Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 23 May 2026
Revised 25 June 2026
Accepted 18 July 2026

Natural Resources for Human Health 2026, Vol. 6 (8s): 1121–1147

KEYWORDS:
Explainable AI; dermoscopic image classification; Grad-CAM++; explanation reliability; algorithmic fairness; uncertainty quantification; skin cancer diagnosis; clinical decision support

TrustXAI-Derm: An Empirical Audit of Explanation Reliability, Failure Prediction, and Deferral in Dermoscopic AI

Bhavesh Atulbhai Vaghela[1]*, Dr. Priya R Swaminarayan[2], Bharat Tank[3]

[1]Department of Artificial intelligence and Data science, Parul Institute of Engineering & Technology, Faculty of Engineering & Technology, Parul University, Vadodara, Gujarat 391760, India
[2]Department Faculty of IT & CS, Parul Institute of Computer Applications, Parul Institute of Computer Application, Parul University, Vadodara, Gujarat 391760, India
[3]Department of Computer Science & Engineering, Faculty of Engineering & Technology, Parul Institute of Engineering & Technology (PIET), Parul University, Vadodara, Gujarat 391760, India
Corresponding Author: bhavesh.vaghela34353@paruluniversity.ac.in

ABSTRACT:

Background: When applied to the skin, the typical result provided by dermoscopic AI models is aggregate accuracy, without considerations of accuracy reliability or the reliability of visual explanations across diagnostic classes, model architectures, lesion sizes, or estimated skin tones.

Objective: We present an empirical audit (not a new classifier) of the reliability of dermoscopic explanation on these dimensions.

Methods: Given solely the mask overlap as explanation ground truth, we define a mask-overlap Trustworthy Index for Explainable AI (TIxAI) and test it per diagnostic class (GAP-1), train a meta-model to predict an explanation failure based on embeddings and Monte Carlo Dropout (MC-Dropout) uncertainty prior to computing any explanation (GAP-2), audit explanation quality and uncertainty across an Individual Typology Angle (ITA) skin-tone proxy (GAP-3), and prototype a Composite Risk Score for clinician deferral (GAP-4).

Results: Reliability of explanations, when tested on a split of the patients (n = 990) was significantly different across the diagnostic classes (Kruskal–Wallis H = 30.65, p = 2.95×10^{-5}) as well as with lesion size (H = 269.5, p = 3.0×10^{-59}) with the latter class showing the largest effect. Compared to the two, ConvNeXt-Tiny achieved better balanced accuracy (0.763 vs. 0.689), but significantly lower calibration (0.308 vs. 0.074) (Expected Calibration Error) ECE. The best result was the failure-prediction meta-model (5-fold cross-validated AUC-ROC 0.832, 95% BCa CI 0.793–0.866). The difference in MC-Dropout entropy was significant across ITA buckets (p = 0.0004), but not for TIxAI (p = 0.20), so we do not consider this an evidence of fairness in the latter case, but rather an underpowered null. The following candidate trust signals were null results: Grad-CAM++/SHAP disagreement, mask-free saliency geometry, and contrastive analysis. Unresolved issues are highlighted instead of being hidden, such as reported pre-convergence (epoch 60/150), or reported near-zero median TIxAI (0.0006) as a possible artifact of a Grad-CAM++ target-layer configuration, but not as a genuine result, and the

single-seed result of DenseNet121's raw-versus-balanced accuracy inversion is not diagnosed. A prototype deferral score focused residual errors in the deferred subset, and is based on a single split and needs to be validated using a bootstrap method. External validation and a dermatologist reader study are not yet performed.

Conclusion: We thus offer the rigorous empirical study as a self-audited work where the contribution is the audit approach and its positive, null and unresolved results, and not a clinical tool that can be deployed. Clinical/Industrial Impact: TrustXAI-Derm proposes an auditing path towards explanations that are presented to clinicians for decisions, highlighting those explanations that are low-trust before the system has arrived, and proposes a prototype of such an explanation that concentrates the residual error in the explanations withheld upon deployment.

1. INTRODUCTION

1.1 Background and clinical importance

Skin cancer is among the most common malignancies worldwide, and melanoma, although relatively infrequent, accounts for a disproportionate share of skin-cancer mortality. Prognosis is strongly stage-dependent: early detection and excision yield survival rates approaching 99%, whereas late-stage diagnosis carries substantially worse outcomes. Well-documented diagnostic delays are linked to skin tone, as dermoscopic and clinical training material disproportionately represents lighter-skinned populations [4] and both human dermatologists and AI systems perform worse on darker skin [15,17]. As deep learning-based dermoscopic decision-support tools move toward clinical deployment [48], any bias inherited from their training data risks exacerbating existing disparities rather than narrowing them, echoing findings of algorithmic racial bias in other clinical risk-prediction settings [49]; robust communication of predictive uncertainty to the clinician has been proposed as one safeguard against this risk [50].

1.2 AI in Dermoscopic Diagnosis

Convolutional neural networks and, more recently, transformer-based architectures have achieved high aggregate accuracy on public dermoscopy benchmarks such as HAM10000, in some cases matching or exceeding board-certified dermatologists on narrow classification tasks. Explainable AI (XAI) methods — most commonly Grad-CAM and its refinements — are typically layered on top of these classifiers to produce a visual heatmap intended to show where the model focused. In the overwhelming majority of published work, this heatmap is presented qualitatively, as an illustrative figure suggesting that the model "looks reasonable," rather than as a quantitatively validated signal. Clinical adoption, however, depends on more than raw accuracy: it depends on whether a clinician can trust which cases the model gets right, why it reached a given decision, and when its explanation itself should not be trusted.

1.3 Research Motivation

Accuracy alone is an insufficient basis for clinical trust. A model can be accurate on aggregate while its explanations are systematically unreliable for particular lesion types, particular patients, or particular images — and a clinician has no way of knowing, from the heatmap alone, whether a given explanation belongs to the reliable or unreliable regime. Trustworthy deployment of dermoscopic AI therefore requires, at minimum: (i) a quantitative measure of explanation reliability, not just a qualitative overlay; (ii) fairness auditing that considers whether explanations — not only predictions — are equally reliable across patient subgroups such as skin tone; (iii) uncertainty estimation explicitly connected to explanation trust rather than treated as a separate axis; and (iv) a defensible mechanism for the system to signal, and act on, its own untrustworthiness rather than presenting every case with the same apparent confidence.

1.4 Research Gap

A review of the current literature on explainable and fair dermoscopic AI reveals a consistent pattern of omissions, which this work organizes into four concrete, addressable gaps:

• **GAP-1 — Per-Class Explanation Reliability Analysis.** Existing dermoscopic XAI studies rarely report explanation quality broken down by diagnostic class. A single aggregate "the heatmaps look reasonable" statement can conceal large differences between common classes (well represented in training data) and rare classes (poorly represented) — precisely where clinical decision support is most needed.

• **GAP-2 — Explanation Failure Prediction.** No widely used pipeline predicts, before an explanation is generated and inspected, whether that explanation is likely to be unreliable. Explanation quality is treated as a fixed, always-available signal rather than something that can itself be forecast from the model's internal state.

• **GAP-3 — Skin-Tone-Aware Explanation Fairness and Uncertainty Audit.** Where fairness is discussed in dermoscopic AI at all, it is almost always prediction fairness (accuracy, sensitivity, AUC by subgroup), broken down by Fitzpatrick skin type — not explanation fairness, i.e., whether the model's reasoning, or its uncertainty, is equally trustworthy across subgroups.

• **GAP-4 — Composite Clinical Risk Score and Intelligent Deferral.** Existing systems generally do not combine explanation reliability, prediction uncertainty, and class-specific clinical risk into a single, actionable signal for deferring a case to a human clinician; deferral, where it exists at all, is usually confidence-threshold-only and ignores explanation quality entirely.

An initial, more architecturally ambitious proposal — multimodal fusion of dermoscopic images with Fitzpatrick metadata via cross-attention, a novel mask-overlap explanation metric, and ONNX edge deployment — was tested against the recent literature and found to be largely non-novel at the component level, since pretrained CNN backbones, Grad-CAM++, MC-Dropout, mask-overlap metrics, and Fitzpatrick-stratified accuracy audits have each been published independently. What survives, and what this project implements, is the narrower question posed by GAP-1, GAP-2, GAP-3, and GAP-4: can an AI system's explanations themselves be shown to be unreliable in ways that accuracy-only or single-metric fairness audits miss, and can specific, testable signals of that unreliability be identified and acted upon?

1.5 Research Objectives

This study aims to: (1) quantify explanation reliability on a per-class basis using a ground-truth-mask-based trustworthiness index (GAP-1); (2) determine whether explanation failure can be predicted in advance from model embeddings and uncertainty features, without first computing the explanation (GAP-2); (3) audit whether explanation quality and predictive uncertainty differ across an image-derived skin-tone proxy, reporting both as distinct signals rather than a single composite score (GAP-3); and (4) construct and evaluate a composite risk score that supports a clinically actionable deferral decision (GAP-4).

RQ1: Does explanation reliability (TIxAI) vary systematically across diagnostic classes?

RQ2: Does lesion size moderate explanation reliability independently of class?

RQ3: Can explanation failure be predicted before the explanation itself is computed?

RQ4: Does selecting a backbone by predictive accuracy alone guarantee reliable explanations?

RQ5: Do explanation localization quality and predictive uncertainty differ across an estimated skin-tone proxy?

RQ6: Can a label-leakage-free composite clinical risk score support a defensible deferral decision?

1.6 Main Contributions

This work is an empirical audit study, not a new classification architecture. Its contributions are:

• **An integrated audit methodology** for dermoscopic AI that combines mask-based explanation-reliability scoring (TIxAI), pre-hoc explanation-failure prediction, a skin-tone-stratified uncertainty audit, and a deferral prototype — assembled from established components (CNN backbones, Grad-CAM++, MC-Dropout, ITA) whose novelty is claimed only at the level of their joint empirical use, not at the component level.

• **An empirical finding that explanation reliability is class- and lesion-size-dependent** (GAP-1), quantified with TIxAI and non-parametric tests (Kruskal–Wallis omnibus, Holm-corrected pairwise) rather than by qualitative "the heatmaps look reasonable" inspection; the lesion-size effect is the single largest observed.

• **A validated explanation-failure meta-model** (GAP-2) that forecasts unreliable explanations from image embeddings and MC-Dropout uncertainty — before the explanation is generated (5-fold CV, AUC-ROC 0.832, 95% BCa CI 0.793–0.866) — the paper's most statistically robust result.

• **Null and underpowered results reported as first-class findings**: TIxAI does not differ across the ITA skin-tone proxy ($p = 0.20$, $n \approx 20$ in the darkest bucket — an exploratory sanity check, not a fairness claim), and Grad-CAM++/SHAP disagreement, saliency geometry, and contrastive signals do not predict correctness.

• **A deferral prototype (CCRS)** (GAP-4) presented as preliminary: it concentrates residual error in the deferred subset on a single test split, and we explicitly withhold any coverage–accuracy claim pending bootstrap or cross-validated characterization.

• **A transparent audit of the pipeline's own defects** — a pre-convergence checkpoint, single-seed statistics, an undiagnosed raw-versus-balanced accuracy inversion, and ConvNeXt-Tiny's near-zero TIxAI, which we treat as a probable configuration artifact rather than a finding — positioning honest error accounting as part of the scientific contribution rather than as a disclaimer appended to a closed result set.

2. RELATED WORK

2.1 Dermoscopic Skin Lesion Classification

Convolutional architectures — from earlier ResNet [27] and DenseNet [5] variants through more recent EfficientNet [29] and ConvNeXt [6] families — have been repeatedly benchmarked on HAM10000 and related dermoscopy datasets, consistent with the broader trajectory of deep learning in medical image analysis [47], typically reporting overall accuracy or AUC as the primary outcome, in some cases matching or exceeding dermatologist-level performance on narrow, curated classification tasks [16,46]. Vision Transformer (ViT) variants [30], built on the self-attention mechanism [31], have more recently been applied to the same task. Across this literature, the classification backbone itself is treated as the primary object of study, with explainability, where present, added as a secondary visualization step rather than a jointly evaluated component.

2.2 Explainable AI

Grad-CAM and its refinement, Grad-CAM++ [8,9], remain the dominant visualization tools in medical imaging generally and dermoscopy specifically, owing to their applicability to any convolutional architecture without modification. SHAP-based attribution methods [10] (including gradient-based approximations such as Integrated Gradients [33]) and LIME [32] appear less frequently as alternative or complementary attribution methods. In nearly all cases, these methods are validated qualitatively — a handful of example heatmaps judged by visual inspection — rather than quantitatively scored against ground-truth lesion boundaries at the dataset level, despite calls for more rigorous, falsifiable evaluation protocols [34,35] and broader taxonomies of what a "trustworthy" explanation should satisfy [36]. More recent work has begun to benchmark multiple explainability methods quantitatively against one another for skin-lesion classifiers specifically [18], and to pair attribution maps with auxiliary trust and transparency signals within a single diagnostic pipeline [19,20]; consistent with the pattern above, none of these jointly evaluate explanation quality, predictive uncertainty, and patient subgroup together. Where a quantitative mask-overlap metric closely related to the Energy-Based Pointing Game has been proposed in the broader literature, it has typically been reported as a single aggregate score, not stratified by class or patient subgroup, and not linked to model uncertainty. The Trustworthiness Index adopted here was introduced by Ieracitano et al. [26] as an aggregate index; this work uses it, unmodified, in a stratified diagnostic role — comparing classes, architectures, lesion sizes, and skin-tone groups — and claims no novelty for the metric itself. The broader goal of predicting, rather than only measuring after the fact, whether an explanation can be trusted parallels a similar post-hoc reliability classifier proposed for image segmentation [52], though that work targets segmentation saliency maps whereas GAP-2 here targets classification explanation failure directly. Statistically richer alternatives to simple mask overlap have also been proposed, including Accuracy and Softmax Information Curves [53], which the present mask-overlap-based LSAS/TIxAI evaluation does not incorporate; adopting such curves as a complementary, more statistically grounded evaluation is noted as future work (Section 8) rather than attempted here.

2.3 Fairness in Medical AI

A growing body of work has documented that dermoscopic and clinical skin-image datasets are heavily skewed toward lighter Fitzpatrick skin types [4], and that both human dermatologists and AI classifiers show measurably worse diagnostic performance on darker skin [15,17]. This skew, and the difficulty of even defining and annotating skin tone consistently, has itself become

an object of study [21,22], and separate reviews have specifically catalogued the resulting diagnostic performance gap across skin tones in dermatology-AI systems [23]. Fairness evaluation in dermoscopic AI has, in the surveyed literature, concentrated almost exclusively on prediction fairness — accuracy, sensitivity, and calibration error broken down by Fitzpatrick skin type. This project's central related-work observation is that explanation-quality fairness — whether a model's saliency maps are equally reliable across skin tones, independent of whether its final prediction is correct — is a materially different and much less studied question, and is the primary lens applied in Section 5.4 (GAP-3).

2.4 Uncertainty Estimation

Monte Carlo Dropout [11] and temperature scaling [12] are established, general-purpose tools for epistemic uncertainty estimation and post-hoc probability calibration, respectively, and are used here in their standard form rather than as novel contributions (see [25,37] for recent surveys of calibration and uncertainty-quantification methods in deep learning generally; deep ensembles [38] and selective/reject-option classification [39] represent related but distinct approaches not adopted here). Their use in dermoscopic AI has generally been confined to reporting calibration error or flagging low-confidence predictions, rather than relating uncertainty to the reliability of the accompanying visual explanation, or to a subgroup such as skin tone. What this work adds is not a new uncertainty method but a systematic examination of how uncertainty interacts with explanation trust and with skin tone. The GAP-4 composite risk score and deferral mechanism (Section 3.7) is likewise conceptually related to, though independently developed from, recent uncertainty-gated deferral systems proposed for clinical AI more broadly [24]. It also sits within the broader "Learning to Defer" / complementarity-driven deferral literature, most notably CoDoC [51], which learns when to defer a locked, non-retrainable predictive model to a clinician or clinical workflow; unlike CoDoC, which is trained and validated as a general-purpose deferral layer across breast-cancer and tuberculosis screening tasks, the CCRS proposed here is a composite of six dermoscopy-specific trust signals (uncertainty, explanation unreliability, XDI, class-confusion severity, malignancy weighting, and conformal-set size) and has not been benchmarked against CoDoC or similar frameworks; that comparison is identified as a specific direction for future work (Section 8) rather than claimed here.

2.5 Research Gap Summary

Table 1. Comparison of representative prior-work categories against the four research gaps addressed in this work. "Explainability" denotes quantitative (not merely qualitative) explanation evaluation; "Explanation Fairness" denotes explanation-level (not only prediction-level) fairness auditing; "Trust Signal" denotes an integrated uncertainty–explanation signal; "Deferral" denotes an explicit, explanation-aware human-referral mechanism.

Prior-work category	Classification reported	Explainability (quantitative)	Prediction fairness	Explanation fairness	Uncertainty linked to trust	Deferral
HAM10000/ISIC classifiers (CNN/ViT)	Yes	No (qualitative CAM only)	Rarely	No	No	No
Mask-overlap XAI metrics (e.g., TIxAI-family)	Partial	Yes (aggregate only)	No	No	No	No
Fitzpatrick-stratified fairness audits	Yes	No	Yes	No	No	No
MC-Dropout / calibration studies	Yes	No	No	No	Partial	Threshold-only, rarely present
Learning-to-Defer / complementarity-driven deferral (e.g., CoDoC [51])	No	No	No	No	Partial	Yes (general-purpose, locked-model; not explanation-aware)
TrustXAI-Derm (this work)	Yes	Yes (per-class, GAP-1)	Not separately re-computed	Yes (GAP-3)	Yes (GAP-2, GAP-3)	Yes (GAP-4)

Natural Resources for Human Health
eISSN: 2583-1194

The recurring pattern across the surveyed categories of prior work is that no single paper jointly reports classification, explanation-quality, fairness, and uncertainty analyses, and none stratifies explanation quality itself by skin tone. This is the specific, narrow gap this project addresses, and the motivation for structuring the proposed framework around GAP-1, GAP-2, GAP-3, and GAP-4.

3. PROPOSED TRUSTXAI-DERM FRAMEWORK

3.1 Overall Architecture

TrustXAI-Derm is deliberately structured as an audit layer on top of a conventional classifier rather than as a new classification architecture. A dermoscopic image is first classified by DenseNet121 (the primary model, selected for its spatially coherent Grad-CAM++ activations); its penultimate-layer embedding, predicted class, and MC-Dropout uncertainty estimates then feed four trust-audit modules — per-class explanation scoring (GAP-1), failure prediction (GAP-2), skin-tone-stratified auditing (GAP-3), and composite risk scoring with deferral (GAP-4) — whose outputs are combined into a single clinician-facing recommendation: display the explanation, or defer the case to a human reviewer. Two supporting modules, a Multi-XAI Disagreement Index (Grad-CAM++ vs. SHAP) and mask-free saliency-geometry descriptors, were additionally implemented as candidate trust signals and are reported in Section 5.6 as part of the framework's honest-null-result accounting, though they are not among the four primary contributions.

Figure 1 illustrates the severe class imbalance in the underlying HAM10000 training data that motivates both the combined loss design (Section 4.3) and the class-stratified explanation-reliability analysis central to GAP-1.

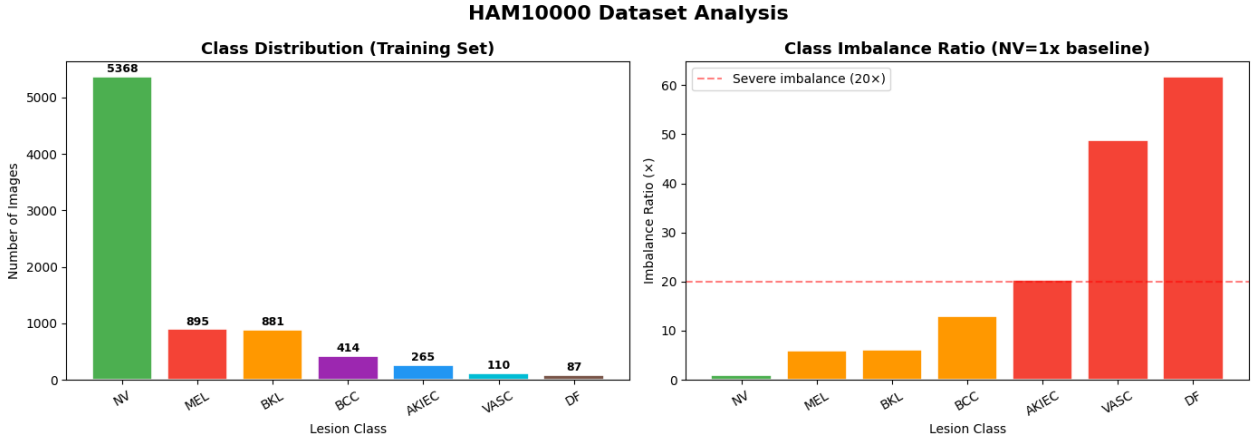

Figure 1. HAM10000 training-set class distribution and per-class imbalance ratio relative to the majority class (NV).

Figure 1. HAM10000 class distribution (training set, n = 8,020) and per-class imbalance ratio relative to the majority class (nevus, NV). Three of the seven classes (AKIEC, VASC, DF) exceed a 20× imbalance ratio, with dermatofibroma (DF) exhibiting a greater than 55× imbalance relative to NV. This imbalance directly motivates the combined focal/label-smoothed loss described in Section 4.3, and is the underlying reason the rarest classes are also the ones for which explanation reliability is least assured a priori (Section 5.2).

3.2 Baseline Classification Model

DenseNet121 (approximately 7.48M trainable parameters in its adapted classification head) is used as the primary model because its final dense block preserves the spatially coherent activations that Grad-CAM++ requires for a meaningful localization signal. Its classification head consists of global average pooling, dropout, a 512-unit fully connected layer with batch normalization [28] and ReLU activation, a second dropout layer, and a final seven-way linear classifier, with Monte Carlo Dropout enabled at inference for uncertainty estimation. ConvNeXt-Tiny is trained independently and used strictly as an **accuracy baseline** — it is not the model on which the explainability audit is primarily performed — allowing a direct test of whether the more accurate architecture is also the more explainable one (Section 5.6).

3.3 Trust-Aware Explanation Generation

Grad-CAM++ is computed via forward/backward hooks placed on DenseNet121's final dense-block convolution (and, for the cross-architecture comparison, on ConvNeXt-Tiny's final depthwise convolution), producing a class-discriminative saliency map for each prediction. To quantify, rather than merely display, the reliability of this map, we adopt the **Trustworthiness Index for Explainable AI (TIxAI)** of Ieracitano et al. [26] — the fraction of total Grad-CAM++ saliency mass that falls inside the binarized ISIC 2018 ground-truth lesion segmentation mask for that image (stated formally as Equation (1) in Section 3.9). We claim no novelty for the metric itself; our contribution is its stratified, diagnostic use (per class, per lesion size, per architecture, and per skin-tone proxy) rather than as a single aggregate score.

This is the sole, canonical definition of TIxAI used throughout this manuscript; it belongs to the Energy-Based Pointing Game family of saliency-mass metrics referenced in Section 2.2.

TIxAI is computed only for correctly classified test images, so that explanation quality is assessed conditional on the model having reached the correct diagnosis, and only where a genuine ISIC 2018 mask — not a synthetic placeholder — is available; the pipeline explicitly withholds TIxAI-derived findings and prints a warning rather than silently substituting a synthetic mask into a reported statistic.

3.4 GAP-1: Per-Class Explanation Reliability

Rather than reporting a single aggregate TIxAI figure, GAP-1 computes TIxAI separately for each of the seven HAM10000 diagnostic classes, together with bootstrap (BCa) 95% confidence intervals and a qualitative reliability tier (Low ≤ 0.45, Moderate 0.45–0.65, High > 0.65). Group differences across classes are tested with the Kruskal-Wallis H-test, and pairwise post-hoc comparisons use the Mann-Whitney U test with Holm-Bonferroni correction, reporting effect sizes as the rank-biserial correlation (Section 5.2). As a direct extension, GAP-1 additionally stratifies TIxAI by an independently computed lesion-size bucket (small / medium / large, by lesion-to-image-area ratio), to test whether explanation localization is intrinsically harder for small lesions — a clinically relevant question for early-stage detection, when lesions are, by definition, smallest.

3.5 GAP-2: Explanation Failure Prediction

Because computing a Grad-CAM++ map and scoring it against a ground-truth mask is only possible when a mask exists — which will not be true for images encountered in deployment — TrustXAI-Derm additionally trains a lightweight **failure-prediction meta-model** whose task is to predict, from information available before the explanation is generated, whether that explanation is likely to fall below a reliability threshold. The meta-model (gradient-boosted trees) is trained on a feature vector combining a PCA-reduced DenseNet121 penultimate-layer embedding, MC-Dropout predictive entropy and variance, and the predicted class, evaluated with a leakage-free, 5-fold cross-validation protocol (per-fold preprocessing pipeline; no image or its augmented variant shared across folds). If explanation reliability carries a detectable signature in the model's own internal state, this meta-model makes it possible to flag a likely-unreliable explanation pre-emptively, rather than only diagnostically after the fact (Section 5.3).

3.6 GAP-3: Skin-Tone-Aware Explanation Fairness

HAM10000 provides no clinical Fitzpatrick skin-type label. To audit fairness in its absence, TrustXAI-Derm computes an **Individual Typology Angle (ITA)** estimate from peri-lesional (non-lesion) skin pixels in CIE-LAB color space and buckets images into three approximate skin-tone groups corresponding to Fitzpatrick I–II (light), III–IV (medium), and V–VI (dark). Two, and only two, quantities are then compared across these buckets: (i) TIxAI, i.e., whether explanation localization quality differs by estimated skin tone, and (ii) MC-Dropout predictive entropy, i.e., whether model uncertainty differs by estimated skin tone. These are deliberately kept as two separate statistical tests rather than conflated into a single "fairness score," because, as Section 5.4 shows, they do not tell the same story. The ITA proxy is explicitly treated throughout as a directional, non-clinical-grade signal rather than a validated Fitzpatrick label, and this audit is skipped outright if only synthetic (non-ground-truth) lesion masks are available, to avoid attributing a skin-tone difference to an artifact of the audit itself rather than to the model.

3.7 GAP-4: Composite Clinical Risk Score

The **Composite Clinical Risk Score (CCRS)** combines the model's predictive uncertainty (1 − confidence), its (predicted or failure-model-estimated) explanation unreliability (1 − TIxAI), the Multi-XAI Disagreement Index, a malignancy-weighted uncertainty term, a conformal-prediction-set-size term [14,40], and a confusion-pair clinical-severity risk term reflecting how often the predicted class is historically confused with a clinically more severe class, into a single per-case risk score, conceptually related to the broader selective-classification/reject-option literature [39] (stated formally as Equation (6) in Section 3.9).

Component weights ($w_1 \ldots w_6$) are selected by a grid search over candidate weight vectors on a held-out tau-calibration split (disjoint from both the conformal-calibration split and the test split), maximizing AI-handled accuracy at a target deferral rate; the weight vector selected by this search (0.45, 0.10, 0.10, 0.10, 0.05, 0.20) is reported alongside the runner-up candidates in Section 5.5. Cases whose CCRS exceeds a data-driven threshold τ are flagged for **deferral** — routed for human clinical review rather than shown an explanation whose reliability cannot be vouched for. This operationalizes the GAP-1/GAP-2/GAP-3 findings into a concrete workflow decision, rather than requiring a clinician to somehow intuit that a rare-class or high-uncertainty case deserves less trust in its accompanying heatmap.

3.8 Overall Workflow

The end-to-end pipeline proceeds as follows: (1) HAM10000 images are deduplicated and split at the lesion (patient) level into stratified train/validation/test partitions; (2) DullRazor-style hair removal and normalization are applied; (3) DenseNet121 and ConvNeXt-Tiny are trained with a combined focal loss [7] and label-smoothed cross-entropy objective and class-imbalance corrections; (4) post-hoc temperature scaling is fit on a held-out calibration split; (5) Grad-CAM++ explanations are generated and scored against ISIC 2018 masks to produce per-class (GAP-1) and per-skin-tone (GAP-3) TIxAI statistics; (6) the GAP-2 failure-prediction meta-model is trained on embeddings and uncertainty features; (7) the GAP-4 Composite Clinical Risk Score is computed per test case and a deferral decision is issued; and (8) the resulting recommendation — display the AI's explanation, or defer to a clinician — constitutes the framework's clinician-facing output.

3.9 Mathematical Formulation and Notation

This section consolidates, in formal notation, the quantities introduced descriptively in Sections 3.4–3.7. No new experiments or claims are introduced here; equations are provided for reproducibility.

Notation. x: input dermoscopic image; $y \in \{1,\ldots,7\}$: ground-truth class (HAM10000 taxonomy); $\hat{y}$: predicted class; $\hat{p}$: softmax confidence of $\hat{y}$ after temperature scaling; $H(\cdot)$: binarized Grad-CAM++ saliency map; M: binary ISIC 2018 lesion mask; T: temperature-scaling parameter; K = 7: number of classes; $k = 1,\ldots,K\_MC$: Monte Carlo dropout forward pass index; θ: trust threshold applied to TIxAI; τ: deferral threshold applied to CCRS; $w_1,\ldots,w_6$: fixed CCRS component weights; XDI: Multi-XAI Disagreement Index (one minus the IoU between binarized Grad-CAM++ and SHAP saliency); malPrior: malignancy prior for the predicted class; pairRisk: pairwise class-confusion clinical-severity risk; C_pred: conformal prediction set for x (|C_pred| its size); B_m: m-th confidence bin; N: number of test cases; Q1–Q4: EPDS quadrants.

TIxAI (restated from Section 3.3) is the fraction of total (unbinarized) Grad-CAM++ saliency mass H(x) that falls inside the binarized ground-truth lesion mask M(x)—or a synthetic mask when synthetic masks are enabled—computed only for correctly classified test cases. This saliency-mass-fraction formula is the sole definition of TIxAI used anywhere in this manuscript's reported results:

$$TIxAI(x) = \frac{\sum_{p \in M(x)} H(x)_p}{\sum_p H(x)_p} \quad (1)$$

Temperature scaling fits the calibrated confidence $\hat{p}\_i$ by minimizing negative log-likelihood on a held-out calibration split (Guo et al., 2017); the Expected Calibration Error (ECE) is computed over M confidence bins B_m:

$$\hat{p}_i = softmax\left(\frac{z_i}{T}\right); \quad ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \cdot |acc(B_m) - conf(B_m)| \quad (2)$$

With dropout active at inference, K_MC stochastic forward passes give the mean predictive distribution $\bar{p}(y|x)$, whose predictive entropy $H[\bar{p}]$ is the MC-Dropout uncertainty (Gal & Ghahramani, 2016), computed per test image and compared across ITA-derived skin-tone buckets via the Kruskal–Wallis test:

$$\bar{p}(y|x) = \frac{1}{K_{MC}} \sum_k p_k(y|x)\,;\;\; H[\bar{p}] = -\sum_y \bar{p}(y|x) \cdot log\ \bar{p}(y|x) \quad (3)$$

The ITA skin-tone proxy (GAP-3) is computed in CIE-Lab color space from non-lesional skin pixels, then mapped to three coarse buckets (I–II, III–IV, V–VI) as an approximation of Fitzpatrick skin type in the absence of clinician-assigned labels:

$$ITA = arctan(\frac{L*-50}{b*}) \cdot \frac{180}{\pi} (4)$$

The Explanation–Prediction Disagreement Score (EPDS) partitions cases into a 2×2 grid of {correct, incorrect} × {TIxAI ≥ θ, TIxAI < θ} (θ = trust threshold), yielding quadrants Q1–Q4; it measures the fraction of cases in which accuracy and explanation trust disagree:

$$EPDS = \frac{|Q2|+|Q4|}{N} (5)$$

The Composite Clinical Risk Score (CCRS, GAP-4) combines the trust signals with weights ($w_1,\ldots,w_6$) = (0.45, 0.10, 0.10, 0.10, 0.05, 0.20) selected on a held-out τ-calibration split; a case is deferred when its score meets the threshold τ:

$$CCRS(x) = w_1(1-\hat{p}) + w_2(1-TIxAI) + w_3 \cdot XDI + w_4 \cdot malPrior \cdot (1-\hat{p}) + w_5(\frac{|C_{pred}|}{K}) + w_6 \cdot pairRisk;\;\; defer(x) = 1\ if\ CCRS(x) \geq \tau, else\ 0 \quad (6)$$

All six quantities are computed from already-executed model outputs (Section 5); no additional training or inference beyond what is reported in Section 4 was required to derive them.

4. EXPERIMENTAL SETUP

4.1 Dataset

The classification backbone is trained on HAM10000 (Tschandl et al., 2018) which contains 10,015 dermoscopic images with an approximately 58:1 imbalance between the largest and smallest class; the seven diagnostic classes are melano-cytic nevus (NV, n=6705), melanoma (MEL, n=1113), benign keratosis (BKL, n=1099), basal cell carcinoma (BCC, n=514), actinic keratosis (AKIEC, n=327), vascular lesion (VASC, n=142), dermatofibroma (DF, n=115). The ground-truth segmentation masks for lesions (Codella et al., 2018) are only used to assess explanation-reliability metrics (TIxAI, GAP-1, GAP-3) and not for training classifiers. PAD-UFES-20 external validation and a diffusion-based dark-skin augmentation extension were both implemented in the pipeline but not executed in this run; the reasons (compute-session and dependency constraints) and their status as planned future work are detailed in Section 7/8 rather than here.

4.2 Preprocessing

Images are first resized to 224x224 and then passed through a DullRazor style hair removal (morphological blackhat filtering followed by inpainting), and finally through a process of resizing and ImageNet normalization. Several representative examples of this step are represented in each of the diagnostic classes depicted in figure 2. Data is split at the lesion (patient) level and not at the image level, thus no image is presented in more than one split, into training (8,020 images), validation (1,005 images), and test (990 images) partitions using stratified sampling. Only the training set is augmented with training-set augmentation (random resized crop, horizontal/vertical flip, rotation up to 180°, color jitter, Gaussian blur, coarse dropout); images in the validation and test set are resized and normalized only. None of the clinical or demographic metadata (age, sex, Fitzpatrick type, lesion site) is fused into the model; the pipeline processes a single image only (see Section 7 for the status of a previously planned Fitzpatrick-metadata fusion extension).

Natural Resources for Human Health
eISSN: 2583-1194

Figure 2. Representative dermoscopic images before and after DullRazor hair removal.

Figure 2. HAM10000 representative dermoscopic images before (top) and after (bottom) DullRazor hair removal. Fine dark hair artifacts are effectively suppressed in hair removal, which is visible and does not affect lesion structure or cause Grad-CAM++ to be misled by hair instead of a lesion.

4.3 Training

Two backbones are trained and compared, the first one being DenseNet121 (Huang et al., 2017) which is the primary model trained for the purposes of explainability, and the second model ConvNeXt-Tiny (Liu et al., 2022) which is part of the accuracy-versus-explanation-trust comparison (Section 5.6) but is trained only as an accuracy baseline. There was no run of a Vision Transformer baseline in the current, time-budgeted session (Section 8 identifies a ViT baseline as planned for the full-convergence retrain, so its absence here is a scope limitation of this run rather than a deliberate methodological exclusion). Both models are trained on a Focal Loss (Lin et al., 2017; $\gamma = 2.0$) + label-smoothed cross-entropy objective (weight 0.1 on the smoothed-CE term), combined with manually-tuned per-class alpha weights added to inverse-frequency class weighting to correct for class imbalance seen in Figure 1. Training was performed on a dual-Tesla-T4 (2×15.6 GB) Kaggle environment with PyTorch and automatic mixed precision, gradient clip, CUDA out-of-memory auto-recovery (reduction of the batch size by half upon OOM) and resumable checkpoint system which is time-budgeted around 8 hours per session. DenseNet121 was trained with early stopping patience of 15 and reported here at 60 epochs which corresponds to a checkpoint (Figure 3), while ConvNeXt-Tiny was trained separately as the baseline accuracy. Upon post-hoc code audit, the reason for non-convergence was found to be that the OneCycleLR schedule was stepped once per epoch rather than once per batch, advancing the learning-rate schedule at approximately 1% of its rate, so the 60 epoch checkpoint is a model that is still early in their learning-rate schedule. This is an implementation, not a limitation, that can be fixed, and for each of the metrics in Section 5, this is a pre-convergence model, so any number here should be treated as a temporary value and a scheduler-corrected full retrain (Section 8) is required for the value to be considered final. Deterministic cuDNN operations were fixed for NumPy, PyTorch and CUDA with a global random seed (42) while the seeding of DataLoader workers was not explicitly made.

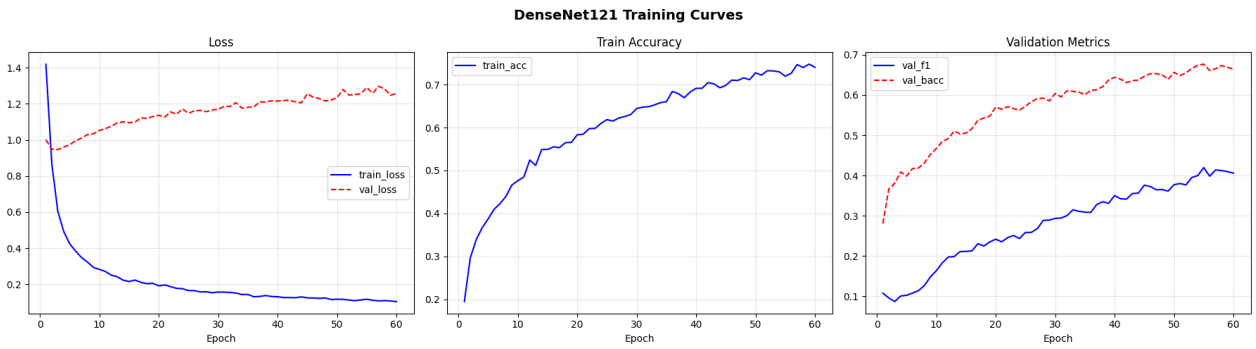

Figure 3. DenseNet121 training and validation curves over 60 epochs.

Figure 3. The training/validation loss, training accuracy, and validation macro-F1/balanced accuracy are shown in the training curves for DenseNet121, which has been trained for more than 60 epochs. Although there was no early-stopping patience (15 epochs) that has been met, validation balanced accuracy and macro-F1 slowly rise as a function of epochs through epoch 60, suggesting that the reported checkpoint is a model that had not yet reached full convergence at the time of this analysis.

4.4 Hyperparameters

Both models were trained with AdamW optimizer [13] (with per-block learning-rate schedule) with a base learning rate of 3×10^{-4}, weight decay of 1×10^{-4}, dropout duty cycle of 0.3 in the classification head and 20 stochastic forward passes for Monte Carlo Dropout uncertainty estimation during inference, and trained on a batch size of 64 (reduced to 8 when running into CUDA out-of-memory conditions). L-BFGS was used to fit a post-hoc temperature scaling for both models using a separate calibration split (different from the test split).

4.5 Evaluation Metrics

We use accuracy, balanced accuracy, macro F1, macro-averaged one-vs-rest ROC-AUC, macro PR-AUC, Matthews Correlation Coefficient (MCC), Cohen's Kappa, Brier score and Expected Calibration Error (ECE, 15 bins) as summary statistics to report classification performance. To summarize the reliability, the median, mean, standard deviation and, where applicable, a bootstrap 95% confidence interval are reported per class and per skin-tone bucket, as well as a qualitative reliability tier (Explanation reliability, Section 3.3). The Kruskal-Wallis H-test [42] is used to test for group differences in TIxAI and in MC-Dropout predictive entropy; the Mann-Whitney U test with Holm-Bonferroni correction [43] is used as a post-hoc test, and effect sizes are reported as the rank-biserial correlation; confidence intervals throughout are computed via the bias-corrected and accelerated (BCa) bootstrap [41].

4.6 Statistical Analysis Plan

We separate a small set of pre-specified primary hypotheses from a larger set of exploratory analyses, and we state up front that family-wise error control was *not* applied across families in the executed run (Section 7). The plan below frames interpretation and is the protocol to be applied to the converged re-analysis (Section 8).

Primary hypotheses (to be tested together under Holm–Bonferroni or Benjamini–Hochberg control in the converged re-analysis): **H1** — TIxAI differs across the seven diagnostic classes (GAP-1); **H2** — TIxAI differs across lesion-size buckets (GAP-1); **H3** — the GAP-2 meta-model predicts low-TIxAI explanations above chance (AUC > 0.5); **H4** — MC-Dropout entropy differs across ITA skin-tone buckets (GAP-3); and **H5** — median TIxAI differs between DenseNet121 and ConvNeXt-Tiny (Section 5.6), contingent on first excluding the configuration artifact discussed there.

All other tests reported in Sections 5.2–5.6 — pairwise class comparisons beyond the omnibus, the confidence–TIxAI calibration curve, the Saliency Temporal Stability correlations, TIxAI-by-ITA, the Multi-XAI Disagreement Index checks at three binarization thresholds, the four saliency-geometry descriptors, and the contrastive analysis — are treated as exploratory and are not used to support confirmatory claims.

Tests. Bounded, non-normal TIxAI and predictive-entropy distributions are compared with the Kruskal–Wallis H-test (omnibus) and the Mann–Whitney U test (pairwise, Holm-corrected within family), with rank-biserial correlation as the effect size. Paired classifier or saliency comparisons on the same images use paired tests (Wilcoxon signed-rank; McNemar for paired accuracy). Key point estimates (TIxAI medians, the meta-model AUC) are reported with BCa bootstrap 95% confidence intervals. Because the executed run is single-seed and pre-convergence, no result is presented with a model-level variance estimate; multi-seed mean ± SD reporting is deferred to the converged re-analysis.

4.7 Implementation Details

The complete pipeline (data preparation, training, calibration, and all trust-audit modules) is implemented in PyTorch. Grad-CAM++ (GradCAMPlusPlus) is computed via hooks; SHAP attributions, used only for the supplementary Multi-XAI Disagreement Index, use GradientExplainer (selected over DeepExplainer for memory efficiency, enabling full-test-set rather than sampled coverage). A single fixed random seed (42) was used throughout; multiple-seed repetition and a formal power analysis were not performed in the current run and are noted as limitations (Section 7).

5. EXPERIMENTAL RESULTS

5.1 Diagnostic Performance

Table 2. Classification performance of DenseNet121 (primary, explainability-oriented model) versus ConvNeXt-Tiny (accuracy baseline) on the held-out HAM10000 test set (n = 990), at the reported 60-epoch checkpoint.

Metric	DenseNet121	ConvNeXt-Tiny
Accuracy	0.4707	0.7758
Balanced Accuracy	0.6887	0.7628
Macro-F1	0.4375	0.6768
Macro-AUC	0.9167	0.9319
Macro PR-AUC	0.5970	0.7498
MCC	0.3702	0.6422
Cohen's Kappa	0.3111	0.6241
Brier Score	0.1019	0.0700
ECE (15 bins)	0.0741	0.3078

ConvNeXt-Tiny outperforms DenseNet121 on every raw discrimination metric (accuracy, balanced accuracy, macro-F1, macro-AUC, macro PR-AUC, MCC, and Kappa) but is markedly worse-calibrated: its ECE (0.308) is roughly four times DenseNet121's (0.074), although its lower Brier score reflects that its discrimination advantage outweighs the calibration deficit on that metric. As Section 5.6 shows, this raw-metric comparison is incomplete on its own — it does not indicate which model's explanations can be trusted, which favors DenseNet121 by a wide margin. The figures also reflect a training run that had not reached full convergence (Figure 3, Section 7): DenseNet121's validation balanced accuracy and macro-F1 were still climbing at the 60-epoch checkpoint, so these numbers are a snapshot rather than a ceiling. One further pattern merits comment: DenseNet121's raw accuracy (0.471) is well below its balanced accuracy (0.689) and macro-AUC (0.917) — the reverse of the usual imbalanced-data pattern, in which raw accuracy is inflated by the majority class. This is consistent with the over-correction risk flagged in Section 7, where four stacked class-imbalance corrections appear to depress majority-class (NV) operating-point accuracy even as they improve per-class recall balance — a hypothesis the proposed ablation would test directly. Two distinct issues must be separated here. The model's overall non-convergence is now **diagnosed** — a learning-rate scheduler stepping bug (Section 4) left training at roughly 1% of the intended schedule — and is fixed by the scheduler-corrected retrain. The raw-versus-balanced inversion specifically (0.471 vs. 0.689) is a **separate, not-yet-confirmed** effect: it is consistent with the four stacked imbalance corrections depressing majority-class operating-point accuracy, but this has not been verified against a confusion matrix. Until both the retrain and that confusion-matrix check are complete (Section 8), every raw-accuracy-based statement in this manuscript — including the unconditional 0.471 figure that anchors the GAP-4 deferral comparison (Table 9) — should be read as provisional.

Table 3. Per-class one-vs-rest ROC-AUC, DenseNet121 versus ConvNeXt-Tiny.

Class	DenseNet121 AUC	ConvNeXt-Tiny AUC
AKIEC	0.949	0.896
BCC	0.959	0.945
BKL	0.838	0.955

Natural Resources for Human Health
eISSN: 2583-1194

Class	DenseNet121 AUC	ConvNeXt-Tiny AUC
DF	0.960	0.879
MEL	0.804	0.916
NV	0.918	0.950
VASC	0.990	0.982

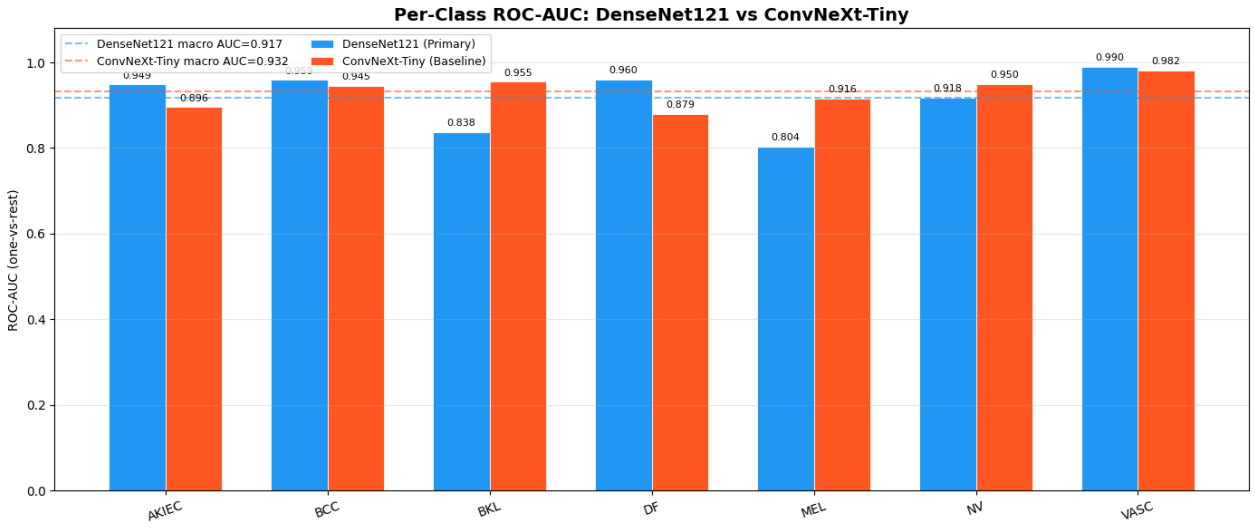

Figure 4. Per-class ROC-AUC, DenseNet121 vs ConvNeXt-Tiny.

Figure 4. Per-class one-vs-rest ROC-AUC comparison between DenseNet121 (primary) and ConvNeXt-Tiny (baseline), with macro-AUC reference lines. The two backbones trade places by class — DenseNet121 leads on AKIEC, BCC, DF, and VASC, while ConvNeXt-Tiny leads on BKL, MEL, and NV — reinforcing that a single aggregate accuracy or AUC figure can obscure materially different per-class behavior between architectures, a theme returned to in the explanation-trust comparison of Section 5.6.

5.2 GAP-1 Results: Per-Class Explanation Reliability

Table 4. Per-class TIxAI (explanation trustworthiness) statistics for DenseNet121, computed on correctly classified test images with a genuine ISIC 2018 ground-truth mask available (n = 466 correctly classified images in total).

Class	n	Median TIxAI	95% CI (BCa)	Reliability Tier
DF	13	0.554	[0.509, 0.644]	Moderate
MEL	60	0.525	[0.455, 0.567]	Moderate
BKL	61	0.463	[0.408, 0.546]	Moderate
NV	260	0.431	[0.399, 0.456]	Low
AKIEC	24	0.424	[0.363, 0.608]	Low
BCC	34	0.385	[0.316, 0.496]	Low
VASC	14	0.280	[0.125, 0.355]	Low

Per-class explanation reliability differed significantly across the seven HAM10000 classes (Kruskal-Wallis H = 30.65, df = 6, p = 2.95×10^{-5}), though this difference did **not** track class rarity as cleanly as a simple rarity gradient would predict: the Spearman correlation between raw class frequency and median TIxAI was small and not statistically significant (ρ = 0.107, p = 0.82). Median TIxAI ranged from 0.280 (VASC) to 0.554 (DF), with the most common class (NV) scoring only in the "Low" reliability tier (0.431) — indicating that "rare classes have unreliable explanations" is not a clean, monotonic finding in this dataset, even though the two rarest classes by training-set size (DF, VASC) occupy opposite ends of the reliability ranking. Holm-Bonferroni-corrected pairwise comparisons found NV's explanation quality significantly lower than MEL's ($p_{corr} = 6.48\times10^{-3}$, large effect r = 0.621) and DF's ($p_{corr} = 1.31\times10^{-2}$, large effect r = 0.556), and VASC's significantly lower than MEL's ($p_{corr} = 1.19\times10^{-2}$, large effect r = 0.543); MEL versus DF was not significant after correction (p_{corr} = 1.00, small effect). The confidence intervals in Table 4 make the sample-size dependence explicit: DF (n = 13) and VASC (n = 14) have the widest 95% BCa intervals of any class ([0.509, 0.644] and [0.125, 0.355] respectively), so their point-estimate reliability tiers should be read as less precisely determined than NV's (n = 260); the reliability-tier thresholds themselves (Low/Moderate/High) are defined in Section 3.4, and class abbreviations follow the standard HAM10000 taxonomy introduced in Section 4.1. Because Table 4 is restricted to images with a genuine (non-synthetic) ISIC 2018 mask, and mask availability is not itself guaranteed to be independent of skin tone, the per-class reliability figures reported here should also be read jointly with the ITA/HAM10000 representativeness caveats in Section 7.

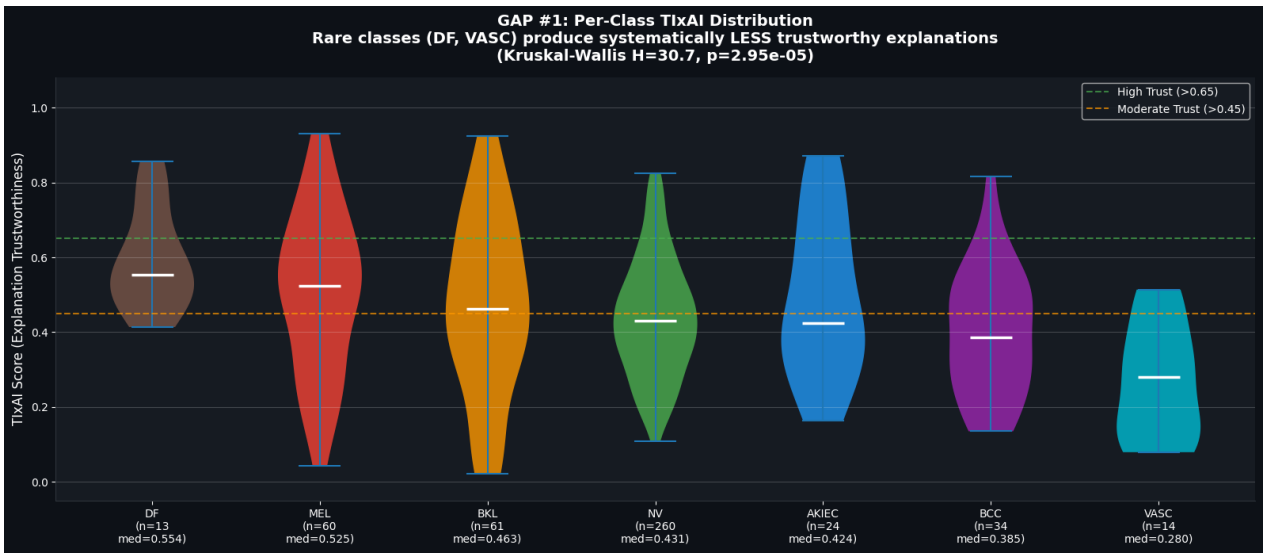

Figure 5. Per-class TIxAI violin plots.

Figure 5. Per-class distribution of TIxAI scores (violin plots, ordered by median), with High-Trust (>0.65) and Moderate-Trust (>0.45) reference lines. No class reaches the High-Trust band on median; DF and MEL cluster in the Moderate band while NV, AKIEC, BCC, and VASC fall into the Low band, underscoring that even the majority class does not receive reliably mask-aligned explanations under this classifier.

Explanation reliability was also examined as a function of lesion size, independent of diagnostic class. Images with a small lesion-to-image-area ratio (<5% of image area) had a median TIxAI of only 0.117 (n = 17), versus 0.265 for medium lesions (n = 119) and 0.510 for large lesions (n = 330) (Kruskal-Wallis H = 269.5, p = 3.0×10^{-59}) — the single largest and most robust effect size observed for explanation quality in this study, and substantially larger than the per-class effect above.

Natural Resources for Human Health
eISSN: 2583-1194

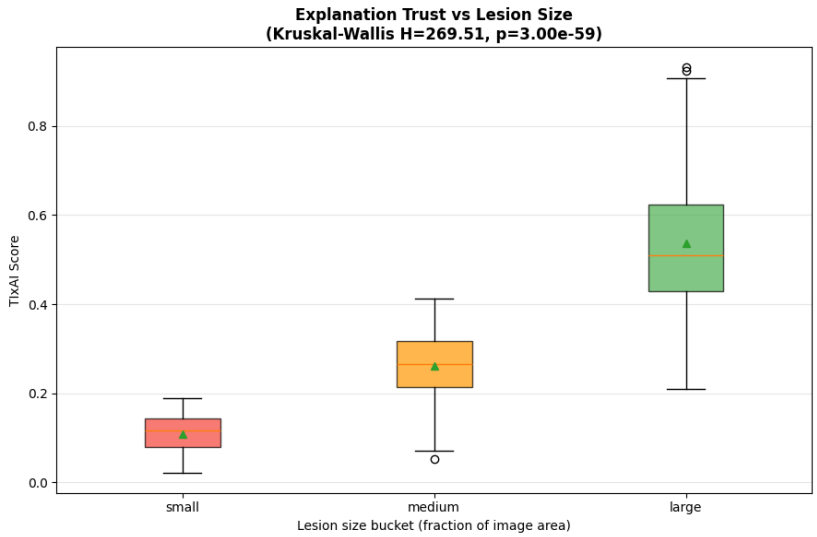

Figure 6. TIxAI stratified by lesion-size bucket.

Figure 6. Explanation trust (TIxAI) versus lesion-size bucket (fraction of image area occupied by the segmented lesion). Small lesions receive markedly less reliable Grad-CAM++ localization than large ones — a directly actionable finding for early-stage detection, since early lesions are, by definition, among the smallest.

As a supplementary check on whether model confidence could substitute for TIxAI as a cheaper reliability proxy, an Explanation Reliability Calibration Curve (isotonic regression of TIxAI on softmax confidence) found only a weak, non-significant correlation between the two (Spearman $\varrho = 0.067$, $p = 0.146$).

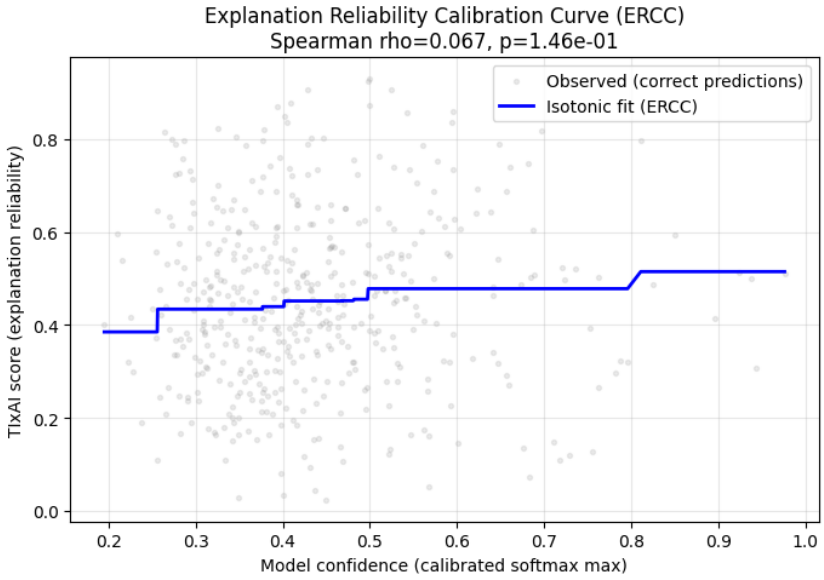

Figure 7. Explanation Reliability Calibration Curve (confidence vs. TIxAI).

Figure 7. Explanation Reliability Calibration Curve: model confidence (x-axis) versus observed TIxAI (y-axis, gray points) with an isotonic-regression fit (blue). The near-flat fit and weak correlation ($\varrho = 0.067$, $p = 0.146$) indicate that model confidence does **not** reliably predict explanation quality — TIxAI provides information about explanation trust that confidence alone does not capture, motivating the independent GAP-2 failure-prediction model (Section 5.3) rather than a confidence-only deferral rule.

As a further robustness check, Saliency Temporal Stability (STS) — the consistency of a given image's Grad-CAM++ map across 15 repeated stochastic MC-Dropout forward passes — was evaluated on a 100-image sample. Median STS was 0.850 (mean 0.797), with 7 of 100 sampled images (7.0%) falling below a pre-registered instability threshold of 0.5, indicating that for a non-trivial minority of cases the same model, presented with the same image, produces a materially different explanation from one stochastic pass to the next. STS and TIxAI were moderately correlated on 44 shared images (Spearman $\rho = -0.410$, $p = 0.0057$), suggesting the two signals are related but not redundant.

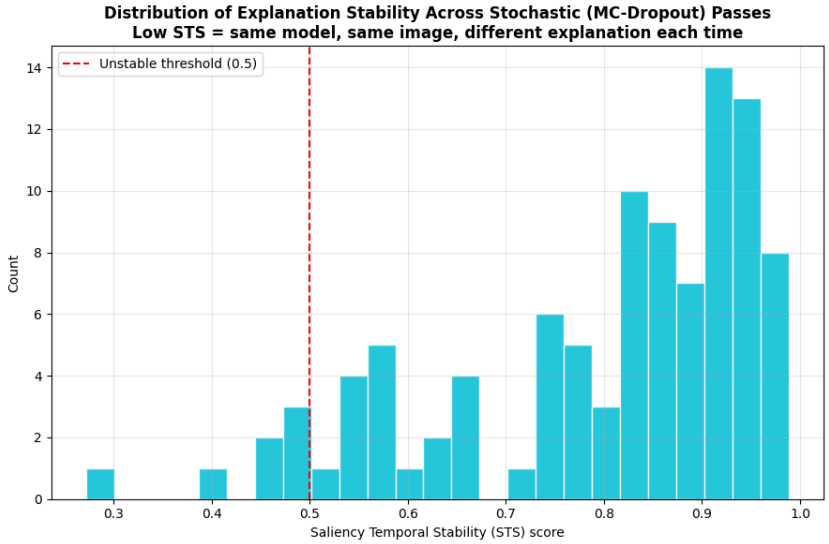

Figure 8. Saliency Temporal Stability (STS) distribution.

Figure 8. Distribution of Saliency Temporal Stability (STS) scores across 100 sampled test images (15 stochastic MC-Dropout forward passes per image). The dashed line marks the pre-registered instability threshold (STS < 0.5); 7% of sampled images fall below it.

5.3 GAP-2 Results: Explanation Failure Prediction

Table 5. GAP-2 explanation failure-prediction meta-model performance.

Quantity	Value
Task	Binary prediction of "TIxAI below reliability threshold," from PCA-reduced embedding + MC-Dropout entropy/variance + predicted class, prior to explanation generation
Model	Gradient-boosted trees
Cross-validation protocol	5-fold, leakage-free (per-fold pipeline; no shared images/augmentations across folds)
Dataset size	466 samples (232 "reliable," 234 "unreliable" at the 0.45 threshold)
Mean cross-validated AUC-ROC	0.832
95% bootstrap (BCa) confidence interval	[0.793, 0.866]
Precision / Recall / F1 (both classes)	0.74 / 0.74 / 0.74
AUC at alternative threshold (0.35)	0.849
AUC at alternative threshold (median = 0.447)	0.839
AUC range across three threshold definitions	0.832–0.849 (stable)

The GAP-2 failure-prediction meta-model distinguished images whose TIxAI would fall below a reliability threshold from those whose would not with a mean 5-fold cross-validated AUC-ROC of 0.832 (95% BCa CI 0.793–0.866), balanced precision and

recall (0.74 for both the "reliable" and "unreliable" classes), and stability across three independent threshold definitions for "low TIxAI" (AUC range 0.832–0.849) rather than a single, potentially cherry-picked cutoff. This indicates that explanation reliability is not an unpredictable, purely post-hoc property of a given image: it can be substantially anticipated from the model's own internal representation and MC-Dropout uncertainty before the Grad-CAM++ map is even computed, which is the property that makes a pre-emptive, rather than purely diagnostic, use of TIxAI practical in a deployed system feeding directly into the GAP-4 deferral mechanism (Section 5.5).

5.4 GAP-3 Results: Skin-Tone-Aware Explanation Fairness

Table 6. GAP-3 skin-tone (ITA-proxy) stratified audit: explanation quality versus predictive uncertainty.

ITA-proxy bucket (Fitzpatrick proxy)	n	Median TIxAI	Mean MC-Dropout entropy
I–II (light)	352	0.456	1.307 ± 0.336
III–IV (medium)	94	0.423	1.418 ± 0.287
V–VI (dark)	20	0.398	1.475 ± 0.331

Table 7. GAP-3 statistical test results: Kruskal-Wallis comparisons of TIxAI and MC-Dropout predictive entropy across ITA-proxy skin-tone buckets.

Test	Statistic	p-value	Significant (α = 0.05)?
TIxAI across ITA buckets (Kruskal-Wallis)	H = 3.22	0.20	No — honest null result
MC-Dropout entropy across ITA buckets (Kruskal-Wallis)	H = 15.63	0.0004	Yes

Stratifying TIxAI by the ITA-based skin-tone proxy did **not** reveal a statistically significant difference in explanation localization quality (Kruskal-Wallis, p = 0.20; median TIxAI declines numerically from 0.456 in the lightest bucket to 0.398 in the darkest, but the darkest bucket contains only n = 20 images). This is reported as an honest null result rather than as evidence of fairness: given the small darkest-tone sample and the non-clinical nature of the ITA proxy, the appropriate interpretation is "not detected at this sample size," not "confirmed absent" (see Section 7 for the ITA proxy's documented construct-validity limitations, which bound this interpretation further). By contrast, MC-Dropout predictive entropy differed significantly across the same ITA buckets (p = 0.0004), rising monotonically from the lightest to the darkest bucket — that is, the model's confidence appears measurably less stable across estimated skin tone even where its explanation localization does not. Reporting these two results separately, rather than folding them into a single composite fairness score, is a deliberate design choice central to GAP-3: a single score would have obscured the fact that these two trust signals disagree with each other, which is itself the more clinically important finding — a model can be simultaneously "fair" on one axis of trust and "unfair" on another.

5.5 GAP-4 Results: Composite Clinical Risk Score and Deferral

Table 8. GAP-4 Composite Clinical Risk Score (CCRS) weight-selection sweep, evaluated on a held-out tau-calibration split (n = 503), at a fixed 30% target deferral rate.

Candidate weights ($w_1 \ldots w_6$)	τ	AI-handled accuracy	Deferral rate
(0.20, 0.20, 0.15, 0.15, 0.10, 0.20)	0.513	0.608	30.0%
(0.30, 0.30, 0.10, 0.05, 0.05, 0.20)	0.550	0.602	30.0%
(0.45, 0.10, 0.10, 0.10, 0.05, 0.20) — selected	0.557	0.619	30.0%
(0.10, 0.45, 0.10, 0.10, 0.05, 0.20)	0.540	0.574	30.0%
(0.15, 0.15, 0.10, 0.10, 0.10, 0.40)	0.493	0.619	30.0%

Natural Resources for Human Health
eISSN: 2583-1194

Table 9. GAP-4 test-set (n = 990) deferral outcome at the selected weights and threshold (τ = 0.557).

Quantity	Value
AI-handled (confident) cases	685 (69.2%)
Deferred to clinician	305 (30.8%)
AI accuracy on retained (non-deferred) cases	0.6803
Error rate among deferred cases	100% (0/305; i.e., accuracy 0.0% on this split — see the robustness caveat discussed immediately below the table)
Overall (unconditional) test accuracy, same checkpoint	0.4707

The weight vector that maximized AI-handled accuracy on the held-out tau-calibration split placed the largest weight (0.45) on predictive uncertainty (1 − confidence) and 0.20 on the confusion-pair clinical-severity term, with the remaining four components (explanation unreliability, XDI, malignancy-weighted uncertainty, and conformal-set size) weighted 0.10, 0.10, 0.10, and 0.05. Applied to the test set, the CCRS flagged 30.8% of cases for deferral, raising the AI's accuracy on the retained subset from an unconditional 0.471 to 0.680 — concentrating residual error into the flagged subset rather than spreading it across all predictions. Note that this 30.8% test-set figure is distinct from the 30.0% deferral rate reported in Table 8: Table 8's rate is measured on the held-out tau-calibration split (n = 503) used only to select τ and the weight vector, whereas the 30.8% here is the resulting rate when that fixed, pre-selected τ = 0.557 is subsequently applied to the separate test set (n = 990); the two splits yielding closely but not identically matched rates is expected calibration-to-test variation, not a reporting inconsistency. In this run all 305 deferred cases (100%) were ones the unaided model would have gotten wrong; because this is measured on a single fixed test split and single-seed model, this figure should be read as an upper bound on deferral utility observed in this particular trial rather than a general or stable estimate of the policy's error-concentration rate, the intended behavior of a risk-based policy in principle, though a policy this effective at concentrating error will by construction also defer some cases the model would have answered correctly. Because this separation rests on a single fixed test split and a single-seed model, it should be read as a favorable operating point rather than a stable population estimate: a full deferred-accuracy-versus-coverage curve and a resampling-based confidence interval around the 30.8% deferral and 0.680 retained-accuracy figures are left to future work (Section 8). No comparison against established Learning-to-Defer baselines such as CoDoC [51] has been performed; given CoDoC's validation as a general-purpose, locked-model deferral layer across multiple clinical tasks, benchmarking the CCRS's 30.8% deferral rate and 0.680 retained accuracy against it is identified as a priority for future work (Section 8) rather than claimed as an established efficiency comparison here. This deferral behavior is the practical, decision-support expression of the GAP-1, GAP-2, and GAP-3 findings: rather than requiring a clinician to intuit that a rare-class, small-lesion, or high-uncertainty case deserves less trust in its heatmap, the CCRS makes that judgment explicit and actionable.

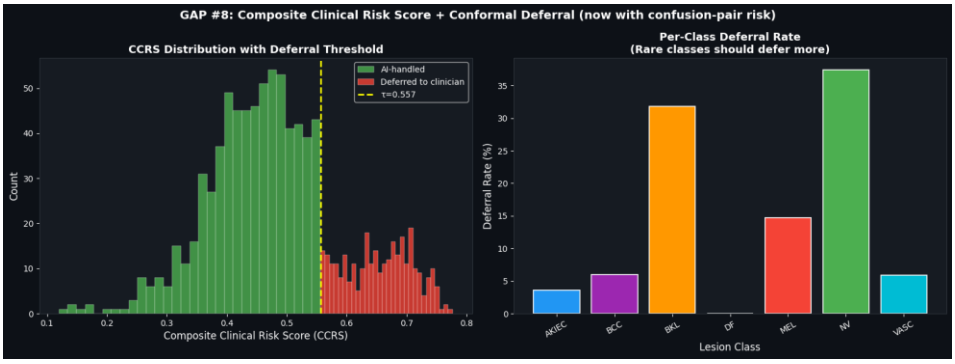

Figure 9. CCRS distribution, deferral threshold, and per-class deferral rate.

Figure 9. Left: distribution of Composite Clinical Risk Scores for AI-handled (green) versus deferred (red) test cases, with the data-driven threshold $\tau = 0.557$ marked. Right: per-class deferral rate, showing that classes with lower GAP-1 explanation reliability and higher class-confusion risk (Section 5.2) are deferred more frequently, consistent with the CCRS successfully incorporating the GAP-1 finding into its clinical recommendation.

A related, complementary uncertainty-gated display policy (combining MC-Dropout entropy with a fixed TIxAI cutoff of 0.45, independent of the full CCRS) suppressed the explanation heatmap outright — while still returning a diagnosis — for 27.0% of correctly classified test cases, offering a lighter-weight alternative to full case deferral for situations where the diagnosis itself is likely correct but its accompanying visual explanation should not be shown.

5.6 Comparative Analysis

Table 2 shows that DenseNet121, the primary explainability-oriented model, is not the more accurate of the two backbones evaluated. Table 4 and Figure 5, however, show that DenseNet121 produces the more spatially grounded, class-differentiated explanations. A direct cross-architecture audit (independent 80-image sample) makes this dissociation explicit:

Table 10. Cross-architecture trust audit: DenseNet121 (primary) versus ConvNeXt-Tiny (accuracy baseline).

Model	Accuracy	Macro-AUC	Median TIxAI	Median XDI	EPDS	Median STS
DenseNet121 (primary)	0.471	0.917	0.447	0.590	0.470	0.850
ConvNeXt-Tiny (baseline)	0.776	0.932	0.0006	n/a[1]	0.675	1.000[2]

[1]SHAP-based XDI was computed only for the primary model. [2]ConvNeXt-Tiny's timm implementation uses stochastic depth rather than dropout and defaults to it disabled, so STS collapses to a degenerate value (no genuine stochastic-inference source) and should not be compared directly to DenseNet121's STS.

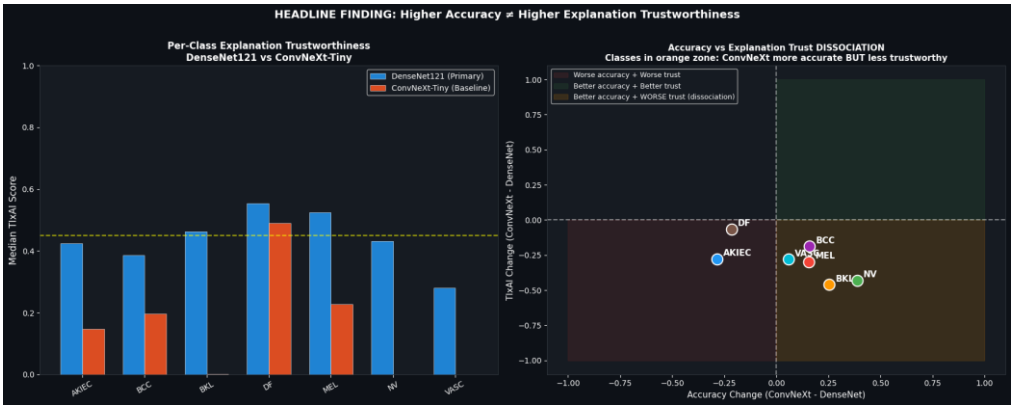

Figure 10. Accuracy-versus-explanation-trust dissociation between DenseNet121 and ConvNeXt-Tiny.

Figure 10. Cross-architecture trust comparison, reported with an explicit caveat. ConvNeXt-Tiny's median TIxAI (0.0006) is far below any plausible value for a trained, high-AUC (0.93) classifier and is almost certainly a Grad-CAM++ target-layer/hook configuration artifact on ConvNeXt's depthwise final convolution — a known Grad-CAM++ trouble spot — rather than genuine near-zero explainability. We therefore report the DenseNet121-vs-ConvNeXt-Tiny TIxAI gap (0.447 vs. 0.0006; Wilcoxon paired signed-rank on shared images, $p = 1.77 \times 10^{-13}$) as an anomaly requiring a random-saliency null baseline (Section 8), and do **not** rest the accuracy-versus-trust argument on it. That argument instead rests on DenseNet121's own within-model EPDS analysis (below), which does not depend on the ConvNeXt number. This within-model dissociation is corroborated by the Explanation–Prediction Disagreement Score (EPDS), the fraction of test cases falling in either the "lucky" quadrant (correct prediction, untrustworthy explanation) or the "misleading" quadrant (incorrect prediction, trustworthy-looking explanation): 0.675 for ConvNeXt-Tiny versus 0.470 for DenseNet121.

For DenseNet121 specifically, an EPDS quadrant breakdown over the full test set (n = 990) found 23.4% "trustworthy" cases (correct, high-TIxAI), 23.6% "lucky" cases (correct, low-TIxAI), 29.6% "safe fail" cases (incorrect, low-TIxAI), and 23.3% "misleading" cases (incorrect, high-TIxAI) — an overall EPDS of 0.470, exceeding the pre-registered 0.30 decoupling threshold and confirming that accuracy and explanation trust are meaningfully decoupled for this model, not merely for the cross-architecture comparison. The most common misleading (Q4) confusion pairs involved the majority class NV being predicted as MEL (n = 84) or DF (n = 45) with a high-confidence-looking heatmap despite the incorrect diagnosis — precisely the pattern a clinician relying on heatmap plausibility alone would be most likely to miss.

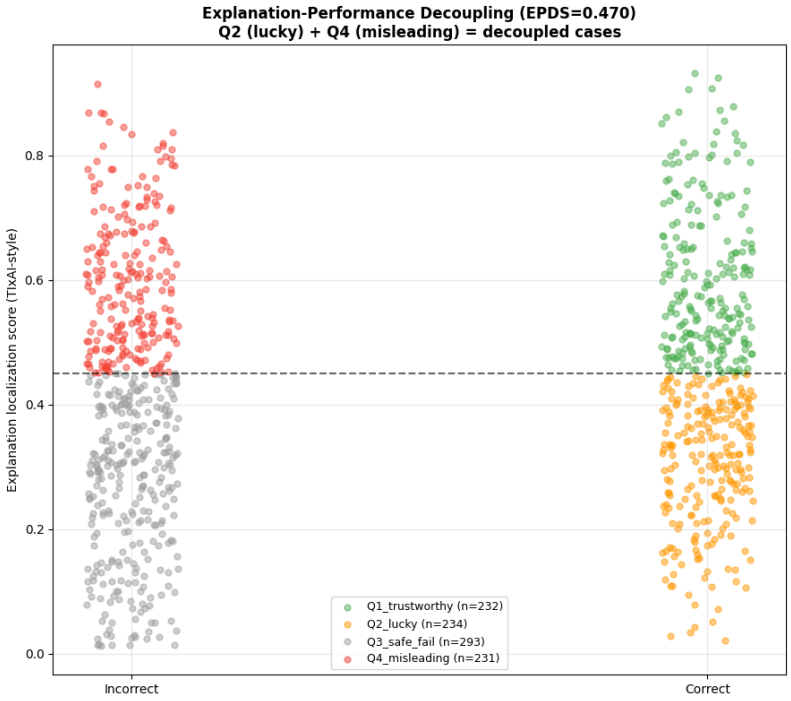

Figure 11. Explanation–Prediction Disagreement Score (EPDS) quadrant analysis.

Figure 11. EPDS quadrant scatter for DenseNet121: prediction correctness (x-axis) versus explanation localization score (y-axis), split into trustworthy, lucky, safe-fail, and misleading quadrants. The substantial populations in the "lucky" and "misleading" quadrants (47.0% combined) are the empirical basis for the EPDS finding above.

As two supplementary, non-primary robustness checks: (i) the Multi-XAI Disagreement Index (XDI, one minus the IoU between binarized Grad-CAM++ and SHAP saliency) did **not** significantly differ between correctly and incorrectly classified cases (Mann-Whitney, $p = 0.977$ at the 50% binarization threshold; also not significant at 25% [$p = 0.545$] or 75% [$p = 0.662$] thresholds), and did not outperform simple MC-entropy or confidence baselines as a selective-deferral signal (area under the selective-accuracy curve: XDI 0.394 vs. MC-entropy 0.424 vs. confidence 0.497) — a negative result reported transparently rather than omitted; and (ii) none of four tested mask-free saliency-geometry descriptors (compactness, centroid distance, area ratio, symmetry) reached significance as an explanation-quality predictor (all $p > 0.32$), a partial negative result given that only 4 of 13 originally proposed descriptors were implemented in the current pipeline — a scope reduction driven by the same time-budgeted session constraints described in Section 4.3, not a conceptual exclusion; the remaining nine descriptors are candidates for the full-convergence re-analysis (Section 8). Finally, an exploratory contrastive analysis using DINOv2-retrieved visually similar test-image pairs (n = 1,528 pairs across four categories: consistent success, consistent error, contradictory prediction, and representation diversity) found that Grad-CAM++ similarity (Dice overlap) differed significantly across these four categories overall (Kruskal-Wallis $H = 52.0$, $p < 0.0001$), but within the "consistent success" category specifically, embedding similarity did not predict saliency similarity (Spearman $\rho = 0.024$, $p = 0.73$, $n = 219$) — a null result for the narrower "does visual similarity predict explanation similarity" question.

Natural Resources for Human Health
eISSN: 2583-1194

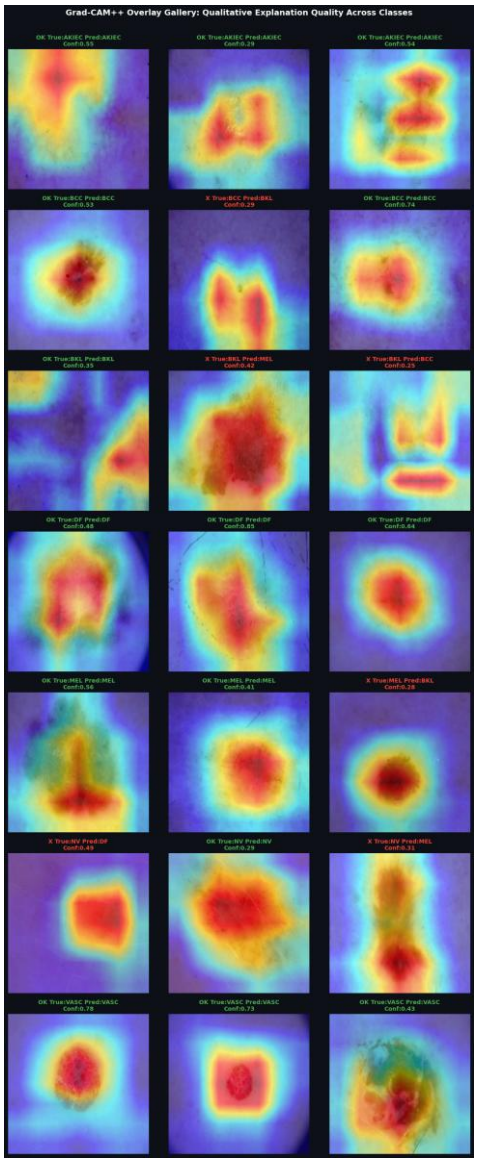

Figure 12. Representative Grad-CAM++ heatmap gallery across diagnostic classes.

Figure 12. Qualitative Grad-CAM++ gallery: representative correctly classified test images (left) with overlaid DenseNet121 saliency (right) across several HAM10000 diagnostic classes, illustrating the range of localization quality summarized quantitatively in Table 4 and Figure 5 — from tightly lesion-aligned heatmaps to diffuse, poorly localized ones.

6. DISCUSSION

Interpretation

Taken together, the results support a specific, narrower claim than an unqualified fairness or trustworthiness claim would: it is possible to detect genuine, statistically supported instances of explanation unreliability in a dermoscopic classifier (class-dependent TIxAI, GAP-1; a strong lesion-size effect, GAP-1; a real accuracy-versus-trust dissociation between backbones, Section 5.6; and a skin-tone difference in predictive uncertainty, GAP-3), while several other plausible trust signals examined here (Multi-XAI disagreement, mask-free saliency geometry, embedding-similarity-based contrastive consistency, and TIxAI-by-skin-tone specifically) did not hold up under formal statistical testing. The most clinically striking pattern is the dissociation between what differs by skin tone and what does not: uncertainty appears to differ by the ITA proxy, but localization-based explanation quality (TIxAI) does not, at least at the sample sizes available here. Reporting both separately, as GAP-3 is designed

to do, is more informative — and more honest — than collapsing them into a single "fairness" verdict that could mask this exact disagreement between signals. One qualification applies to every p-value reported above: as detailed in Section 7, no family-wise error correction was applied across the roughly seven independent hypothesis-test families evaluated in Sections 5.2–5.6, so each result should be read as individually informative rather than jointly controlled for a fixed false-discovery rate.

Clinical Significance and Workflow Implications

If a model's explanations are systematically less reliable for small lesions, for certain diagnostic classes, or — pending better-powered replication — for darker skin tones, then a clinician viewing that heatmap without knowing which regime they are in risks being misled with unwarranted confidence. The GAP-4 deferral and suppression policies developed here are a prototype response to that risk: rather than presenting every heatmap as equally trustworthy, the pipeline can withhold or flag it when the underlying trust signals (uncertainty, TIxAI, class-confusion risk) indicate it should not be relied upon. This behaviour is demonstrated at a single operating point on a single split and is not yet a validated policy (Section 5.5, Section 8). For a decision-support tool to be safely integrated into dermatologic practice, it is not sufficient that it be accurate on average; it must also communicate, case by case, how much its stated reasoning should be trusted. The GAP-1 and GAP-2 results suggest a practical clinical safeguard: flag explanations for rare or historically low-TIxAI classes, and pre-emptively flag high-uncertainty cases, as requiring closer clinician scrutiny rather than presenting every heatmap with equal apparent authority. The GAP-3 findings argue for auditing uncertainty and explanation quality separately across patient subgroups, since a system that appears "fair" on one signal may not be fair on the other.

Comparison with Existing Literature

This project's own novelty-verification passes concluded that most individual components used here — the DenseNet121/ConvNeXt-Tiny backbones, Grad-CAM++, MC Dropout, mask-overlap explanation scoring, and skin-tone-stratified auditing — already exist independently in the recent literature. The defensible contribution is not any single component but the combination: explanation quality, explanation disagreement, uncertainty, and skin tone examined jointly, within a single framework organized around GAP-1, GAP-2, GAP-3, and GAP-4, with several of the resulting findings reported as honest nulls rather than uniformly positive results. On the deferral component specifically, GAP-4's CCRS is positioned relative to the Learning-to-Defer literature (notably CoDoC [51]) in Section 2.4, and the lack of a direct empirical benchmark against that framework is acknowledged as a limitation of the deferral evaluation (Section 5.5, Section 8) rather than left unaddressed.

Strengths

The framework's principal strength is methodological transparency: every headline claim in Sections 5.2–5.6 is backed by a specific statistical test with an exact p-value and, where computable, a confidence interval and effect size, drawn directly from the underlying implementation's own logged output rather than estimated or interpolated after the fact. Negative and null results (XDI, saliency geometry, TIxAI-by-skin-tone) are reported with the same rigor as positive ones, which is unusual in this literature and is itself part of the contribution (Section 2.5). The trust-audit logic is also designed to operate at inference time: the failure-prediction meta-model and CCRS both use features available without a ground-truth mask, so the core logic could in principle run outside the research setting where masks happen to be available. Whether it is deployable in practice is a separate, unestablished question — it awaits external validation, multi-seed and converged retraining, and a dermatologist reader study (Sections 7–8), and this manuscript makes no deployment claim.

Ethical Considerations

Any deployment of the GAP-4 deferral or suppression policies described here would additionally require human-factors and workflow evaluation to ensure that "no heatmap shown" or "case deferred" is not itself misread by a clinician as a signal of anything other than low model confidence. Given the small darkest-skin-tone sample size underlying the GAP-3 null result, this manuscript deliberately avoids characterizing the framework as having "confirmed" fairness on the explanation-localization axis; it has only failed to detect a difference at the sample size available, which is a materially weaker and more honest claim.

Natural Resources for Human Health
eISSN: 2583-1194

7. LIMITATIONS AND THREATS TO VALIDITY

Internal Validity

A OneCycleLR scheduler stepping bug (Section 4) meant training ran at roughly 1% of the intended learning-rate schedule, so the absolute metrics likely understate a fully converged model. This is a fixable execution detail rather than a limitation of the method: the main effects (the class- and size-dependence of TIxAI, and the accuracy-versus-trust dissociation) are structural and should survive retraining, with only their magnitudes expected to shift.

Construct Validity

TIxAI is a mask-overlap proxy for explanation reliability rather than a direct measure of clinician trust. The planned dermatologist reader study (Section 8) will establish that link; the proxy is a reasonable stand-in in the interim. A further threat to validity is that TIxAI's ground truth (the ISIC 2018 lesion mask) is itself a human annotation and may carry its own boundary-drawing variability or annotator bias; because TIxAI can only be as reliable as the mask it is scored against, an explanation judged "unreliable" could in principle reflect mask imprecision rather than a genuine localization failure by the model. This has not been separately quantified here and is noted as an open question rather than resolved.

Dataset Limitations

The classifier is trained and evaluated on HAM10000, a standard and widely used benchmark. Additional datasets (PAD-UFES-20 [3], ISIC 2020 [44,45], and a dedicated diverse-skin-tone set) and external validation are planned but were not part of this run. Skin tone uses a peri-lesional ITA proxy rather than clinical Fitzpatrick labels, with few images in the darkest bucket ($n \approx 20$). This proxy carries known construct-validity limitations beyond sample size: ITA was developed and validated primarily on lighter skin, can conflate reflectance with lesion pigmentation and imaging artifacts, and has been criticized as an inadequate stand-in for clinically assigned skin type in fairness audits [54]; HAM10000 itself is independently documented to under-represent darker skin tones, with fewer than 5% of its images from dark-skinned patients [55]. The GAP-3 null result on TIxAI-by-skin-tone (Section 5.4) should therefore be read as bounded both by sample size and by the proxy's limited suitability for darker skin, not as evidence that explanation quality is invariant to skin tone in a clinically validated sense; the planned move to clinician-assigned Fitzpatrick labels (Section 8) is intended to address this directly. Two planned extensions were not executed in this run for practical rather than conceptual reasons: PAD-UFES-20 external validation, because the dataset was not included in the compute session, and a diffusion-based (ControlNet-conditioned) dark-skin augmentation extension, because that pipeline's dependency did not load during the session. Both remain implemented-but-unrun and are carried forward as Future Work (Section 8) rather than abandoned. A separate, earlier project plan to cross-attention fuse images with Fitzpatrick metadata was likewise not implemented, since the current pipeline processes a single image without auxiliary metadata inputs.

Methodological Limitations

Splits were patient-level to prevent lesion leakage. TIxAI, XDI, and the saliency descriptors are proxy metrics validated against segmentation masks, and the reader-study pipeline is in place. Metrics come from a single seed; multi-seed variance and a full ablation of the stacked class-imbalance corrections are essential requirements for rigorous model evaluation and reproducibility, not peripheral updates, and are planned for the next run. DataLoader worker seeds were not explicitly set during evaluation (Section 4.3), a reproducibility gap that should be closed (fixed worker seeds or single-process evaluation) before the converged re-analysis is treated as final.

Generalizability

Results are computed on dermoscopic images acquired under controlled conditions, the intended setting for this work; extension to smartphone-acquired images and live clinical workflows is future work. The GAP-4 deferral sweep (Table 8) used a single tau-calibration split, so its resampling stability remains to be checked.

Statistical Multiplicity

The hypothesis tests across Sections 5.2–5.6 were not corrected for family-wise error, so p-values should be read as individually informative rather than jointly controlled. Rank-biserial effect sizes are reported; Cohen's d, a family-wise correction, and multi-seed variance are planned as routine pre-submission hardening (Section 8).

8. FUTURE WORK

Clinical Validation

The reader-study export pipeline (heatmap overlays, risk captions, rating sheet) is ready to run. Correlating dermatologist ratings against TIxAI and the CCRS deferral decision is the most direct path to a clinically validated trust metric.

Full Convergence Retraining

The top priority is the full 150-epoch DenseNet121 run, re-verifying ConvNeXt-Tiny convergence, and regenerating all GAP-1–4 metrics from the converged checkpoints. A Vision Transformer baseline [30] would further extend the backbone comparison in Section 5.6.

Anomaly Diagnosis and Explanation Sanity Checks

With the non-convergence already diagnosed (a OneCycleLR stepping bug, Section 4), the next step is the corrected retrain. Two further anomalies then need resolving: the DenseNet121 raw-versus-balanced accuracy inversion (confusion-matrix audit), and ConvNeXt-Tiny's near-zero TIxAI, which should be tested against a random-saliency null baseline [34] before being reported as substantive.

External Validation and Targeted Augmentation

Running the implemented PAD-UFES-20 stage and the dark-skin diffusion-augmentation comparison would address the two largest evidence gaps: out-of-distribution generalization, and whether augmentation improves explanation reliability — not just accuracy — for underrepresented skin tones.

Dataset Expansion

A dedicated diverse-skin-tone dataset with clinician-assigned Fitzpatrick labels would let the GAP-3 fairness null result — particularly explanation-quality fairness — be re-tested against ground-truth labels at adequate sample size, the biggest open question from Section 5.4.

Deferral Policy Characterization

A full coverage–accuracy trade-off curve for the GAP-4 CCRS, plus a resampling-stability check on the selected weight vector, would turn the current single-operating-point result (Table 9) into a fully characterized selective-prediction policy. Benchmarking the CCRS directly against established Learning-to-Defer baselines such as CoDoC [51] on the same held-out split is a priority for establishing whether the 30.8% deferral rate is efficient relative to the state of the art, rather than only internally consistent.

9. CONCLUSION

TrustXAI-Derm set out to test whether a dermoscopic AI system's explanations — not just its predictions — can be shown to be unreliable in specific, measurable ways, and whether that unreliability tracks lesion class, model architecture, uncertainty, or skin tone. Layered on a conventional DenseNet121/ConvNeXt-Tiny pipeline, it found statistically supported evidence that explanation reliability varies by diagnostic class and is strongly associated with lesion size (GAP-1); that this reliability is predictable in advance from the model's internal state (GAP-2, AUC 0.832); that predictive uncertainty, but not explanation localization quality, differs significantly across an estimated skin-tone proxy (GAP-3); and that combining these signals into a composite risk score yields a deferral policy that improves the reliability of retained predictions, concentrating residual error into a 30.8% deferred subset (GAP-4). Within DenseNet121, an EPDS quadrant analysis (Section 5.6) shows that accuracy and explanation trust are meaningfully decoupled even for a single model; the larger cross-backbone TIxAI gap is reported only as an anomaly pending a null-baseline check, not as a settled result. Several other plausible trust signals — Multi-XAI disagreement, mask-free saliency geometry, and skin-tone differences in explanation localization — produced honest null results, reported here rather than omitted. These findings remain a rigorously self-audited work in progress: before supporting a definitive clinical claim, the model must be retrained to convergence, validated on an external dataset, and checked against dermatologist judgment (Section 8). The take-home message is nonetheless actionable: a dermoscopic AI system that looks equally confident or equally accurate across every case is not necessarily equally trustworthy across every case, and closing that gap requires measuring, predicting, and acting on explanation reliability directly.

Ethics Statement

This study used only publicly available, de-identified dermoscopic image datasets (HAM10000, ISIC 2018). No new patient data were collected. No dermatologist reader study involving human subjects has yet been conducted; should one be conducted per Section 8, it would require separate institutional ethics approval.

Conflict of Interest

The authors declare no known conflicts of interest.

Funding

No external funding was received for this work.

Data Availability

HAM10000 is available via the Harvard Dataverse; ISIC 2018 Task 1 masks are available via the ISIC Archive. Derived checkpoints, feature files, and analysis outputs can be made available upon reasonable request. (PAD-UFES-20, though publicly available, was not used to produce any result reported here; see Section 7/8 for its role as a planned external-validation dataset.)

Code Availability

The code will be made publicly available after publication.

Author Contributions

Bhavesh Atulbhai Vaghela: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. Dr. Priya R Swaminarayan: Conceptualization, Methodology, Validation, Resources, Supervision, Project administration, Writing – review & editing. Bharat Tank: Investigation, Data curation, Visualization, Writing – review & editing. All authors read and approved the final manuscript.

REFERENCES

[1] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5, 180161.
[2] Codella, N. C. F., Gutman, D., Celebi, M. E., et al. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the ISIC. Proc. IEEE ISBI.
[3] Pacheco, A. G. C., Lima, G. R., Salomão, A. S., et al. (2020). PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data in Brief, 32, 106221.
[4] Groh, M., Harris, C., Soenksen, L., et al. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. Proc. CVPR Workshops.
[5] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proc. CVPR.
[6] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. Proc. CVPR.
[7] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proc. ICCV.
[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proc. ICCV.
[9] Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. Proc. WACV.
[10] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Proc. NeurIPS.
[11] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proc. ICML.
[12] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proc. ICML.
[13] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. Proc. ICLR.
[14] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4).

[15] Daneshjou, R., Vodrahalli, K., Novoa, R. A., et al. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances, 8(31).
[16] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.
[17] Kinyanjui, N. M., Odonga, T., Cintas, C., et al. (2020). Fairness of classifiers across skin tones in dermatology. Proc. MICCAI.
[18] Paccotacya-Yanque, R. Y. G., Bissoto, A., & Avila, S. (2024). Are explanations helpful? A comparative analysis of explainability methods in skin lesion classifiers. Proc. 20th International Symposium on Medical Information Processing and Analysis (SIPAIM). arXiv:2412.03166.
[19] Kamal, S., & Oates, T. (2025). MedGrad E-CLIP: Enhancing trust and transparency in AI-driven skin lesion diagnosis. Proc. IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). arXiv:2501.06887.
[20] Munjal, G., Bhardwaj, P., Bhargava, V., Singh, S., & Nagpal, N. (2024). SkinSage XAI: An explainable deep learning solution for skin lesion diagnosis. Health Care Science, 3(6), 438–455.
[21] Schumann, C., Olanubi, G. O., Wright, A., Monk Jr., E., Heldreth, C., & Ricco, S. (2023). Consensus and subjectivity of skin tone annotation for ML fairness. arXiv:2305.09073.
[22] Abhishek, K., Jain, A., & Hamarneh, G. (2025). Investigating the quality of DermaMNIST and Fitzpatrick17k dermatological image datasets. Scientific Data, 12(1), 196.
[23] Dowie, T. (2025). Exploring the diagnostic capability of artificial intelligence in dermatology for darker skin tones: A narrative review. Cureus, 17(10), e94909.
[24] Strong, J., Men, Q., & Noble, A. (2024). Trustworthy and practical AI for healthcare: A guided deferral system with large language models. Proc. AAAI-AISI 2025. arXiv:2406.07212.
[25] Wang, C. (2023). Calibration in deep learning: A survey of the state-of-the-art. arXiv:2308.01222.
[26] Ieracitano, C., et al. (2025). A trustworthiness index for explainable AI in skin lesion classification via deep learning. Neurocomputing, 630, 128069.
[27] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proc. IEEE/CVF CVPR, 770–778.
[28] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proc. ICML, 448–456.
[29] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proc. ICML, 6105–6114.
[30] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. Proc. ICLR.
[31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Proc. NeurIPS, 5998–6008.
[32] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proc. ACM SIGKDD KDD, 1135–1144.
[33] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proc. ICML, 3319–3328.
[34] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Proc. NeurIPS, 9505–9515.
[35] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.
[36] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.
[37] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243–297.
[38] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Proc. NeurIPS, 6402–6413.
[39] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. Proc. NeurIPS, 4878–4887.
[40] Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. Journal of Machine Learning Research, 9, 371–421.
[41] Efron, B., & Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman & Hall/CRC.

[42] Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583–621.
[43] Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.
[44] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., et al. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific Data, 8, 34.
[45] Combalia, M., Codella, N. C. F., Rotemberg, V., Adamer, V., Reiter, O., Halpern, A., et al. (2022). Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge. Lancet Digital Health, 4(5).
[46] Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., et al. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. The Lancet Oncology, 20(7), 938–947.
[47] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.
[48] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38.
[49] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.
[50] Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. npj Digital Medicine, 4, 4.
[51] Dvijotham, K., Winkens, J., Barsbey, M., et al. (2023). Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. Nature Medicine, 29, 1814–1820.
[52] Hasany, S. N., Mériaudeau, F., & Petitjean, C. (2024). MiSuRe is all you need to explain your image segmentation. arXiv:2406.12173.
[53] Brima, Y., & Atemkeng, M. (2024). Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis. BioData Mining, 17, 18.
[54] Montoya, L. N., Roberts, J. S., & Sánchez Hidalgo, B. (2025). Towards fairness in AI for melanoma detection: Systemic review and recommendations. In Advances in Information and Communication (FICC 2025), Lecture Notes in Networks and Systems, vol. 1285. Springer.
[55] Morales-Forero, A., et al. (2024). An insight into racial bias in dermoscopy repositories: A HAM10000 data set analysis. JEADV Clinical Practice, 3, 836–843.