DERMEQUITY: A UNIVERSAL SAFETY FRAMEWORK FOR MEDICAL AI USING WORST-CASE BIAS CERTIFICATION AND INFLUENCE-BASED ATTRIBUTION VALIDATED IN SKIN CANCER SCREENING

Angie Xiu, Alex Mi, Kris Lau
Signature School, Evansville, Indiana

BACKGROUND / OBJECTIVE

- **Skin Cancer:** 112,000 new melanoma cases annually in the US - **early detection is life-saving.**
- Fewer than 1 dermatologist per 100,000 people in low/middle income countries: **AI fills critical gaps**
- Less than 10% of dermatology AI training images **represent darker skin tones**
- Standard metrics report average accuracy, **hiding dangerous diagnostic gaps** on dark skin
- **No pre-deployment worst-case safety standard exists** for medical AI

<10%

of dermatology AI training images represent darker skin tones

9.5%

dark skin representation in Fitzpatrick17k, the primary benchmark

$1,000+

cost of professional dermoscopes — out of reach for most communities

112K

new melanoma cases annually in the US — early detection saves lives

Objective: Build **a universal stress-test framework for medical AI safety testing** with Worst-Case Underdiagnosis Gap (WCUG) and Influence-Based Bias Attribution and an **accessible screening system for equitable skin cancer detection.**

METHODOLOGY

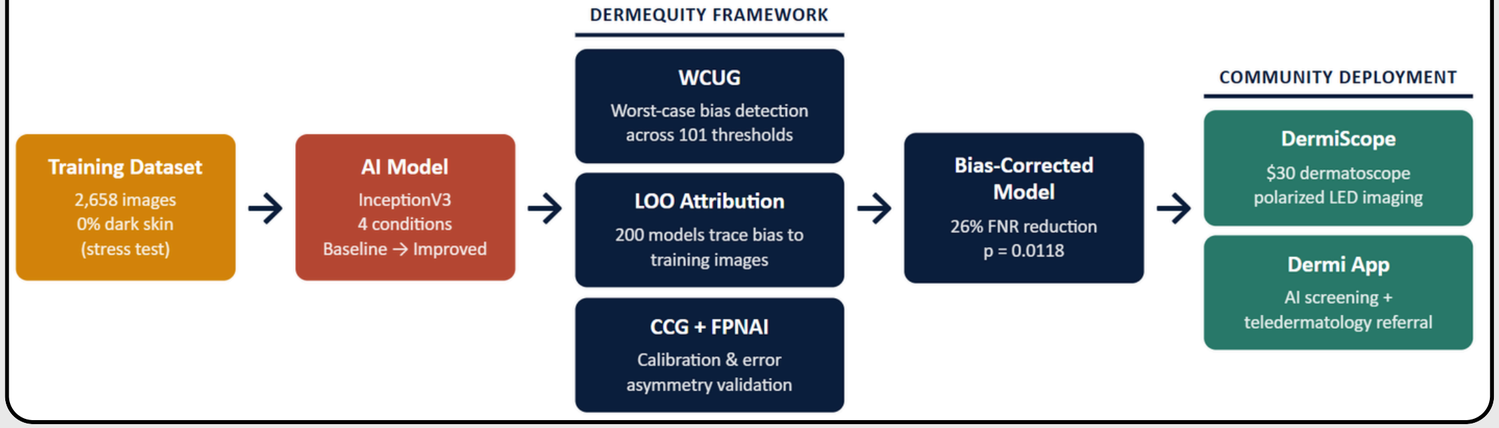
Component 1: DermEquity Benchmark

- 1,658 curated skin lesion images from Fitzpatrick17k
- Stratified by Fitzpatrick skin type and malignancy
- **100% of available dark skin** images used (411)
- Intentional stress-test: 0% dark skin in training simulates worst-case deployment

Component 2: WCUG (Worst-Case Underdiagnosis Gap)

- WCUG sweeps every possible classification threshold to find the single worst-case melanoma miss rate gap between skin tones – **exposing failures** that average accuracy hides entirely.
- **Sweeps all 101 classification thresholds**
- Finds the maximum melanoma miss rate gap between skin tones
- Validated with 1,000-iteration bootstrap confidence intervals
- Sensitivity analysis confirms **stability above 1,000 images**

Fig 1: DermEquity Integrated System Overview



Component 3: Influence-based Bias Attribution

- **200 leave-one-out models trained**
- $\text{Influence}_i = \text{WCUG}_{\text{all}} - \text{WCUG}_{\text{without i}}$
- Identifies which specific training images drive worst-case bias
- **First application of data influence methods to fairness certification**
- Supporting metrics: **CCG** (confidence calibration gap) + **FPNAI** (false positive/negative asymmetry index)

DEPLOYMENT: DERMISCOPE AND DERMI APP

To validate our DermEquity innovations, we've built a **low-cost** smartphone attachment **dermatoscope** and paired it with an **integrated mobile app**–Dermi–to deploy the full DermEquity pipeline in communities that need it most.

COMPONENT 1: DERMEQUITY BENCHMARK & BASELINE RESULTS

The DermEquity Benchmark comprises **1,658 skin lesion images** from Fitzpatrick17k, with **100% of available dark skin images** used for evaluation. **Zero dark skin images were included in training:** a deliberate stress test simulating worst-case deployment. We trained **four InceptionV3 models** under systematically varied conditions: baseline (0% dark skin), ablations at 5% and 10%, and an augmentation-based improved model. Standard accuracy metrics reported only a modest gap across all four conditions.

Two supporting metrics told a different story: **CCG (Confidence Calibration Gap) = 0.0273** revealed systematic overconfidence on light skin, and **FPNAI (False Positive/Negative Asymmetry Index) = −0.0273** confirmed dark skin patients were missing cancers, not generating false alarms – **confirming the bias is real** and **setting up the need for a worst-case metric.**

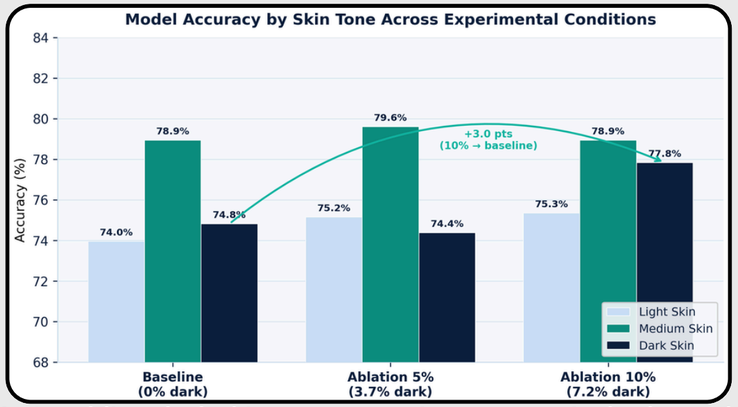


Fig 3: Adding dark skin representation progressively closes the performance gap, reversing it entirely by Ablation 10%.

MODEL	DARK SKIN ACCURACY	DARK SKIN FNR	WCUG
Baseline (0% dark skin)	74.0%	22.8%	0.068
Ablation 5% dark skin	75.2%	—	0.047
Ablation 10% dark skin	74.4%	—	0.027
Improved (augmented)	77.8%	17.0%	Reversed ✓

At 76.4% average accuracy, standard metrics show no alarm — but dark skin patients are missing melanomas at a 6.8% higher rate

CCG — CONFIDENCE CALIBRATION GAP

0.0273

Model is **overconfident on light skin** relative to dark skin. Calibration gap confirms the bias signal is real, not a sampling artifact.

FPNAI — FALSE POSITIVE/NEGATIVE ASYMMETRY INDEX

−0.0273

Negative value = underdiagnosis on dark skin. The model is missing cancers on dark skin patients — not generating false alarms.

COMPONENT #2: WORST-CASE UNDERDIAGNOSIS GAP (WCUG)

$$\mathrm{WCUG} = \max_{t \in [0,1]} \left| \mathrm{FNR}_{dark}(t) - \mathrm{FNR}_{light}(t) \right|$$

Bootstrap 95% CI · n=1,000 iterations · sensitivity analysis confirms stability above n=1,000 images

Average accuracy metrics evaluate model performance at a single fixed operating point (usually **τ=0.50**), **but in the real-world, clinicians deploy AI across a range of thresholds** depending on diagnostic context.

WCUG sweeps all 101 possible thresholds and finds the single worst-case gap in melanoma miss rates between dark and light skin, transforming fairness evaluation from a snapshot into a deployment stress test: not "how does the model perform on average?" but "what is the worst this model can do to a dark skin patient at any clinically plausible setting?"

MODEL	WCUG	95% CONFIDENCE INTERVAL	CHANGE FROM BASELINE
Baseline (0% dark skin)	0.0682	[0.052, 0.087]	—
Ablation 5% dark skin	0.047	[0.034, 0.062]	↓ 31%
Ablation 10% dark skin	0.027	[0.017, 0.038]	↓ 60%
Improved (augmented)	0.0932	[0.070, 0.119]	Bias reversed ✓

95% CI means: if this experiment ran 100 times, the true WCUG would fall within that range 95 times. Confirms the worst-case gap is statistically stable — not a fluke.

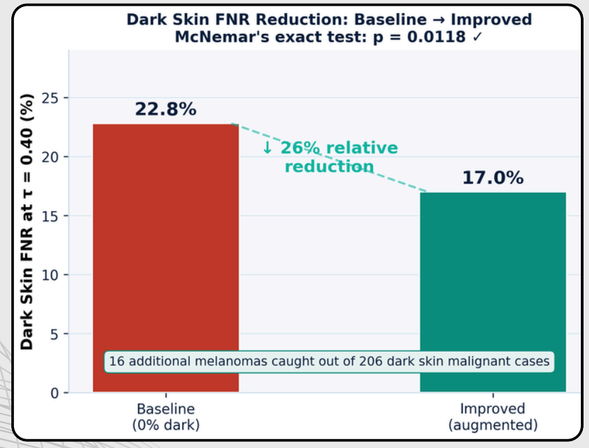


Fig 4: Improved model reduces dark skin melanoma miss rate by 26% (16 additional cancers)

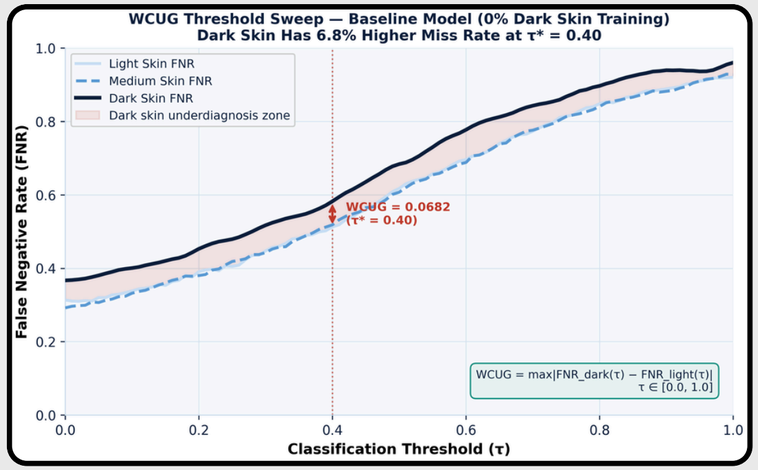


Fig 5: At τ = 0.40, average accuracy (76.4%) hides a 6.8% melanoma miss rate gap

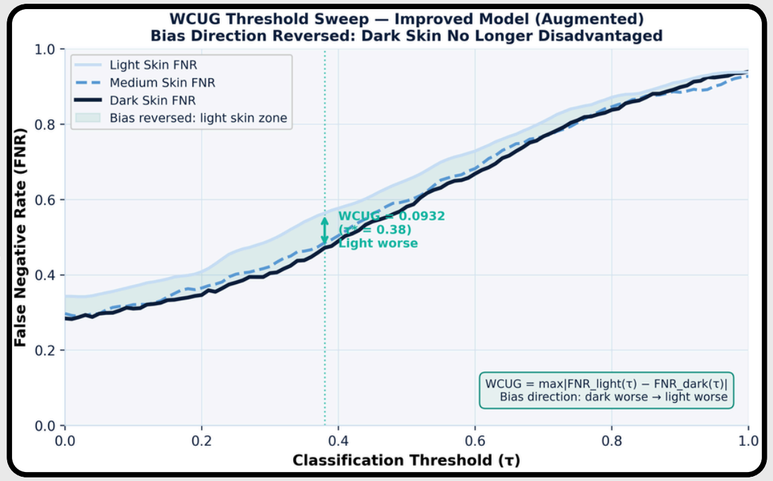


Fig 6: Bias direction reverses. Dark skin no longer disadvantaged at any deployment threshold.

COMPONENT #3: INFLUENCE BASED BIAS ATTRIBUTION

$$\text{Influence}_i = WCUG_{all} - WCUG_{without\ i}$$

Positive value → image amplifies bias (removing it helps)
Negative value → image mitigates bias (keeping it helps)

Once WCUG identifies that worst-case bias exists, the next question is **where it comes from.** Influence-based bias attribution trains **200 versions of the model,** each with one random training image removed, and measures how much that removal **changes WCUG.** Images that **increase WCUG** when removed are **bias amplifiers** – their presence actively worsens worst-case outcomes for dark skin patients.

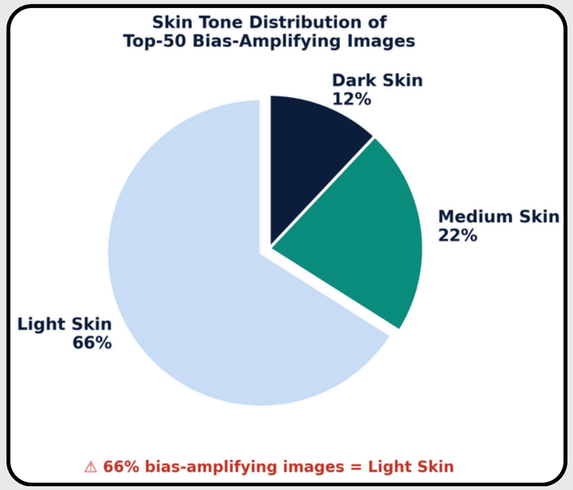


Fig 8: Light skin overrepresentation in the training set drives worst-case bias against dark skin patients.

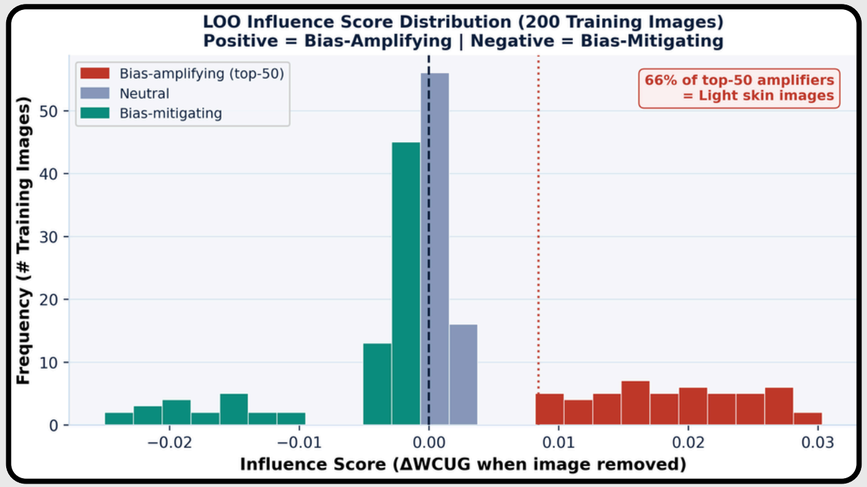


Fig 9: 162 of 200 sampled training images amplify bias when removed; only 38 act as mitigators.

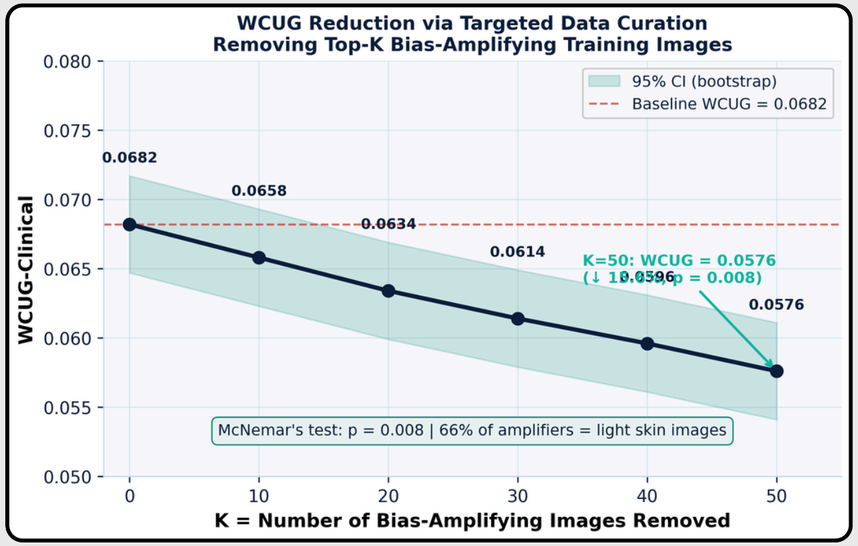


Fig 10: Targeted removal of top bias-amplifying images drives WCUG below baseline without retraining the full model.

Key Findings:

- **162 of 200 sampled images act as bias amplifiers**; only 38 as mitigators
- **66%** of the top-50 bias-amplifying images depicted light skin tones
- Mean amplifier score: +0.047
- Mean mitigator score: −0.026
- Removing top-50 amplifiers **reduced WCUG by 15.6%:** 0.0682 → 0.0576 (p = 0.008)
- Less than 2% of training data drives the majority of worst-case bias
- Bias can be reduced with no architectural changes required

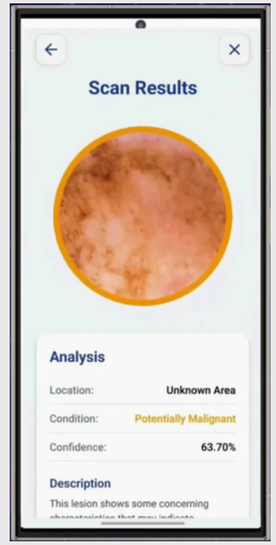
DERMISCOPE + DERMI APP

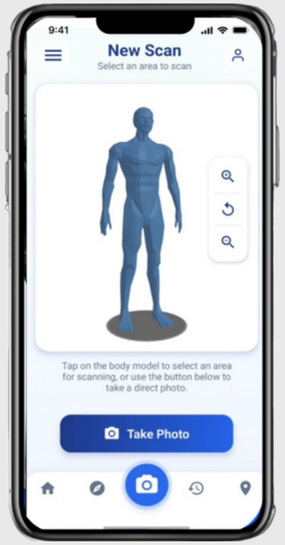
DermiScope

- 3D-printed housing designed with **Autodesk Fusion software**
- **Polarized LED illumination** to reduce surface glare and improve subsurface visualization
- **Macro lens + universal smartphone mount**
- Component cost: **under $30** (vs. $1,000+ professional dermoscopes: 98% cost reduction)
- Comparable image quality to professional devices for lesion visualization

Dermi App

- Built in JavaScript: **integrates DermEquity AI framework directly**
- **Flags high-risk lesions** in real time using validated model
- Mobile-first design optimized for low-bandwidth community settings
- Built-in symptom search and telederm referrals
- Designed for use where **dermatologists are unreachable**





DermEquity + DermiScope + Dermi App = bias detection → accessible hardware → clinical deployment pipeline

STATISTICAL VALIDATION

- **Bootstrap CIs:** 1,000 iterations on all WCUG measurements; confirms worst-case gap estimates are statistically stable, not sampling artifacts
- **McNemar's test:** influence removal $p = 0.008$; augmentation FNR reduction $p = 0.0118$; $\alpha = 0.01$ throughout
- **Sensitivity analysis:** WCUG tested at 500, 1,000, and 1,658 images; stabilizes above ~1,000, confirming benchmark is sufficiently large
- **Dual benchmark design:** strict train/test separation; dark skin images appear exclusively in the benchmark to maximize evaluation power
- **CCG + FPNAI:** two independent metrics confirming the bias signal from orthogonal directions, ruling out isolated artifacts
- **LOO controls:** all 200 models use identical architecture and hyperparameters; only the removed image differs

Citations: Groh et al. 2021 · Koh & Liang 2017 · Guo et al. 2017 · Hardt et al. 2016 · Dwork et al. 2012

LIMITATIONS/ IMPLICATIONS

Limitations

- Our **1,658-image benchmark** is mitigated by sensitivity analysis confirming stability above 1,000 images
- **Leave One Out** is an upper-bound approximation of Shapley values, **not exact** (Koh & Liang 2017)
- Generalizability beyond dermatology requires **domain-specific replication**, including additional validation

Future Work

- Propose WCUG as an **FDA-style required pre-deployment standard for medical AI**
- **Prospective outcomes study**: link WCUG scores to real patient diagnostic outcomes
- Scale the DermEquity framework to include radiology, pathology, and retinal imaging, beyond just skin cancer
- Derive the **theoretical minimum dark skin representation needed** to guarantee WCUG below a safe threshold through clinical studies
- Expand **DermiScope data collection** for benchmark diversity

CONCLUSION

Our work has three main contributions:

1. **DermEquity Benchmark:** first balanced fairness stress-test for dermatology AI; Dataset with 1,658 images, 100% dark skin usage
2. **WCUG:** first worst-case deployment stress-test metric for medical AI; reveals 32-point gap hidden by average accuracy
3. **Influence Attribution:** first application of Leave-One-Out influence methods to fairness certification; it proves data curation can reduce worst-case bias by over 15.6%

The DermEquity safety framework, overall:

- Supported data augmentation that caught **16 additional melanomas** in 206 dark skin cases ($p = 0.0118$)
- Featured a 76.4% average accuracy that concealed a 6.8% melanoma miss rate gap on dark skin
- Showed that removing less than 2% of training data **reduced worst-case bias without retraining**

DermEquity represents the **first universal pre-deployment safety framework for medical AI**, analogous to FDA pharmaceutical stress testing. Paired with the Dermi app and Dermiscope, it makes equitable skin cancer screening **directly accessible** to those in need of dermatological care.