset using the same 271 test samples. Below we summarize the pairwise statistical results separately for each evaluation metric.

Average Increase (AI). Every pairwise AI comparison was significant after Holm correction ($p_{\mathrm{Holm}} \leq 0.00316$). CoughLIME achieved AI success on all 271 samples and significantly out-

performed AudioLIME, LMAC-TD, and L2I. AudioLIME significantly outperformed LMAC-TD and L2I, while LMAC-TD significantly outperformed L2I. The statistically supported ordering was

$$\text{CoughLIME} > \text{AudioLIME} > \text{LMAC–TD} > \text{L2I}.$$

The zero standard deviation for CoughLIME reflects a ceiling effect rather than uncertainty in the estimate: every per-sample AI value equaled 100.

Average Decrease (AD). Every pairwise AD comparison was significant ($p_{\text{Holm}} \leq 1.56 \times 10^{-8}$). Because lower AD is better, the supported ordering was

$$\text{CoughLIME} < \text{AudioLIME} < \text{LMAC–TD} < \text{L2I}.$$

CoughLIME achieved zero AD for every test sample. Its zero mean and standard deviation therefore represent a floor effect. AudioLIME also had significantly lower AD than both LMAC–TD and L2I.

Average Gain (AG). All six AG comparisons remained significant after correction ($p_{\text{Holm}} \leq 0.0313$). CoughLIME achieved substantially greater AG than every other method, including AudioLIME ($p_{\text{Holm}} = 2.13 \times 10^{-44}$, $r_{\text{rb}} = 0.988$). AudioLIME ranked second, followed by LMAC–TD and L2I:

$$\text{CoughLIME} > \text{AudioLIME} > \text{LMAC–TD} > \text{L2I}.$$

Faithfulness. All faithfulness comparisons were significant ($p_{\text{Holm}} \leq 0.0274$). LMAC–TD was significantly more faithful than every alternative, with large rank-biserial effects against L2I, CoughLIME, and AudioLIME ($r_{\text{rb}} = 0.866$–0.947). AudioLIME ranked second, followed by L2I and CoughLIME:

$$\text{LMAC–TD} > \text{AudioLIME} > \text{L2I} > \text{CoughLIME}.$$

Thus, CoughLIME's strong AI, AD, and AG performance did not translate into high faithfulness.

Sparseness. Every sparseness comparison was significant ($p_{\text{Holm}} \leq 4.36 \times 10^{-44}$). LMAC–TD produced the sparsest explanations, followed by AudioLIME, CoughLIME, and L2I:

$$\text{LMAC–TD} > \text{AudioLIME} > \text{CoughLIME} > \text{L2I}.$$

The comparisons involving LMAC–TD had rank-biserial effects between 0.976 and 1.000, indicating highly consistent sample-level differences.

Complexity. All pairwise complexity comparisons were significant ($p_{\text{Holm}} = 2.03 \times 10^{-45}$). Since lower complexity is preferred, the complete ordering was

$$\text{AudioLIME} < \text{LMAC–TD} < \text{L2I} < \text{CoughLIME}.$$

AudioLIME therefore produced significantly less complex explanations than every alternative. CoughLIME had the greatest complexity despite its strong AI, AD, AG, and fidelity performance.

`AGRI-Fidelity` fidelity. Every pairwise fidelity comparison was significant ($p_{\text{Holm}} \leq 9.42 \times 10^{-4}$). CoughLIME achieved the greatest fidelity, followed by AudioLIME, L2I, and LMAC–TD:

$$\text{CoughLIME} > \text{AudioLIME} > \text{L2I} > \text{LMAC–TD}.$$

CoughLIME significantly exceeded AudioLIME ($p_{\text{Holm}} = 4.48 \times 10^{-9}$, $r_{\text{rb}} = 0.419$), while AudioLIME significantly exceeded L2I and LMAC–TD.

`AGRI-Fidelity` reliability. AudioLIME achieved significantly greater reliability than LMAC–TD ($p_{\text{Holm}} = 1.65 \times 10^{-8}$), L2I ($p_{\text{Holm}} = 1.37 \times 10^{-12}$), and CoughLIME ($p_{\text{Holm}} = 1.76 \times 10^{-17}$). LMAC–TD and CoughLIME were both significantly more reliable than L2I. However, LMAC–TD and CoughLIME did not significantly differ ($p_{\text{Holm}} = 0.2435$). The supported reliability grouping was therefore

$$\text{AudioLIME} > \{\text{LMAC–TD}, \text{CoughLIME}\} > \text{L2I}.$$

Overall, the SmartEars results showed strong separation among the four explanation methods: $47/48$ pairwise comparisons remained significant after Holm correction. No method dominated every evaluation dimension. CoughLIME led AI, AD, AG, and fidelity, but had the lowest faithfulness and the greatest complexity. LMAC–TD led faithfulness and sparseness. AudioLIME achieved the lowest complexity and the greatest consensus reliability. L2I did not lead any metric on this dataset. These results demonstrate why explanation methods should be evaluated across complementary criteria rather than ranked using a single metric.

Table 11: Quantitative comparison on the SmartEars Denoised dataset. Values are mean ± standard deviation across 271 paired test samples. Within each metric, methods sharing a superscript letter were not significantly different after Holm correction. Methods with no letter in common differed at $p_{\text{Holm}} < 0.05$. AI was assessed using the exact McNemar test; all other metrics were assessed using two-sided Wilcoxon signed-rank tests. Superscript letters are metric-specific.

Baseline	AI↑	AD↓	AG↑	Faithfulness↑
LMAC–TD	47.9705 ± 50.0512^{c}	10.1437 ± 24.2249^{c}	40.3661 ± 45.3392^{c}	$\mathbf{0.8793 \pm 0.2997}^{a}$
L2I	35.0554 ± 47.8026^{d}	18.8788 ± 33.3563^{d}	26.9938 ± 39.5217^{d}	0.1164 ± 0.2739^{d}
CoughLIME	$\mathbf{99.6310 \pm 6.0746}^{a}$	$\mathbf{0.0014 \pm 0.0226}^{a}$	$\mathbf{92.9236 \pm 11.1646}^{a}$	0.2015 ± 0.2210^{c}
AudioLIME	83.0258 ± 37.6100^{b}	1.8072 ± 8.3919^{b}	63.1785 ± 35.7665^{b}	0.3087 ± 0.3475^{b}

Baseline	Sparse↑	Comp.↓	AGRI-Fidelity	
			Fidelity↑	Reliability↑
LMAC–TD	$\mathbf{0.6026 \pm 0.0533}^{a}$	10.4535 ± 0.2859^{b}	0.1095 ± 0.2118^{d}	0.9472 ± 0.0843^{a}
L2I	0.0999 ± 0.0631^{d}	12.1748 ± 0.1093^{c}	0.1877 ± 0.2912^{c}	0.9071 ± 0.1553^{b}
CoughLIME	0.2329 ± 0.0767^{c}	13.0715 ± 0.1307^{d}	$\mathbf{0.2673 \pm 0.2234}^{a}$	$\mathbf{0.9657 \pm 0.0176}^{a}$
AudioLIME	0.3600 ± 0.0985^{b}	$\mathbf{2.1031 \pm 0.1267}^{a}$	0.2311 ± 0.2340^{b}	0.9277 ± 0.1613^{a}

H.2.3 Statistical comparison on the SmartEars Denoised dataset

Table 11 presents the statistical comparison of the four explanation methods on the matched SmartEars Denoised test subset comprising 271 samples. Below we summarize the pairwise statistical results separately for each evaluation metric.

AI. Every pairwise AI comparison was significant after Holm correction ($p_{\text{Holm}} \leq 0.00262$). CoughLIME achieved an AI increase on 270 of the 271 test samples and significantly outperformed AudioLIME, LMAC–TD, and L2I. AudioLIME significantly outperformed LMAC–TD and L2I, while LMAC–TD significantly outperformed L2I. The statistically supported ordering was

$$\text{CoughLIME} > \text{AudioLIME} > \text{LMAC–TD} > \text{L2I}.$$

CoughLIME's mean of 99.6310 and relatively small standard deviation reflect a near-ceiling result, with only one sample failing to exhibit an increase.

AD. Every pairwise AD comparison was significant ($p_{\text{Holm}} \leq 2.01 \times 10^{-4}$). Because lower AD is better, the statistically supported ordering was

$$\text{CoughLIME} < \text{AudioLIME} < \text{LMAC–TD} < \text{L2I}.$$

CoughLIME achieved an AD close to zero (0.0014 ± 0.0226), significantly outperforming all other methods. AudioLIME also achieved significantly lower AD than both LMAC–TD and L2I. The near-zero CoughLIME result indicates a strong floor effect, although it was not exactly zero for every test sample.

AG. All six AG comparisons remained significant after Holm correction ($p_{\text{Holm}} \leq 8.56 \times 10^{-5}$). CoughLIME achieved significantly greater AG than every alternative, including AudioLIME ($p_{\text{Holm}} = 2.52 \times 10^{-34}$, $r_{\text{rb}} = 0.867$). AudioLIME ranked second, followed by LMAC–TD and L2I:

$$\text{CoughLIME} > \text{AudioLIME} > \text{LMAC–TD} > \text{L2I}.$$

Faithfulness. All faithfulness comparisons were significant ($p_{\text{Holm}} \leq 2.83 \times 10^{-7}$). LMAC–TD was significantly more faithful than every alternative, with large rank-biserial effects against L2I, CoughLIME, and AudioLIME ($r_{\text{rb}} = 0.878$–0.959). AudioLIME ranked second, followed by CoughLIME and L2I:

$$\text{LMAC–TD} > \text{AudioLIME} > \text{CoughLIME} > \text{L2I}.$$

Thus, CoughLIME's strong AI, AD, and AG performance did not translate into the greatest faithfulness.

Sparseness. Every sparseness comparison was significant ($p_{\mathrm{Holm}} \leq 1.47 \times 10^{-37}$). LMAC–TD produced the sparsest explanations, followed by AudioLIME, CoughLIME, and L2I:

$$\text{LMAC–TD} > \text{AudioLIME} > \text{CoughLIME} > \text{L2I}.$$

The comparisons between LMAC–TD and the three alternatives had rank-biserial effects between 0.999 and 1.000, indicating extremely consistent sample-level differences.

Complexity. All pairwise complexity comparisons were significant ($p_{\mathrm{Holm}} = 2.03 \times 10^{-45}$). Since lower complexity is preferred, the complete ordering was

$$\text{AudioLIME} < \text{LMAC–TD} < \text{L2I} < \text{CoughLIME}.$$

AudioLIME therefore produced significantly less complex explanations than every alternative. CoughLIME had the greatest complexity despite its strong AI, AD, AG, and fidelity performance. All complexity comparisons had an absolute rank-biserial correlation of 1.000, indicating that the direction of the paired difference was completely consistent across samples.

`AGRI-Fidelity` fidelity. Every pairwise fidelity comparison was significant ($p_{\mathrm{Holm}} \leq 0.00202$). CoughLIME achieved the greatest fidelity, followed by AudioLIME, L2I, and LMAC–TD:

$$\text{CoughLIME} > \text{AudioLIME} > \text{L2I} > \text{LMAC–TD}.$$

CoughLIME significantly exceeded AudioLIME ($p_{\mathrm{Holm}} = 8.88 \times 10^{-4}$, $r_{\mathrm{rb}} = 0.246$), L2I ($p_{\mathrm{Holm}} = 4.88 \times 10^{-8}$, $r_{\mathrm{rb}} = 0.399$), and LMAC–TD ($p_{\mathrm{Holm}} = 1.93 \times 10^{-20}$, $r_{\mathrm{rb}} = 0.663$). Although CoughLIME had the greatest mean fidelity, its advantage over AudioLIME was smaller than its advantages over L2I and LMAC–TD.

`AGRI-Fidelity` reliability. CoughLIME achieved the greatest mean reliability (0.9657 ± 0.0176), but it did not differ significantly from LMAC–TD ($p_{\mathrm{Holm}} = 0.0903$) or AudioLIME ($p_{\mathrm{Holm}} = 0.5756$). LMAC–TD and AudioLIME also did not differ significantly ($p_{\mathrm{Holm}} = 0.2685$). These three methods therefore formed a common statistical group.

In contrast, L2I was significantly less reliable than LMAC–TD ($p_{\mathrm{Holm}} = 1.89 \times 10^{-7}$, $r_{\mathrm{rb}} = 0.386$), CoughLIME ($p_{\mathrm{Holm}} = 6.03 \times 10^{-17}$, $r_{\mathrm{rb}} = 0.602$), and AudioLIME ($p_{\mathrm{Holm}} = 1.67 \times 10^{-6}$, $r_{\mathrm{rb}} = 0.355$). The statistically supported reliability grouping was therefore

$$\{\text{CoughLIME}, \text{LMAC–TD}, \text{AudioLIME}\} > \text{L2I}.$$

Although CoughLIME had the largest numerical mean, the statistical results do not support describing it as uniquely more reliable than LMAC–TD or AudioLIME.

Overall, the SmartEars Denoised results showed strong separation among the four explanation methods: $45/48$ pairwise comparisons remained significant after Holm correction. The only nonsignificant comparisons were the three reliability comparisons among CoughLIME, LMAC–TD, and AudioLIME. No method dominated every evaluation dimension. CoughLIME led AI, AD, AG, and fidelity; LMAC-TD led faithfulness and sparseness; and AudioLIME achieved the lowest complexity. L2I did not lead any metric and had significantly lower reliability than the other three methods. These findings demonstrate that strong perturbation-based performance does not necessarily imply greater faithfulness, lower complexity, or uniquely greater consensus reliability. Explanation methods should therefore be evaluated across complementary criteria rather than ranked using a single metric.

H.2.4 Statistical comparison on the Chicken Spurious dataset

AI. Four of the six AI comparisons remained significant after Holm correction. CoughLIME and LMAC-TD did not differ significantly ($p_{\mathrm{Holm}} = 0.8318$), nor did L2I and AudioLIME ($p_{\mathrm{Holm}} = 0.7666$). However, both CoughLIME and LMAC–TD significantly outperformed both L2I and AudioLIME. For example, the CoughLIME–AudioLIME comparison contained 30 discordant samples favouring CoughLIME and none favouring AudioLIME ($p_{\mathrm{Holm}} = 1.12 \times 10^{-8}$). The statistically supported grouping was therefore

$$\{\text{CoughLIME}, \text{LMAC–TD}\} > \{\text{L2I}, \text{AudioLIME}\}.$$

The large AI standard deviations follow from its binary sample-level definition and do not indicate computational instability.

AD. Four of the six AD comparisons were significant ($p_{\mathrm{Holm}} \leq 0.00971$). CoughLIME and LMAC–TD did not significantly differ ($p_{\mathrm{Holm}} = 1.000$), despite CoughLIME having the smaller numerical mean. L2I and AudioLIME also did not differ significantly ($p_{\mathrm{Holm}} = 1.000$). Since lower AD is preferred, the supported grouping was

$$\{\text{CoughLIME}, \text{LMAC–TD}\} < \{\text{AudioLIME}, \text{L2I}\}.$$

Thus, CoughLIME's numerically lowest AD should not be interpreted as a statistically significant advantage over LMAC–TD.

AG. Four of the six AG comparisons were significant ($p_{\mathrm{Holm}} \leq 1.29 \times 10^{-5}$). LMAC–TD and CoughLIME did not significantly differ ($p_{\mathrm{Holm}} = 0.7119$), while L2I and AudioLIME also did not differ ($p_{\mathrm{Holm}} = 0.8484$). Both LMAC–TD and CoughLIME significantly exceeded L2I and AudioLIME. The supported grouping was therefore

$$\{\text{LMAC–TD}, \text{CoughLIME}\} > \{\text{AudioLIME}, \text{L2I}\}.$$

Although LMAC–TD had the largest mean AG, its advantage over CoughLIME was not statistically significant.

Faithfulness. Five of the six faithfulness comparisons were significant ($p_{\mathrm{Holm}} \leq 0.00404$). AudioLIME achieved significantly greater faithfulness than LMAC–TD, L2I, and CoughLIME. L2I and CoughLIME did not significantly differ ($p_{\mathrm{Holm}} = 0.5974$), while both were significantly more faithful than LMAC–TD. The statistically supported ordering was

$$\text{AudioLIME} > \{\text{CoughLIME}, \text{L2I}\} > \text{LMAC–TD}.$$

The AudioLIME–LMAC–TD comparison produced a large rank-biserial effect ($r_{\mathrm{rb}} = 0.906$ in favour of AudioLIME).

Sparseness. All six sparseness comparisons were significant after Holm correction ($p_{\mathrm{Holm}} \leq 4.17 \times 10^{-5}$). LMAC–TD produced the sparsest explanations, followed by CoughLIME, AudioLIME, and L2I:

$$\text{LMAC–TD} > \text{CoughLIME} > \text{AudioLIME} > \text{L2I}.$$

The comparisons between LMAC–TD and every alternative had $r_{\mathrm{rb}} \geq 0.999$, indicating highly consistent paired differences across the test samples.

Complexity. All pairwise complexity comparisons were significant after correction ($p_{\mathrm{Holm}} \leq 8.76 \times 10^{-7}$). Since lower complexity is preferred, the complete ordering was

$$\text{AudioLIME} < \text{LMAC–TD} < \text{L2I} < \text{CoughLIME}.$$

AudioLIME produced significantly less complex explanations than every alternative. Although L2I and CoughLIME had similar numerical means, their difference remained significant ($p_{\mathrm{Holm}} = 8.76 \times 10^{-7}$, $|r_{\mathrm{rb}}| = 0.784$).

`AGRI-Fidelity` fidelity. Only two of the six fidelity comparisons remained significant after Holm correction. CoughLIME had significantly lower fidelity than L2I ($p_{\mathrm{Holm}} = 0.0399$, $r_{\mathrm{rb}} = 0.433$ in favour of L2I) and AudioLIME ($p_{\mathrm{Holm}} = 0.0436$, $r_{\mathrm{rb}} = 0.418$ in favour of AudioLIME). The remaining four comparisons were not significant. In particular, LMAC–TD did not significantly differ from L2I ($p_{\mathrm{Holm}} = 0.9760$), CoughLIME ($p_{\mathrm{Holm}} = 0.0809$), or AudioLIME ($p_{\mathrm{Holm}} = 0.9760$).

The compact-letter grouping was

$$\{\text{L2I}, \text{AudioLIME}\}^{a}, \qquad \text{LMAC–TD}^{ab}, \qquad \text{CoughLIME}^{b}.$$

Thus, although LMAC–TD had the largest mean fidelity, the paired analysis does not establish it as significantly better than any competing method.

`AGRI-Fidelity` reliability. LMAC–TD achieved significantly greater reliability than L2I ($p_{\mathrm{Holm}} = 0.00757$, $r_{\mathrm{rb}} = 0.519$), CoughLIME ($p_{\mathrm{Holm}} = 0.0135$, $r_{\mathrm{rb}} = 0.483$), and AudioLIME ($p_{\mathrm{Holm}} = 0.0179$, $r_{\mathrm{rb}} = 0.462$). No significant differences were detected among L2I, CoughLIME, and AudioLIME ($p_{\mathrm{Holm}} \geq 0.4084$). The supported reliability grouping was therefore

$$\text{LMAC–TD} > \{\text{AudioLIME}, \text{CoughLIME}, \text{L2I}\}.$$

This result is especially important because CoughLIME's strong AI and AD did not translate into reliability significantly greater than that of L2I or AudioLIME.

Overall, the Chicken Spurious analysis showed less complete separation than the SmartEars datasets, with $34/48$ pairwise comparisons remaining significant after Holm correction. No explanation method dominated every evaluation dimension. CoughLIME and LMAC–TD formed the leading group for AI, AD, and AG; AudioLIME achieved the highest faithfulness and lowest complexity; and LMAC–TD produced the greatest sparseness and uniquely greater reliability. Fidelity differences were comparatively limited, with only two significant comparisons. These findings demonstrate that strong conventional masking performance does not necessarily correspond to greater causal fidelity or statistically validated reliability, particularly in the presence of a spurious feature.

I Class-Dependent Explanation Fidelity and Reliability With 5 Consensus Models

For our Chicken Spurious dataset, we evaluated explanation fidelity and cross-architecture reliability using 5 consensus models. The analysis included 28 Healthy and 24 Unhealthy test samples and classifier architectures: CNN, LSTM, MLP, ResNet18, and Audio Spectrogram Transformer (AST) Gong et al. (2021). For each model, the top 5% of positive integrated-gradient attributions were converted into a binary explanation mask.

At each time-frequency location, the consensus level λ denotes the fraction of models selecting that location. With five models, $\lambda \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, corresponding to agreement among at least one, two, three, four, or all five architectures, respectively. Statistical reliability was evaluated against independently time-shifted explanation masks. A consensus level was considered reliable when its false-discovery rate satisfied $\mathrm{FDR} \leq 0.1$.

For clarity, a "consensus failure" denotes a sample for which no consensus level passed the FDR criterion. A "reliable consensus" denotes a sample for which the highest observed consensus level passed the criterion. A "partial consensus" denotes a sample for which a lower consensus level passed but the highest observed level did not.

Unlike the baseline-explainer comparisons in Table 9, this experiment does not evaluate or rank a particular explanation method for all SOTA metrics. Instead, it assesses the committee-level behavior of AGRI-Fidelity and examines whether the consensus identified across the five model architectures is statistically distinguishable from agreement expected under the cyclic-shift null model. To ensure consistency with the baseline evaluation, the analysis uses the same 52 held-out test samples used to generate Table 9. We present the principal findings and the corresponding empirical evidence below.

I.1 Explanation reliability is strongly associated with class

Explanation reliability differed substantially between Healthy and Unhealthy samples with the addition of AST to the consensus committee. As shown in Table 12, only $2/28$ Healthy samples (7.1%) exhibited a complete consensus failure, compared with $18/24$ Unhealthy samples (75.0%). Conversely, reliable consensus was obtained for $25/28$ Healthy samples (89.3%) but only $5/24$ Unhealthy samples (20.8%).

Table 13 shows a two-sided Fisher exact test comparing complete consensus failure against the presence of at least one accepted consensus level produced $p = 4.38 \times 10^{-7}$. The corresponding contingency table contained 2 failures and 26 non-failures among Healthy samples, compared with 18 failures and 6 non-failures among Unhealthy samples. The estimated odds of consensus failure were therefore approximately 39 times higher for Unhealthy samples than for Healthy samples.

Table 12: Cross-architecture consensus outcomes by class. Percentages are computed within each class. "No threshold passed" means that no λ satisfied $\mathrm{FDR} \leq 0.1$.

	Healthy ($n = 28$)		Unhealthy ($n = 24$)	
Outcome	Count	Percentage	Count	Percentage
No consensus threshold passed	2	7.1%	18	75.0%
Reliable highest-order consensus	25	89.3%	5	20.8%
Reliable lower-order consensus only	1	3.6%	1	4.2%
Total	28	100.0%	24	100.0%

A second Fisher exact test compared near-unit FDR behavior against all other FDR trends. Near-unit FDR occurred in $0/28$ Healthy samples and $15/24$ Unhealthy samples, producing $p = 2.92 \times 10^{-7}$. After Holm correction for these two categorical comparisons, both tests remained significant with $p_{\mathrm{Holm}} = 5.84 \times 10^{-7}$.

Table 13: Two-sided Fisher exact tests for associations between class and categorical explanation-reliability outcomes. Holm correction was applied across the two tests. Entries in the Healthy and Unhealthy columns are event/non-event counts.

Outcome	Healthy	Unhealthy	Raw p	Holm-adjusted p
No consensus threshold passed	$2/26$	$18/6$	4.38×10^{-7}	5.84×10^{-7}
Near-unit FDR trend	$0/28$	$15/9$	2.92×10^{-7}	5.84×10^{-7}

These results provide strong evidence that explanation reliability is not independent of class. In particular, unreliable or chance-like cross-architecture agreement was concentrated in the Unhealthy class. We suspect that this behavior arises because AST disagreed with the other committee models for the majority of the test samples.

I.2 Healthy explanations are substantially more consistent across models

The shape of the FDR curves provided additional evidence of greater cross-model consistency for Healthy samples. As shown in Table 14 shows strictly downward FDR trend was observed for $27/28$ Healthy samples (96.4%), compared with only $5/24$ Unhealthy samples (20.8%). No Healthy sample exhibited a near-unit FDR curve, whereas this behavior occurred in $15/24$ Unhealthy samples (62.5%).

Table 14: FDR trend categories by class. A downward curve indicates that requiring agreement among more architectures progressively reduced the estimated false-discovery rate. A near-unit curve indicates that observed agreement remained similar to agreement under the time-shifted null model.

	Healthy ($n = 28$)		Unhealthy ($n = 24$)	
FDR trend	Count	Percentage	Count	Percentage
Downward	27	96.4%	5	20.8%
Near-unit	0	0.0%	15	62.5%
Mixed or non-monotonic	1	3.6%	4	16.7%
Total	28	100.0%	24	100.0%

A downward FDR curve means that agreement among a larger number of architectures became progressively less compatible with the shifted-null explanations. Thus, the predominance of downward curves among Healthy samples indicates that their salient regions were spatially more consistent across architectures.

A downward trend does not necessarily imply that the FDR crossed the predefined 0.1 threshold. This distinction explains why 27 Healthy samples had downward curves but only 25 were assigned to

the reliable highest-order consensus category. Nevertheless, the combined prevalence of downward FDR curves and accepted high-order consensus demonstrates that the Healthy explanations were substantially more reproducible across models.

I.3 Unhealthy explanations are frequently influential but architecture-dependent

Fidelity was defined as the absolute probability change caused by replacing a model's selected top-5% attribution region with the baseline:

$$\text{Fidelity}_{m,i} = \left| p_m(\mathbf{x}_i) - p_m(\mathbf{x}_{i,\text{masked}}) \right|,$$

where m denotes a model and i denotes a test sample. Larger values therefore indicate that the model's selected region had a greater influence on its prediction.

As shown in Table 15, CNN, LSTM, MLP, and ResNet18 all exhibited greater average fidelity for Unhealthy samples than for Healthy samples. The difference was most pronounced for CNN and LSTM. AST was an exception, exhibiting uniformly low fidelity in both classes.

Table 15: Mean attribution fidelity by architecture and class. Values denote the mean absolute probability change after masking each model's top-5% positive integrated-gradient region.

Architecture	Healthy	Unhealthy
CNN	0.111602	0.679755
LSTM	0.001011	0.264838
MLP	0.051632	0.108750
ResNet18	0.120937	0.284878
AST	0.008978	0.007947

The high Unhealthy fidelity values show that the identified regions were often consequential for the corresponding model predictions. However, the FDR results show that these consequential regions were not spatially consistent across architectures. Therefore, the dominant Unhealthy pattern was not an absence of influential features. Rather, different architectures identified different influential regions.

This combination supports an interpretation of architecture-dependent evidence use: each architecture could identify a region that affected its own prediction, while the locations of those regions failed to agree beyond the shifted-null expectation.

The AST results should be interpreted separately. Its low fidelity indicates that masking the selected top-5% AST attribution region produced little change in the AST prediction under the present perturbation protocol. This does not imply that the AST classifier was inaccurate. It instead indicates that the selected AST region had low perturbation fidelity, potentially because AST evidence was more spatially distributed or because the attribution and masking procedure interacted differently with the transformer architecture.

I.4 Fidelity alone is insufficient for assessing explanation quality

The Unhealthy trend-stratified results provide direct evidence that fidelity alone does not identify reliable explanations. As shown in Table 16, samples with near-unit FDR had higher mean fidelity than samples with downward FDR for every architecture. This occurred even though near-unit FDR indicates that cross-model agreement was approximately indistinguishable from the shifted-null expectation.

For example, the mean CNN fidelity in the near-unit group was 0.787425, compared with 0.397999 in the downward group. Similarly, LSTM fidelity was 0.403936 in the near-unit group but only 0.000919 in the downward group. Thus, large prediction changes after masking did not imply that different architectures selected the same evidence.

These findings demonstrate that fidelity evaluates whether an explanation is influential for its own model, not whether that explanation is stable, reproducible, or shared across architectures. Consequently, fidelity should not be interpreted as a complete measure of explanation quality.

Table 16: Mean model-specific fidelity for Unhealthy samples, stratified by FDR trend. The near-unit group had poor cross-model reliability despite having comparatively high fidelity.

Architecture	Downward ($n = 5$)	Near-unit ($n = 15$)	Mixed ($n = 4$)
CNN	0.397999	0.787425	0.628189
LSTM	0.000919	0.403936	0.073118
MLP	0.000346	0.162515	0.042637
ResNet18	0.128104	0.344533	0.257140
AST	0.003081	0.010082	0.006027

I.5 Consensus reliability adds information not captured by fidelity

Fidelity and consensus reliability quantify different explanation properties. Fidelity measures the prediction sensitivity of an individual model to its own explanation mask. Consensus reliability measures whether the salient regions identified by multiple architectures overlap more than expected under a time-shifted null model.

The near-unit Unhealthy group demonstrates why both quantities are necessary. These samples exhibited the highest average fidelity for most architectures, but their cross-model agreement remained approximately chance-like. A fidelity-only evaluation would therefore classify many of these explanations as strong because masking them substantially changed their respective model predictions. The consensus analysis revealed that the explanations were nevertheless not reproducible across architectures. The joint interpretation is summarized in Table 17.

Table 17: Interpretation of fidelity and consensus reliability as complementary explanation properties.

Fidelity	Consensus reliability	Interpretation
High	High	Influential and reproducible evidence
High	Low	Influential but architecture-dependent evidence
Low	High	Shared region with limited prediction effect
Low	Low	Neither influential nor reproducible under the protocol

The dominant Unhealthy behavior corresponded most closely to the high-fidelity, low-reliability condition, particularly for CNN, LSTM, MLP, and ResNet18. In contrast, Healthy samples exhibited much stronger cross-architecture reliability, although their mean fidelity was generally lower.

Accordingly, consensus reliability contributes information that cannot be recovered from fidelity alone. The two measures should be reported jointly: fidelity establishes whether a selected region affects a model's prediction, while reliability establishes whether the explanatory evidence is stable across architectures.

Taken together, **the results show that explanation behavior was strongly class-dependent.** Healthy samples generally produced stable and statistically reliable cross-architecture explanations. Unhealthy samples, by contrast, frequently produced model-specific regions that strongly affected individual predictions but did not align reliably across architectures.

The Unhealthy explanations were often **individually influential but collectively inconsistent**. This distinction demonstrates why fidelity alone is insufficient and why cross-model consensus reliability provides an essential complementary measure of explanation quality.

J Illustrative Numerical Example

Throughout our paper, consistent with our goals and motivation, we evaluate `AGRI-Fidelity` against other XAI metrics rather than comparing classifier architectures or explanation algorithms. The models and explainers used in the framework serve only as sources of attribution masks required to calculate `AGRI-Fidelity`. Accordingly, differences in their individual predictions or attribution maps are not interpreted as evidence that one model or explainer is superior to another. Instead, we examine whether conventional fidelity alone is sufficient to characterize an explanation and

whether the reliability component introduced by `AGRI-Fidelity` provides additional information about statistically validated cross-source agreement. Below, we illustrate how `AGRI-Fidelity` works using a numerical example.

Consider an input spectrogram

$$X \in \mathbb{R}^{T \times F}$$

containing $T \times F = 200$ time–frequency bins. Let

$$\mathcal{M} = \{M_1, M_2, M_3\}$$

denote a simplified committee consisting classifiers. For each model M_k, the top 5% of the corresponding attribution map is retained and binarized to obtain

$$E_k \in \{0,1\}^{T \times F}.$$

Because the spectrogram contains 200 bins, each binary mask contains

$$0.05 \times 200 = 10$$

selected bins.

Let $E_* \in \{0,1\}^{T \times F}$ denote the binarized explanation created by $M*$, that is for which AGRI-Fidelity components need to be computed. In this example, the selected bins are

$$\begin{aligned} E_1 &= \{4,5,6,7,8,20,21,22,40,60\} \\ E_2 &= \{4,5,6,7,8,20,21,23,41,61\} \\ E_3 &= \{4,5,6,7,8,20,22,24,42,62\} \\ E_* &= \{4,5,6,7,8,20,21,22,43,63\} \end{aligned}$$

The consensus is computed over the four masks

$$\mathcal{E} = \{E_1, E_2, E_3, E_*\}.$$

Thus, the observed pixel-wise consensus is

$$S_{\text{obs}}(t,f) = \frac{1}{4}\left[E_1(t,f) + E_2(t,f) + E_3(t,f) + E_*(t,f)\right].$$

The set of nonzero consensus tiers is therefore

$$\Lambda = \left\{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{4}{4}\right\} = \{0.25, 0.50, 0.75, 1.00\}.$$

Bins

$$\{4,5,6,7,8,20\}$$

are selected by all four masks and consequently satisfy

$$S_{\text{obs}}(t,f) = 1.$$

Bins 21 and 22 are each selected by three of the four masks and therefore satisfy

$$S_{\text{obs}}(t,f) = 0.75.$$

The remaining selected bins occur in only one mask and have consensus

$$S_{\text{obs}}(t,f) = 0.25.$$

For each $\lambda \in \Lambda$, the tier-specific observed consensus mask is

$$C_\lambda(t,f) = \mathbb{I}\left[S_{\text{obs}}(t,f) \geq \lambda\right].$$

The corresponding observed counts are

$$N_{\text{obs}}(0.25) = \sum_{t,f} C_{0.25}(t,f) = 18,$$
$$N_{\text{obs}}(0.50) = \sum_{t,f} C_{0.50}(t,f) = 8,$$
$$N_{\text{obs}}(0.75) = \sum_{t,f} C_{0.75}(t,f) = 8,$$
$$N_{\text{obs}}(1.00) = \sum_{t,f} C_{1.00}(t,f) = 6.$$

To determine whether this agreement exceeds chance alignment, each mask is independently shifted along the time axis. For permutation $b \in \{1,\ldots,B\}$, let

$$E_k^{(b)}(t,f) = E_k\left((t + \Delta_k^{(b)}) \bmod T, f\right),$$

where $\Delta_k^{(b)}$ is a randomly sampled temporal offset. The mask E_* is shifted independently in the same manner. The null consensus for permutation b is

$$S_{\text{null}}^{(b)}(t,f) = \frac{1}{4}\left[E_1^{(b)}(t,f) + E_2^{(b)}(t,f) + E_3^{(b)}(t,f) + E_*^{(b)}(t,f)\right].$$

The corresponding tier-specific null mask is

$$C_\lambda^{\text{null},(b)}(t,f) = \mathbb{I}\left[S_{\text{null}}^{(b)}(t,f) \geq \lambda\right].$$

Suppose that $B = 50$ cyclic temporal permutations produce the following average null counts:

$$\overline{N}_{\text{null}}(0.25) = 33.7,$$
$$\overline{N}_{\text{null}}(0.50) = 4.5,$$
$$\overline{N}_{\text{null}}(0.75) = 1.7,$$
$$\overline{N}_{\text{null}}(1.00) = 0.1,$$

where

$$\overline{N}_{\text{null}}(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\sum_{t,f} C_\lambda^{\text{null},(b)}(t,f).$$

Following the empirical FDR formulation, we compute

$$\widehat{\text{FDR}}(\lambda) = \min\left\{1,\ \frac{\overline{N}_{\text{null}}(\lambda)}{N_{\text{obs}}(\lambda)}\right\}.$$

The FDR values at the four consensus tiers are therefore

$$\widehat{\mathrm{FDR}}(0.25) = \min\left\{1, \frac{33.7}{18}\right\} = 1.000,$$
$$\widehat{\mathrm{FDR}}(0.50) = \frac{4.5}{8} = 0.562,$$
$$\widehat{\mathrm{FDR}}(0.75) = \frac{1.7}{8} = 0.212,$$
$$\widehat{\mathrm{FDR}}(1.00) = \frac{0.1}{6} = 0.0166.$$

Thus, the resulting FDR vector is

$$\widehat{\mathbf{FDR}} = [1.000,\ 0.562,\ 0.212,\ 0.0166].$$

The highest observed consensus tier is

$$\lambda_{\max} = 1.00.$$

The reliability score is consequently

$$\begin{aligned} R &= 1 - \widehat{\mathrm{FDR}}(\lambda_{\max}) \\ &= 1 - \widehat{\mathrm{FDR}}(1.00) \\ &= 1 - 0.0166 \\ &= 0.984. \end{aligned}$$

To evaluate fidelity, let p_{orig} denote the target-class confidence for the original spectrogram X, and let p_{mask} denote the confidence after removing the CoughLIME-selected region. Suppose that

$$p_{\mathrm{orig}} = 0.91 \qquad \text{and} \qquad p_{\mathrm{mask}} = 0.36.$$

Let the fidelity score be defined as

$$\begin{aligned} F &= |p_{\mathrm{orig}} - p_{\mathrm{mask}}| \\ &= |0.91 - 0.36| \\ &= 0.550. \end{aligned}$$

Accordingly, AGRI-Fidelity assigns the evaluated CoughLIME explanation the two component scores

$$\boxed{F = 0.550, \qquad R = 0.984}.$$

The fidelity score indicates that the $M*$-selected region has a substantial effect on the target-class prediction. The reliability score indicates that the full-consensus region shared by $M*$ and the three committee models occurs substantially more frequently in the observed explanations than under random temporal alignment. This example illustrates how AGRI-Fidelity distinguishes causal influence, measured by fidelity, from statistically validated cross-model agreement, measured by reliability.