Information Sheet

Systematic Vulnerability Audit of Post-Hoc XAI under Common Corruptions: A Factor Analysis Across Vision Benchmarks

Minyeong Kim | Independent Researcher, Gyeonggi-do, South Korea | win198@naver.com

Target journal: *Machine Learning* (Springer Nature, journal 10994).

1. Area and type. *Explainable AI (XAI) attribution methods under distribution shift.* This is an empirical methodology paper that introduces an evaluation protocol and validates it through a factorial audit. It is not a benchmark paper or an algorithm paper — the contribution is the protocol itself, released as an open reference implementation. Primary sub-field: ML evaluation methodology; secondary: computer vision (CIFAR-10-C / CIFAR-100-C) and statistical methodology (factorial ANOVA, mixed-effects, FDR corrections, bounded-metric transformations).

2. Main contribution and novelty. We contribute three artifacts: (i) *the protocol* — the first systematic evaluation protocol for post-hoc attribution stability under distribution shift, specifying factorial axes (method × corruption × severity × architecture × dataset), five stability metrics (Spearman, rank agreement, sign agreement, L_2, Jensen–Shannon), and a four-way statistical decision rule (parametric ANOVA, non-parametric Friedman, linear mixed-effects model, TOST equivalence); (ii) *the reference implementation* — a single-command-reproducible codebase (Zenodo v1.1.0, 10.5281/zenodo.19689329) with Fisher-z, Benjamini–Yekutieli, Holm–Bonferroni, and ω^2 sensitivity scripts; (iii) *the empirical characterisation* — a vulnerability ranking (SmoothGrad > Integrated Gradients > GradientSHAP) consistent across 96% of 150 tested conditions, validated by four-way statistical convergence, operationalised as a deployment-ready decision rule. *What is novel:* prior XAI evaluation has operated predominantly on clean inputs (e.g., Li et al. 2023 M4; Rozanec et al. 2025 XAI-Units; Moradi et al. 2025). The present study is the first to provide a systematic, cross-method, cross-architecture evaluation protocol for the corrupted-input regime, with a $5 \times 15 \times 5 \times 2 \times 2$ factorial yielding over 760,000 clean–corrupted attribution pairs.

3. Relationship to prior work. No part of this paper has been published before; there is no conference version. The closest recent work is Repetto et al. (*Machine Learning*, 2025, DOI 10.1007/s10994-025-06919-6), which evaluated XAI robustness under natural and adversarial corruption in a single domain (medical image recognition). The present manuscript generalises that single-domain enquiry into a cross-method, cross-architecture, cross-dataset evaluation protocol at CIFAR-scale, with explicit reusability instructions (manuscript §6.4, four-step procedure for evaluating any new attribution method). A concurrent work, Anani et al. (ICML 2025, arXiv:2506.15499), provides pixel-level certification for attribution maps via randomised smoothing; our work is complementary, supplying the empirical validation substrate against which such theoretical certificates must be calibrated.

4. Empirical and theoretical support. *Empirical scale:* 760,000+ clean–corrupted attribution pairs generated from a $5 \times 15 \times 5 \times 2 \times 2$ factorial design, run on consumer-class NVIDIA RTX 3080 Ti hardware without institutional compute. Every numerical claim in the Results section is traceable to the public codebase. *Statistical support:* four convergent test families — (i) two-way factorial ANOVA with η^2 and ω^2 reported in parallel; (ii) non-parametric Friedman with Wilcoxon post-hoc; (iii) linear mixed-effects with image random intercept; (iv) TOST equivalence for the H3 architecture-invariance null. Multiple-comparison corrections: Benjamini–Hochberg (default), Benjamini–Yekutieli (sensitivity), Holm–Bonferroni (conservative). A Fisher-z sensitivity analysis (Warton & Hui, 2011) confirms that bounded-metric (Spearman $\in [-1, 1]$) inference preserves all four hypothesis verdicts. *Theoretical support:* off-manifold extrapolation (Janzing et al. 2020; Frye et al. 2020) and axiomatic attribution framing (Sundararajan et al. 2017; Lundberg & Lee 2017) motivate the method selection and the interpretation of the H3 boundary-condition finding under CIFAR-100 complexity.

5. Reproducibility. Reproducibility package: attribution pipeline, aggregated statistics JSON (ANOVA / Friedman / bootstrap Tukey / mixed-effects / TOST / Fisher-z / BY / ω^2), figure-generation code, pinned `requirements.txt`. Permanent archives: GitHub (https://github.com/minyeong5691-svg/paper2-xai-corruption-audit) and Zenodo v1.1.0 (10.5281/zenodo.19689329; concept DOI 10.5281/zenodo.19688326). CIFAR-10-C / CIFAR-100-C benchmarks and the four pretrained ResNet-50 / ViT-B/16 checkpoints are publicly available (specific URLs and accuracies in the manuscript's *Data Availability* statement). Raw attribution tensors (≈ 120 GB) are not deposited due to aggregate size; full reproduction requires re-running the pipeline on the public benchmarks.

6. Generative AI and author information. The author used Anthropic's Claude (Opus 4.7) for English-language drafting, implementation assistance, bibliographic verification, and structural review. Research design, hypotheses, analytical plan, and interpretation reflect the author's own judgement. No AI system is listed as an author. Full disclosure in the manuscript's *Statements and Declarations*. The submission is single-author, unfunded, with no competing interests; the author is an independent researcher based in Gyeonggi-do, South Korea, with no institutional affiliation.