Research Square

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

Spatial resolution alters saliency stability comparisons in skin lesion classifiers across architectures and external datasets

Binod Bhattarai[1]
Islington College
Dipesh Gurung[2]
dipesh.gurung@lbef.edu.np
Lord Buddha Education Foundation
R. N. Thakur[2]
Lord Buddha Education Foundation

Article

Keywords: saliency maps, spatial resolution, skin lesion classification, Integrated Gradients, Grad-CAM, external evaluation

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11371085/v1

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

Additional Declarations: No competing interests reported.

Abstract

Saliency stability comparisons may depend on spatial resolution, numerical eligibility and prediction quality. We compared Grad-CAM and positive Integrated Gradients in HAM10000-trained ResNet-18 and DenseNet-121 classifiers. A frozen ResNet audit and exploratory DenseNet replication evaluated the same 300 corrected lesion groups across three training seeds. Twelve photometric perturbations were assessed at input and pooled 7×7 resolutions. The outcome was the condition-balanced change in the method stability contrast after pooling. Numerical eligibility required completeness and successive-budget convergence. We reported paired-group bootstrap intervals and worst-case missingness bounds. Exploratory external evaluation used 100 BCN20000 groups and 100 PAD-UFES-20 patient groups per architecture. Resolution interactions were 0.287–0.351 on HAM and 0.293–0.334 externally. Eligible-map intervals were positive in all ten summaries, but worst-case bounds included zero in three. External accuracy was 24–32%. On DenseNet PAD, Integrated Gradients insertion advantage changed from 0.180 with black replacement to − 0.017 with blurred replacement. Spatial representation materially affected the comparison, while numerical missingness, replacement sensitivity and poor diagnostic transfer limited interpretation. Stability should be accompanied by matched-resolution comparisons and complementary quality controls.

Introduction

Saliency maps summarize image regions associated with a model output. Grad-CAM derives coarse localization from convolutional features, whereas IG attributes a prediction to input features relative to a reference [1, 2]. These representations differ in spatial detail. Consequently, comparing their perturbation stability at input resolution may reflect both the attribution method and the spatial structure of the resulting map.

Existing research already establishes important limitations of saliency evaluation. Smoothing can alter stability and fidelity [3]; convolutional architectural priors can affect trained-versus-random explanation similarity [4]; and attribution evaluation can depend on resolution and perturbation-induced distribution shift [5]. Medical saliency audits also show that localization, repeatability, reproducibility and parameter sensitivity need not agree [6]. Our purpose is therefore not to introduce smoothing as a new principle or to propose a new attribution algorithm.

We investigate a narrower question: how does matching spatial resolution change the apparent stability difference between Grad-CAM and positive IG in skin-lesion classifiers? We quantify this interaction across two architectures, training seeds and external imaging settings. We additionally examine numerical eligibility and masking controls, and retain diagnostic transfer failures as part of the interpretation. The intended contribution is a reproducible empirical audit of a comparison design, rather than a clinically validated diagnostic system.

Stability as one component of explanation quality

An explanation can be stable without being informative. For example, an image-independent map produces the same ranking under every input perturbation, yet provides no image-specific information. A useful evaluation therefore starts by specifying what the explanation is intended to do. The distinction between evaluating an explanatory representation and evaluating its use in a human task is central to interpretable-machine-learning research [7]. Our study concerns a representation-level comparison; it does not test whether a clinician makes better decisions after viewing a map.

Parameter-randomization checks examine whether explanations depend on learned parameters [8]. Input-invariance arguments examine whether transformations that leave the prediction function unchanged can nevertheless alter an attribution [9]. These are different from robustness under perturbations that may change the model output [10]. The distinctions matter here because our photometric conditions can change predictions even though the explanation target remains fixed. We consequently report both the full eligible sample and an unchanged-prediction secondary analysis rather than assume that all image changes are prediction preserving.

SmoothGrad averages gradient information over noisy inputs [11]. Our Gaussian controls instead filter an already generated attribution map in the image plane. The two operations should not be described interchangeably. Ye and Farnia analyze stability to training randomness and a fidelity trade-off in smoothed gradient interpretations [3]; our fixed-model perturbation comparison addresses a different empirical setting. This distinction narrows the potential contribution but does not establish theoretical novelty.

Perturbation benchmarks and evaluation design

Perturbation explanations and their evaluations depend on what is removed, how missing pixels are replaced and which output is measured. Meaningful Perturbation and RISE exemplify explicit image-masking approaches [12, 13]. Remove-and-retrain evaluation additionally addresses how removing information changes the learning problem [14]. We did not retrain masked-image models, and our inference-only deletion/insertion results should not be interpreted as estimates from that design.

Formal work on sensitivity and infidelity shows why making an explanation insensitive is insufficient on its own [15]. Evaluation toolkits such as Quantus support examining multiple criteria and parameter choices rather than relying on a single visually appealing map or score [16]. We implement a restricted set of such checks using frozen sources and saved records; we do not claim to replace these toolkits or to introduce a general benchmark for all attribution methods.

Medical saliency research has already assessed multiple quality criteria. Venkatesh and colleagues evaluate localization, repeatability, reproducibility and parameter sensitivity in musculoskeletal imaging [6]. Saarela and Georgieva compare Integrated Gradients and local model-agnostic explanations on HAM10000 [17]. Our medical setting is therefore not itself novel. We focus on the measured interaction between method and resolution, and on the eligibility and transfer conditions under which that interaction is interpretable.

Research question and scope

We ask whether the Grad-CAM-minus-IG perturbation-stability contrast changes after both maps are represented on the same coarse grid. A secondary question is whether the observation remains visible across architectures, training seeds and external imaging settings, including settings with poor predictive transfer. A further question is how numerical exclusions and replacement choices alter the strength of the interpretation. These questions concern the evaluation procedure rather than the development of a new diagnostic model.

There are three intended empirical outputs. First, we estimate the resolution interaction with uncertainty at the group level. Second, we provide spatial controls that separate simple smoothing from the effects of shared spatial rearrangement. Third, we expose numerical missingness and predictive transfer failure rather than treating completed attribution runs as automatically valid explanations. The original ResNet analysis was held out; later extensions remain exploratory throughout this paper.

Methods

Study sequence and analysis status

The original ResNet test protocol was frozen before test inference. Seed 42 was primary, with seeds 43 and 44 prespecified sensitivity runs. Development data informed the endpoint: a proposed severity-slope analysis was replaced before test access because its all-conditions completeness requirement retained only 15 of 42 development images. The final endpoint averages condition-specific contrasts and retains an image in conditions where it is eligible.

DenseNet replication, spatial controls, external evaluation and correctness subgroup analyses followed inspection of ResNet test results. Their settings were frozen before their respective inference runs where applicable, but these analyses remain exploratory. There was no public prospective registration. We did not select a model seed using test performance or adapt the models to external outcomes.

Data grouping and sampling

HAM10000 contains 10,015 images of seven lesion categories [18]. Our corrected manifest retained all images and incorporated lesion identifiers, exact duplicates and documented duplicate/longitudinal links from published dataset-quality work [19]. The training, validation and test partitions contained 7,014, 1,482 and 1,519 images, respectively. The corrected test partition contained 1,117 leakage groups. Patient identifiers were unavailable, so grouping should not be interpreted as verified patient independence. The output labels were actinic keratoses and intraepithelial carcinoma (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (vasc). SCC denotes squamous cell carcinoma in the external exclusions.

We selected 300 test groups by deterministic SHA256 ordering of group identifiers, followed by deterministic selection of one representative image per group. Selection preceded test inference and used no class quotas or predictions. The explanation estimand is group-weighted over this sampled test population, distinct from image-level classification on all 1,519 test images. The 300-group budget was a feasibility decision, not a clinical power calculation. The same explanation sample was used for both architectures and all seeds.

External data were available locally from BCN20000, a dermoscopic collection [20], and PAD-UFES-20, a smartphone clinical-image collection [21]. We audited 18,946 BCN and 2,298 PAD files by decoding and hashing. No exact file or decoded-RGB matches to HAM were found. A perceptual screen flagged 315 BCN and 11 PAD images; these were unresolved similarity candidates, not confirmed duplicates. Entire connected groups containing such flags were conservatively quarantined. Lesion, patient where available, and identical-pixel links were merged before selection. The local metadata supplied 5,440 BCN lesion identifiers; completeness relative to the published collection remains a provenance limitation.

Diagnoses and grouping identifiers were reconciled with local source metadata. We excluded unsupported or mixed-diagnosis groups. BCN covered seven corresponding categories, excluding SCC, melanoma metastasis, scar, other diagnoses and solar lentigo. PAD covered five corresponding categories, excluding SCC; actinic keratosis was treated as a subset correspondence to HAM's broader akiec class. The frozen external sample contained 100 groups per dataset: BCN had 14–15 images per class, and PAD had 22 akiec, 21 bcc, 21 bkl, 15 mel and 21 nv images. The PAD allocation reflected only 15 eligible melanoma patient groups under the selection rules.

These approximately class-balanced samples do not estimate natural-prevalence accuracy. External datasets had previously been used in a separate BioPredict project and are external to the present HAM-trained models rather than globally untouched research resources. Residual overlap risk remains despite screening.

Figure 1. **Experimental pipeline and resolution comparison.** Panel A shows the dataset inventories, grouped explanation cohorts and frozen HAM-trained classifiers. External inference uses seed 42 without retraining. Panel B shows fixed-target attribution under 12 photometric perturbations, IG numerical checks, and native versus pooled stability comparisons. Unresolved numerical cases are counted and propagated exclusions remain separate. Here, ρ denotes clean-to-perturbed Spearman correlation. Complementary controls assess clean faithfulness, HAM mask overlap and prediction correctness. Patient independence is not established for HAM or BCN; external similarity flags were conservatively quarantined.

Models and preprocessing

ResNet-18 [22] and DenseNet-121 [23] used ImageNet-pretrained weights [24] and seven-class output heads. Each architecture was trained separately with seeds 42, 43 and 44 for 12 epochs using batch size 32, AdamW [25], learning rate 10^{-4} and weight decay 10^{-4}. Cross-entropy weights were proportional to the

inverse square root of training-class frequency and normalized to mean one. There was no weighted sampler.

Images were converted to RGB, resized as whole images to 224×224, and normalized using ImageNet means and standard deviations. Training-only augmentation comprised horizontal and vertical flips. Checkpoints were selected separately by highest validation macro-F1, retaining the earlier epoch for ties. All explanation analyses used evaluation mode with frozen parameters and non-inplace activations. External analyses reused seed 42 checkpoints without retraining. Inference used PyTorch [26], torchvision and Captum [27] on an RTX 3050 GPU with 8 GB memory.

Attribution and perturbations

For each image and model, the target was the clean predicted-class logit, held fixed across conditions even if the perturbed prediction changed. Grad-CAM targeted the final residual block in ResNet and final normalization feature output in DenseNet; this architecture-specific layer choice was explicit. Coarse Grad-CAM maps were bilinearly resized to 224×224 with aligned corners disabled. IG used a black RGB reference and Gauss–Legendre integration. We retained positive input-channel contributions and normalized nondegenerate maps to unit mass; conclusions therefore do not generalize automatically to signed IG.

We compared processed 224×224 maps with adaptive-average-pooled 7×7 maps. The term native denotes the processed input-resolution map, not the original feature-map resolution of Grad-CAM. Perturbations were brightness factors 0.9, 0.8 and 0.7; gamma values 1.1, 1.2 and 1.3; Gaussian blur standard deviations 0.5, 1.0 and 1.5 with kernel width 11; and RGB noise standard deviations 0.01, 0.02 and 0.03. Noise realization was deterministically image-seeded and shared across doses and model seeds. Clipping was recorded. No geometric transformations were evaluated, and clinical acquisition plausibility was not adjudicated by dermatologists.

IG integration budgets were 256, 512, 1,024 and 2,048 steps. A map was accepted at the earliest budget of at least 512 satisfying both a scaled completeness residual ≤ 0.01 and native/pooled Spearman correlations ≥ 0.995 against the preceding budget. The residual denominator was the larger of one and the absolute target-logit difference. An unresolved clean IG reference also excluded its perturbation comparisons; these propagated exclusions were counted separately from maps' own numerical failures. Nonfinite, zero-mass and constant-map correlations remained undefined.

Outcomes and uncertainty

Let ρ denote Spearman correlation between clean and perturbed maps. For image i and condition c, we defined

$$D\square\square = (\rho\, GC, 224 - \rho\, IG, 224) - (\rho\, GC, 7 - \rho\, IG, 7) \quad (1)$$

Only comparisons valid for both methods and resolutions contributed. We averaged D within each of the 12 conditions and then averaged condition means equally. Positive values indicate that pooling reduces the Grad-CAM-minus-IG stability contrast. They do not directly measure diagnostic accuracy, faithfulness or equivalence between methods.

Paired-group percentile intervals followed the bootstrap principle [28] and used 2,000 resamples, preserving each group's condition vector and method pairing. We also calculated worst-case bounds by assigning missing contrasts − 4 or + 4. These are sensitivity bounds, not confidence intervals. Unchanged-prediction analyses were secondary. Seeds and datasets were reported separately rather than combined as independent studies. Exploratory comparisons had no multiplicity adjustment.

Additional controls and classification checks

On 42 validation groups, exploratory analyses compared grids of width 7, 14, 28, 56, 112 and 224; Gaussian map filters with standard deviations 1, 4 and 16; and shared pixel-shuffle pooling. Fixed and independently redrawn random maps provided stability controls. Gaussian map filtering is distinct from SmoothGrad's input-noise averaging.

Clean faithfulness used deletion/insertion probability AUC across 11 fractions of 49 equal-area regions, comparing attribution ordering with five fixed random orders under black and blurred-image replacement. Random-order results were averaged within image. Positive deletion advantage was random minus attribution AUC; positive insertion advantage was attribution minus random AUC. Descriptive intervals resampled groups. Native and pooled maps share region rankings by construction under this 7×7 procedure, so their agreement is not evidence that pooling preserves pixel-level fidelity. Replacement may induce distribution shift.

HAM lesion-mask localization was recorded as an exploratory control. External masks were unavailable. Localization was treated descriptively because lesion masks are not ground truth for the model reasoning process. We report seed 42 means below without a clinical-validity claim. Classification used accuracy, macro-F1 and balanced accuracy as applicable, computed with scikit-learn [29]. PAD scoring retained all seven prediction outputs despite only five supported truth classes. A subsequent integrity check verified source labels, image hashes, checkpoint label order, preprocessing, confusion matrices and fresh external predictions. Correct/incorrect explanation subgroups were post-hoc descriptive analyses.

Independent units and selection restrictions

The full external inventory was larger than the binary subsets previously used in BioPredict. We therefore began from the complete manifest rather than reuse the binary test files as if they represented a seven-class population. All 21,244 image paths were readable and their newly calculated file hashes matched the local inventory. Within-dataset identical-pixel screening identified 47 BCN duplicate groups and 16 PAD duplicate groups. These links contributed to connected-group construction; they were not merely removed by choosing arbitrary filenames.

We constructed external units before selecting representatives. PAD patient links connected images of the same recorded patient, including different lesions, while lesion and exact-pixel links were also included. BCN used lesion and exact-pixel links because patient identifiers were unavailable. A connected unit was excluded if it contained an unresolved HAM similarity flag, an unreadable image, or a diagnosis that did not map unambiguously into the supported endpoint. Groups with multiple diagnosis categories were excluded under the declared homogeneous-group rule. This sacrifices coverage, particularly of patients with several different lesions, in exchange for a simple independent-unit analysis.

Eligible group counts after these rules were 307 akiec, 1,253 bcc, 375 bkl, 50 df, 663 mel, 1,580 nv and 46 vasc for BCN. PAD had 297 akiec, 289 bcc, 90 bkl, 15 mel and 118 nv groups. These are local audit counts, not revised dataset-provider totals. Selection proceeded in sorted label order with approximately equal quotas, skipping classes whose eligible pool was exhausted, followed by deterministic hash ordering within class. No model predictions were used in this procedure.

Reproducibility and integrity checks

Protocol documents, sample manifests, core source files and checkpoint files were hashed before inference. Each image result was saved atomically and linked to its lock identity. Resumption was permitted only when saved results matched the same lock. This enabled interruptions to preserve completed images without changing the scientific sample. Output folders for the two architectures and external datasets were separate, and source-hash checks were performed before stages. These controls protect against accidental changes but are not a substitute for public protocol registration or independent replication.

The DenseNet attribution adapter used non-inplace activations and was checked against the original evaluation forward pass. A one-epoch smoke run established memory feasibility before full training. Numerical failures were retained as observations of the evaluation process rather than exceptions to be silently retried with relaxed criteria. No tolerance was changed after inspecting a test map. The main IG budget cap remained 2,048 even though an exploratory validation diagnostic later considered higher budgets.

For the external accuracy investigation, labels were reconstructed from source diagnoses and compared with the frozen sample. Checkpoint configuration label order, strict state loading, input conversion and preprocessing were checked. Confusion matrices were rebuilt from per-image records, and classification inference was rerun on every external image-model pair. No prediction disagreed with its saved result. This verifies the checked execution path; it does not independently establish the clinical correctness of all original annotations or exclude every shared systematic implementation problem.

Interpretation of missingness and control intervals

The primary eligible-map contrast requires four finite correlations for an image-condition pair. Own IG numerical failure, clean-reference propagation and degenerate maps are distinct reasons why that contrast may be unavailable. Denominators therefore depend on the question: 3,900 or 1,300 counts refer to all 13 clean/perturbed IG evaluations, whereas primary contrast eligibility concerns 12 perturbation

comparisons per image. An overall execution-completion percentage is not an eligibility percentage. We retain both to avoid confusing computational completion with statistical coverage.

Bootstrap intervals summarize variation over sampled groups conditional on the observed eligibility pattern. The separate missingness bounds deliberately allow extreme contrasts for unavailable pairs and do not assume those pairs resemble observed ones. No confidence level attaches to these worst-case ranges. For clean faithfulness, five random orders were averaged inside each group before group resampling, so the five orders were not treated as five independent lesions. Marginal method means can involve different eligible sets; we do not infer paired method superiority from them.

Results

Resolution interaction across architectures

All six HAM runs yielded positive eligible-map interactions (Table 1; Fig. 2). DenseNet showed the same direction as the original ResNet audit. These results quantify sensitivity to spatial representation, rather than establishing universal superiority of either method.

Table 1
HAM classification and resolution interaction. Classification includes 1,519 test images; explanations include 300 corrected groups per seed.

Architecture	Seed	Accuracy	Macro-F1	Interaction	95% bootstrap interval	Worst-case missingness bounds
ResNet-18	42	0.812	0.742	0.324	[0.317, 0.331]	[0.112, 0.501]
ResNet-18	43	0.804	0.719	0.312	[0.305, 0.320]	[− 0.646, 1.130]
ResNet-18	44	0.804	0.716	0.287	[0.277, 0.296]	[0.147, 0.407]
DenseNet-121	42	0.840	0.738	0.311	[0.304, 0.318]	[0.252, 0.358]
DenseNet-121	43	0.824	0.718	0.351	[0.344, 0.357]	[0.063, 0.581]
DenseNet-121	44	0.824	0.683	0.308	[0.299, 0.317]	[− 0.146, 0.661]

Spatial controls and numerical eligibility

The spatial controls are summarized in Fig. 3. In the ResNet validation sweep, the Grad-CAM-minus-IG stability gap increased from approximately zero at 7×7 to 0.022, 0.066, 0.139, 0.255 and 0.346 at progressively finer grids. DenseNet validation sweeps showed a similar progression. Gaussian filtering substantially attenuated the gap, whereas shared shuffling before pooling retained a larger contrast. A

fixed random map had perfect stability, illustrating that stability alone cannot establish model dependence or informative spatial localization.

IG own numerical failures among 3,900 clean/perturbed evaluations per HAM seed were 115, 410 and 33 for ResNet and 4, 110 and 308 for DenseNet. Additional clean-reference propagated exclusions were 54, 417 and 68 for ResNet and 12, 121 and 40 for DenseNet. Numerical reliability therefore varied across seeds and architectures. Worst-case bounds included zero for ResNet seed 43 and DenseNet seed 44 despite positive eligible-map bootstrap intervals. Degenerate maps were tracked separately; numerical counts should not be equated with all contrast missingness.

External findings and predictive transfer

External eligible-map interactions were positive in both datasets and architectures (Table 2); predictive transfer and correctness subgroups are shown in Fig. 4. However, classification transfer was poor. Balanced accuracy fell from 0.749 and 0.734 on HAM to 0.284 and 0.326 on BCN and 0.262 and 0.236 on PAD for ResNet and DenseNet. This drop cannot be explained solely by changing class prevalence. ResNet and DenseNet predicted nv for 63 and 57 BCN images and 89 and 90 PAD images, respectively, out of 100 per dataset. All 400 fresh external predictions reproduced the saved predictions exactly; no error was identified in the checked scoring and inference components. Per-class external confusion matrices are given in Supplementary Table S3.

Table 2

Exploratory external evaluation, seed 42 only. Each row uses 100 groups. PAD has five supported truth classes.

Model	Dataset	Accuracy	Balanced accuracy	Interaction	95% bootstrap interval	Worst-case missingness bounds
ResNet-18	BCN	0.280	0.284	0.334	[0.323, 0.344]	[0.232, 0.418]
DenseNet-121	BCN	0.320	0.326	0.331	[0.319, 0.343]	[0.305, 0.352]
ResNet-18	PAD	0.260	0.262	0.294	[0.279, 0.307]	[– 0.041, 0.579]
DenseNet-121	PAD	0.240	0.236	0.293	[0.283, 0.303]	[0.267, 0.314]

Post-hoc correct-prediction subgroups contained only 24–32 images per comparison. Their eligible-map interactions were 0.302–0.323, versus 0.290–0.340 among incorrect predictions. We did not formally test a correctness-subgroup difference, and ResNet PAD's correct-subgroup missingness bounds included zero. These results do not establish independence from prediction correctness.

Faithfulness depends on replacement

Figure 5 summarizes clean insertion controls, which provided positive advantages over random ordering in many settings, but magnitudes depended on the replacement. The clearest example was DenseNet PAD IG: mean insertion advantage was 0.180 with black replacement (descriptive 95% interval [0.158, 0.204]) and − 0.017 with blurred replacement ([− 0.044, 0.012]). Grad-CAM in the same model/dataset changed from 0.258 to 0.041. Thus, favorable performance under one replacement did not establish uniformly favorable behavior under another. These controls evaluated the fixed predicted target, including incorrect targets, and did not measure clinical correctness. Across per-image exploratory stability/faithfulness associations, signs varied; we make no general monotonic or causal claim.

Descriptive lesion mask overlap

ResNet-18: Grad-CAM had mean top-decile IoU 0.298 and mean attribution mass inside the mask 0.435 across 300 eligible clean maps; IG had mean top-decile IoU 0.073 and mean attribution mass inside the mask 0.297 across 295 eligible clean maps. DenseNet-121: Grad-CAM had mean top-decile IoU 0.366 and mean attribution mass inside the mask 0.497 across 300 eligible clean maps; IG had mean top-decile IoU 0.083 and mean attribution mass inside the mask 0.315 across 299 eligible clean maps. These means involve eligible clean maps and are not adjusted method comparisons. The selected top decile and lesion area influence overlap, and a model can legitimately or spuriously use contextual pixels outside a lesion boundary. No external localization estimate was available.

Discussion

The main empirical observation is that the apparent stability advantage of Grad-CAM over positive IG changes substantially when spatial resolution is matched. Grad-CAM's input-resolution display derives from a coarse convolutional map, whereas IG retains finer input variation. Our controls are consistent with spatial aggregation influencing the comparison. They do not show that one representation is the uniquely correct evaluation scale or that the effect has a clinical mechanism.

This observation should be distinguished from established smoothing theory [3], architectural saliency priors [4] and prior evaluation critiques [5, 6]. Our candidate incremental contribution is the quantified resolution interaction under fixed-target photometric perturbations, accompanied by explicit numerical-eligibility diagnostics and external-transfer limitations. We do not claim priority for the smoothing effect itself.

Numerical missingness is consequential (Supplementary Fig. S1). Completeness and successive-budget rank checks address different aspects of integration, and clean-reference failures can propagate to otherwise accepted perturbation maps. Reporting only finite correlations creates an eligible-map estimand whose interpretation depends on the excluded cases. The wide worst-case bounds in several runs prevent an unconditional claim that the effect survives arbitrary missingness. Conversely, positive bounds in other runs show that the finding is not equally vulnerable in every configuration. Condition-level coverage and numerical-failure counts are given in Supplementary Tables S1 and S2.

External transfer failure is a substantive limitation rather than a reason to omit the external results. A stable explanation may consistently summarize an incorrect prediction. Acquisition differences, diagnostic case mix, output-category definitions and learned biases are plausible contributors to poor transfer, but this study does not identify their separate causal effects. Any future adaptation should preserve a new unevaluated test set rather than tune on these external samples and relabel them as untouched validation.

Faithfulness controls also require restraint. Black and blur replacement can yield different conclusions, and both may produce inputs unlike training images. Coarse region-ranking evaluations cannot validate pixel-level information preservation. The results support a reporting practice: declare target and layer, compare spatial scales, retain numerical failure counts, and inspect correctness and multiple faithfulness controls alongside stability.

Practical implications for comparative studies

A comparison at a single spatial scale can answer a narrower question than readers assume. Input-resolution Spearman correlation assesses the ranking of many displayed pixels, while a pooled comparison assesses ranking of regional averages. Neither measures all aspects of explanation quality. Matching a grid can expose an imbalance in spatial detail, but may also remove detail that matters for a specific application. Accordingly, we recommend presenting both scales and a sensitivity curve rather than treating the coarse result as the corrected truth or enforcing a single universal resolution.

The spatial-shuffle control reinforces this point. The same permutation applied to clean and perturbed maps preserves their native rank relation, but changes which pixels share a pooled region. A change after pooling therefore diagnoses dependence on spatial arrangement. It does not simulate a plausible lesion transformation, and it does not create a clinically meaningful explanation. Its purpose is to test an evaluation operation under a deliberately artificial null.

Shortcut learning is a plausible concern when models rely on features that transfer poorly [30], but our confusion matrices and saliency controls do not establish that a particular shortcut caused the external failures. Confidence also should not be read as calibrated reliability; calibration is a separate property [31] that we did not estimate or improve in the current audit. These distinctions prevent a stable, high-confidence explanation of a wrong prediction from being interpreted as a trustworthy diagnosis.

The reporting recommendation is therefore procedural. Comparative studies should disclose output target, reference, sign handling and attribution layer, show the spatial representations being compared, and retain numerical exclusions. They should report whether perturbations changed predictions and test masking conclusions under more than one replacement. If clinical utility is the goal, a separate user or outcome study is required. Our experiments address the computational comparison and its limits, not that downstream question.

Limitations

The original audit concerned one dataset and two saliency methods, with positive-only IG and a black reference. Other methods, signed attributions, alternative targets and geometric transformations were not comprehensively evaluated. External samples were small, approximately class-balanced and single-seed, with conservative overlap exclusions that restrict representativeness. Patient identifiers were unavailable for HAM and BCN; residual duplicates and incomplete source identifiers remain possible. Synthetic perturbations were not clinically adjudicated. External localization could not be assessed without verified masks. Multiple follow-ups were exploratory and unadjusted for multiplicity. Existing source metadata were checked for consistency but not independently clinically reannotated. Finally, the work does not establish diagnostic usefulness, clinician benefit or superiority of a new method.

Conclusions

Matching spatial resolution changed the apparent stability advantage of Grad-CAM over positive Integrated Gradients in HAM10000-trained ResNet-18 and DenseNet-121 classifiers. After both maps were represented on a common 7×7 grid, the Grad-CAM-minus-IG stability contrast decreased by 0.287–0.351 across six HAM10000 runs (two architectures, three training seeds) and by 0.293–0.334 in four exploratory external summaries from BCN20000 and PAD-UFES-20. Validation controls showed that the gap shrank progressively as grids became coarser, was largely attenuated by Gaussian filtering and direct pooling, and was partly retained when a shared pixel permutation preceded pooling. These patterns are consistent with spatial aggregation influencing the comparison. They do not identify a mechanism, and they do not show that either the native or the pooled scale is the correct basis for evaluating explanations.

The finding is conditional on several limits that we regard as part of the result. Integrated Gradients maps failed numerical-eligibility checks in a run-dependent number of evaluations, and worst-case missingness bounds included zero in three of the ten architecture-by-dataset-by-seed summaries. Diagnostic transfer to the external samples was poor, with accuracy of 24–32%, so stable explanations were often attached to incorrect predictions. Faithfulness conclusions depended on the replacement used for masking: the insertion advantage of Integrated Gradients for DenseNet-121 on PAD-UFES-20 was 0.180 with black replacement and − 0.017 with blurred replacement. Lesion-mask overlap, available only for HAM10000, was descriptive and favoured Grad-CAM over Integrated Gradients, but lesion masks describe lesion extent rather than diagnostic evidence. The external samples were small, approximately class-balanced and evaluated with a single seed, and the follow-up analyses were exploratory and unadjusted for multiplicity.

These results support a procedural recommendation rather than a verdict on either method. Studies that compare saliency stability should declare the output target, attribution layer, sign handling and baseline; report native and matched-resolution comparisons; retain counts of numerical failures and propagated exclusions; and examine prediction correctness, localization and more than one faithfulness control alongside stability. Stability should not be treated as a stand-alone indicator of explanation reliability, and none of these analyses establishes diagnostic usefulness or clinician benefit. Future work should test whether the interaction generalizes to additional attribution methods, signed attributions, geometric

perturbations and non-medical datasets, use independent external cohorts with verified lesion masks, and preserve a genuinely untouched test set for any model adaptation.

Data availability

The datasets analysed in this study are publicly available: HAM10000 (Harvard Dataverse, https://doi.org/10.7910/DVN/DBW86T) [18], BCN20000 [20] and PAD-UFES-20 [21]; see the reference list for access details and terms of use. The code, fixed sample identifiers, checksums and per-image derived results generated for this study are available from the corresponding author on reasonable request. Redistribution of images and masks is subject to the dataset-specific licences.

Declarations

Data availability

The datasets analysed in this study are publicly available: HAM10000 (Harvard Dataverse, https://doi.org/10.7910/DVN/DBW86T) [18], BCN20000 [20] and PAD-UFES-20 [21]; see the reference list for access details and terms of use. The code, fixed sample identifiers, checksums and per-image derived results generated for this study are available from the corresponding author on reasonable request. Redistribution of images and masks is subject to the dataset-specific licences.

Acknowledgements

The authors thank the creators of the public datasets and open-source software cited in this manuscript.

Author contributions

B.B.: conceptualization, methodology, investigation, formal analysis, visualization, writing—original draft, and writing—review and editing. D.G.: methodology, validation, interpretation of findings, and writing—review and editing. R.N.T.: supervision, conceptual and methodological guidance, and writing—review and editing. All authors reviewed and approved the final manuscript.

Funding

This research received no specific funding.

Competing interests

The authors declare no competing interests.

Ethics statement

This study used only previously published, publicly available image collections. No new participants were recruited, no clinical intervention was performed and no identifiable data were collected.

Use of AI tools

An AI assistant supported code development, computational execution, analysis, literature retrieval and drafting. The authors verified the analyses, citations and text, take full responsibility for the content, and did not list the AI tool as an author.

Additional information

Correspondence and requests for materials should be addressed to D.G. (dipesh.gurung@lbef.edu.np).

References

1. Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of ICCV (2017). https://arxiv.org/abs/1610.02391
2. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. Proceedings of ICML, PMLR 70, 3319–3328 (2017). https://proceedings.mlr.press/v70/sundararajan17a.html
3. Ye, Z. & Farnia, F. Gaussian smoothing in saliency maps: The stability-fidelity trade-off in neural network interpretability. Proceedings of AISTATS, PMLR 258, 2125–2133 (2025). https://proceedings.mlr.press/v258/ye25a.html
4. Kuang, K., Carmichael, I. & Purdom, E. When random saliency looks trained: Architectural center bias in CNN interpretability. Proceedings of ICML, PMLR 306, 61236–61266 (2026). https://proceedings.mlr.press/v306/kuang26a.html
5. Nieradzik, L., Stephani, H. & Keuper, J. Reliable evaluation of attribution maps in CNNs: A perturbation-based approach. *Int. J. Comput. Vision*. **133**, 2392–2409. https://doi.org/10.1007/s11263-024-02282-6 (2025).
6. Venkatesh, K., Mutasa, S., Moore, F., Sulam, J. & Yi, P. H. Gradient-based saliency maps are not trustworthy visual explanations of automated AI musculoskeletal diagnoses. *J. Imaging Inf. Med.* **37**, 2490–2499. https://doi.org/10.1007/s10278-024-01136-4 (2024).
7. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608 (2017). https://arxiv.org/abs/1702.08608
8. Adebayo, J. et al. Sanity checks for saliency maps. *Adv. Neural. Inf. Process. Syst.* **31** (2018). https://arxiv.org/abs/1810.03292
9. Kindermans, P. J. et al. The (un)reliability of saliency methods. arXiv:1711.00867 (2017). https://arxiv.org/abs/1711.00867
10. Alvarez-Melis, D. & Jaakkola, T. S. On the robustness of interpretability methods. arXiv:1806.08049 (2018). https://arxiv.org/abs/1806.08049
11. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: Removing noise by adding noise. arXiv:1706.03825 (2017). https://arxiv.org/abs/1706.03825
12. Fong, R. & Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. Proceedings of ICCV (2017). https://arxiv.org/abs/1704.03296

13. Petsiuk, V., Das, A., Saenko, K. & RISE Randomized input sampling for explanation of black-box models. Proceedings of BMVC (2018). https://arxiv.org/abs/1806.07421
14. Hooker, S., Erhan, D., Kindermans, P. J. & Kim, B. A benchmark for interpretability methods in deep neural networks. Advances in Neural Information Processing Systems 32 (2019). https://proceedings.neurips.cc/paper_files/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html
15. Yeh, C. K., Hsieh, C. Y., Suggala, A. S., Inouye, D. I. & Ravikumar, P. On the (in)fidelity and sensitivity for explanations. arXiv:1901.09392 (2019). https://arxiv.org/abs/1901.09392
16. Hedström, A. et al. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *J. Mach. Learn. Res.* **24** (34), 1–11 (2023). https://jmlr.org/papers/v24/22-0142.html
17. Saarela, M. & Georgieva, L. Robustness, stability, and fidelity of explanations for a deep skin cancer classification model. *Appl. Sci.* **12**, 9545. https://doi.org/10.3390/app12199545 (2022).
18. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data*. **5**, 180161. https://doi.org/10.1038/sdata.2018.161 (2018).
19. Abhishek, K., Jain, A. & Hamarneh, G. Investigating the quality of DermaMNIST and Fitzpatrick17k dermatological image datasets. *Sci. Data*. **12**, 196. https://doi.org/10.1038/s41597-025-04382-5 (2025).
20. Hernández-Pérez, C. et al. BCN20000: Dermoscopic lesions in the wild. *Sci. Data*. **11**, 641. https://doi.org/10.1038/s41597-024-03387-w (2024).
21. Pacheco, A. G. C. et al. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data Brief.* **32**, 106221 https://doi.org/10.1016/j.dib.2020.106221 (2020).
22. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. Proceedings of CVPR (2016). https://arxiv.org/abs/1512.03385
23. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. Proceedings of CVPR (2017). https://arxiv.org/abs/1608.06993
24. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision*. **115**, 211–252 (2015). https://arxiv.org/abs/1409.0575
25. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Proceedings of ICLR (2019). https://arxiv.org/abs/1711.05101
26. Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural. Inf. Process. Syst.* **32** (2019). https://arxiv.org/abs/1912.01703
27. Kokhlikyan, N. et al. Captum: A unified and generic model interpretability library for PyTorch. arXiv:2009.07896 (2020). https://arxiv.org/abs/2009.07896
28. Efron, B. Bootstrap methods: Another look at the jackknife. *Annals Stat.* **7**, 1–26. https://doi.org/10.1214/aos/1176344552 (1979).

29. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011). https://jmlr.org/papers/v12/pedregosa11a.html

30. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673. https://doi.org/10.1038/s42256-020-00257-z (2020).

31. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. Proceedings of ICML, PMLR 70, 1321–1330 (2017). https://proceedings.mlr.press/v70/guo17a.html

Figures

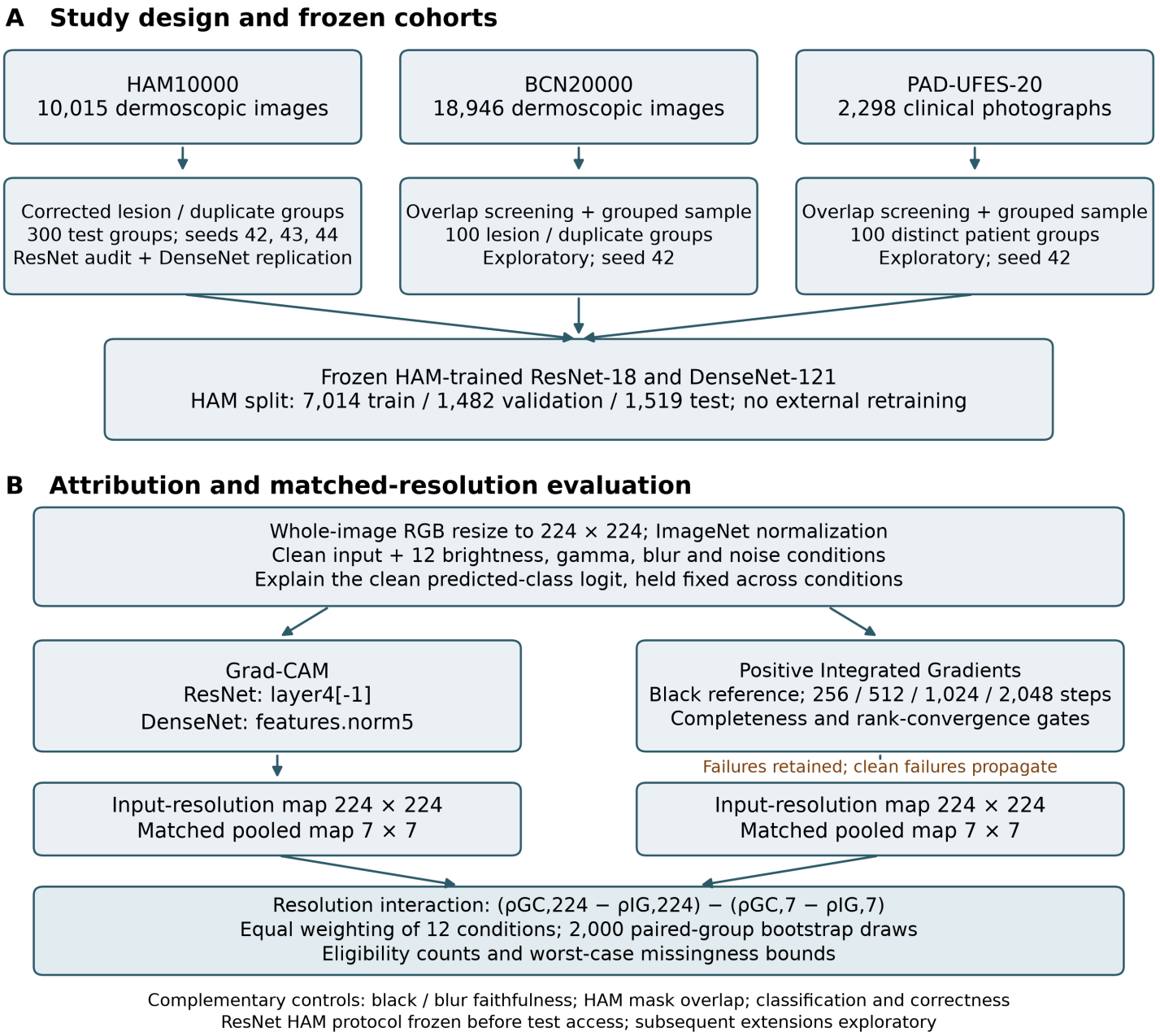

Figure 1

Experimental pipeline and resolution comparison. Panel A shows the dataset inventories, grouped explanation cohorts and frozen HAM-trained classifiers. External inference uses seed 42 without retraining. Panel B shows fixed-target attribution under 12 photometric perturbations, IG numerical checks, and native versus pooled stability comparisons. Unresolved numerical cases are counted and propagated exclusions remain separate. Here, ρ denotes clean-to-perturbed Spearman correlation. Complementary controls assess clean faithfulness, HAM mask overlap and prediction correctness. Patient independence is not established for HAM or BCN; external similarity flags were conservatively quarantined.

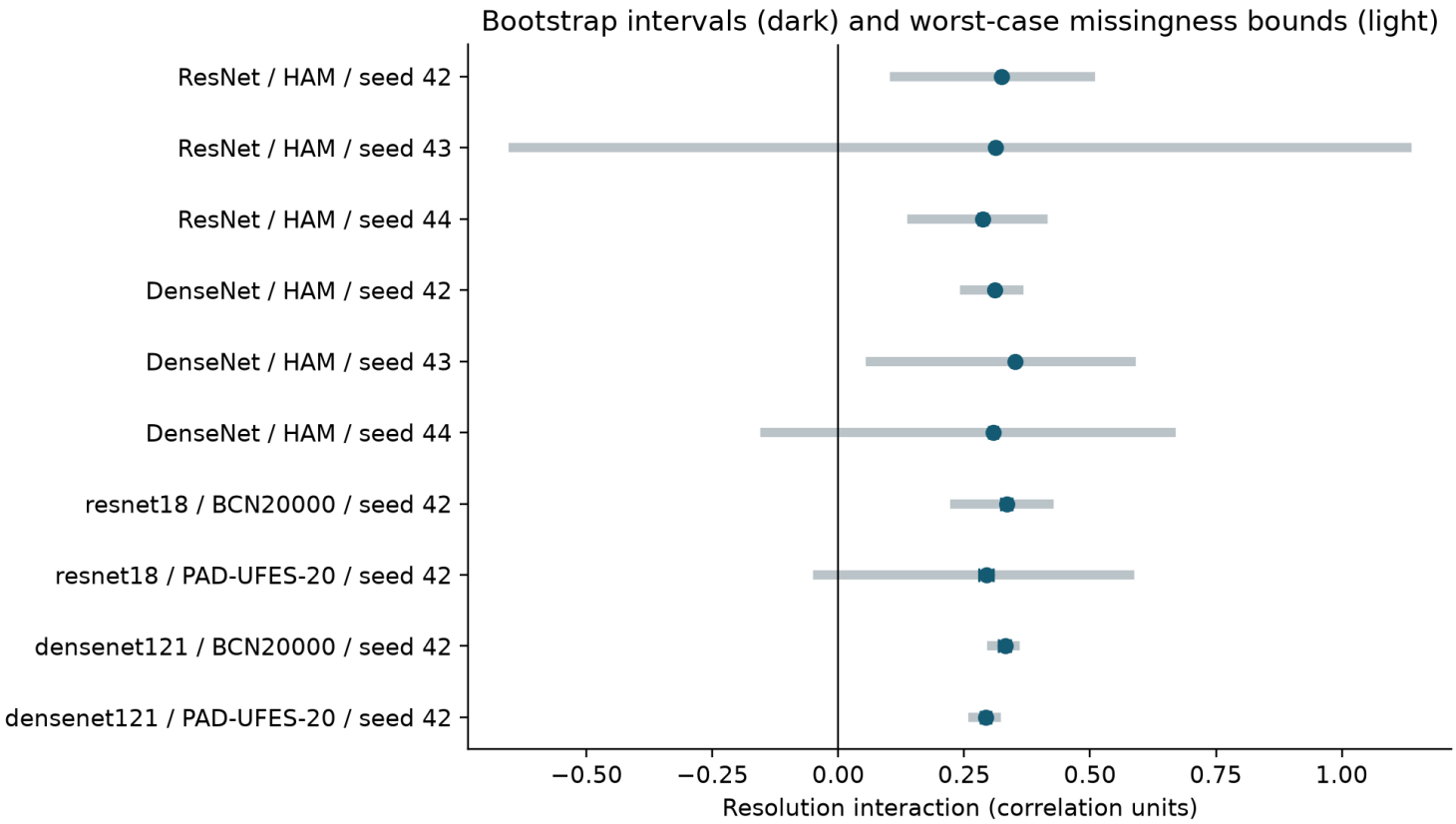

Figure 2

Resolution interaction and numerical missingness. Positive values indicate that pooling reduces the Grad-CAM-minus-IG stability contrast. Dark intervals are paired-group percentile bootstrap intervals conditional on observed eligibility; light ranges assign missing contrasts their theoretical extremes and are not confidence intervals. HAM includes 300 corrected groups per seed. External analyses include 100 groups per dataset and architecture, seed 42 only. Overlapping samples and model seeds are not independent studies for meta-analysis.

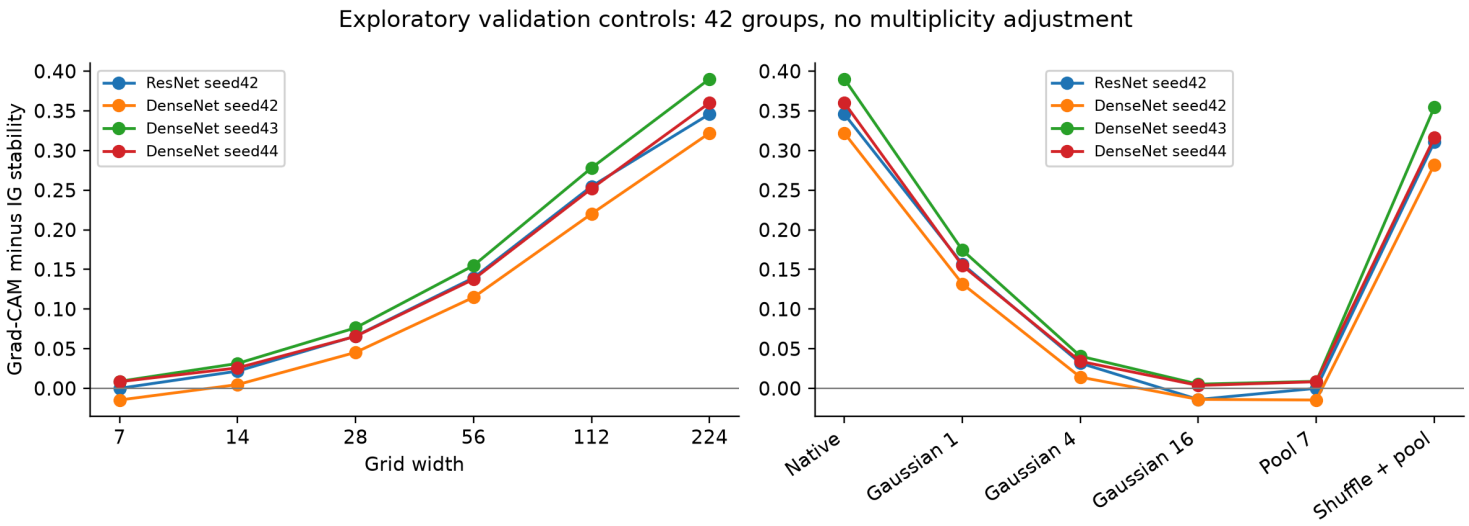

Figure 3

Resolution and spatial controls on validation groups. The left panel shows the mean method stability gap at progressively finer grids. The right panel compares native maps, Gaussian map filters, direct pooling and shared-shuffle pooling. Each sweep uses the same 42 validation groups. Points are descriptive estimates; these exploratory follow-ups are not multiplicity-adjusted significance tests. Gaussian map filtering differs from SmoothGrad input-noise averaging. Shared shuffling preserves native ranks while disrupting spatial arrangement before pooling.

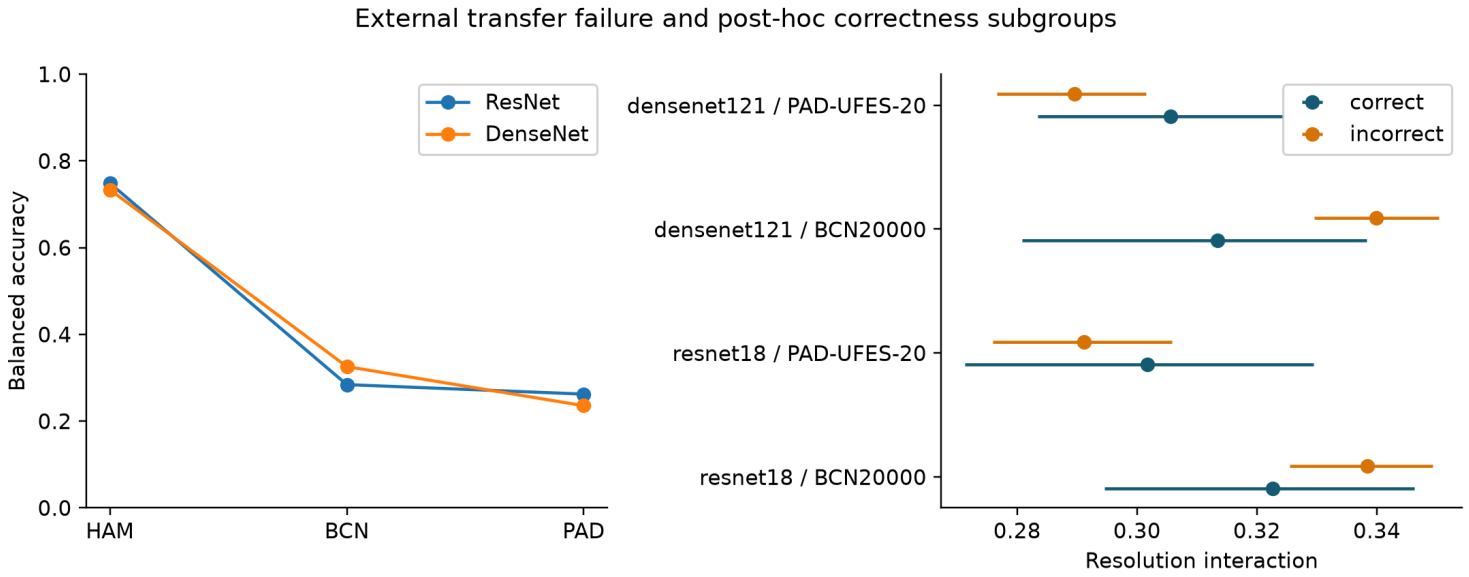

Figure 4

External prediction failure and correctness subgroups. The left panel compares balanced accuracy on HAM and the approximately class-balanced external samples. PAD includes five supported truth classes but all seven prediction outputs remain in scoring. The right panel shows post-hoc resolution interactions for correctly and incorrectly classified external images with eligible-map bootstrap intervals. Correct

subgroups contain only 24–32 images. No formal between-subgroup difference test was performed; worst-case missingness ranges are reported separately.

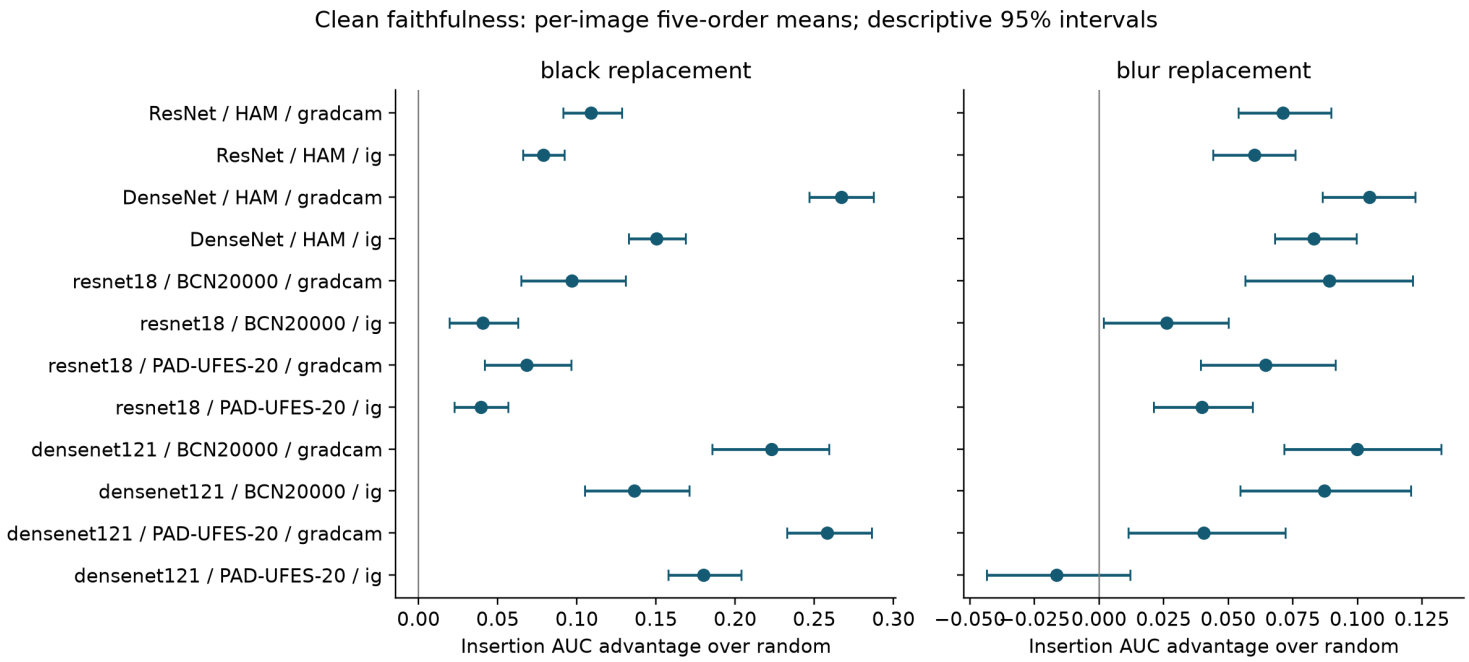

Figure 5

Clean insertion controls under two replacements. Each point summarizes the attribution-order insertion AUC minus random-order AUC, averaging five random orders within image and bootstrapping groups 2,000 times. Positive values indicate advantage under this benchmark, not clinical validity. The figure displays seed 42; saved summaries contain all HAM seeds. IG uses eligible clean maps, so denominators differ. Black and blur replacement can induce distribution shift, and the same 7×7 ranking means native/pool7 agreement is a consistency check rather than pixel-fidelity evidence.

Supplementary Files

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

- 06SupplementaryInformation.docx
- SupplementaryFigureS1.png