



AI EQUITY AUDIT: PUBLIC SAMPLE

Statement of Findings, public edition · Independent evaluation on public data only · no client

Prepared 9 October 2026 · RR-STEM Methods v1.1 · Engine v0.5.0 (SHA 6fdbd691711982be) · Input SHA 2d0dc84fa544ec61

What was audited

An openly available seven-class skin-lesion classifier (Apache-2.0 licence), scored on **PAD-UFES-20** (Pacheco et al., 2020; Mendeley Data; CC BY 4.0): the 1,494 smartphone images that carry a clinician-recorded Fitzpatrick type, from 794 patients. Malignancy score = model probability of basal cell carcinoma, melanoma or actinic keratosis; reference standard = the dataset's recorded diagnosis (BCC, SCC, MEL counted as malignant). 1,485 records analysed after nine exact duplicates were collapsed under a rule fixed before the data was opened. Operating threshold derived at 90% pooled sensitivity (0.2702).

Primary verdict

The evaluation set contains too few positive cases in the darkest skin-tone band for any statement about that band, in either direction, to be supportable. Fitzpatrick V–VI holds 11 images and 5 malignant cases; the narrowest interval any sensitivity estimate there can have is about ± 31 percentage points. The ordered trend in detection across all tone levels does not show a statistically supported gradient (slope +0.08 per level on the logit scale, 95% CI -0.22 to $+0.37$).

Results by skin-tone band (patient-aware intervals; no claim below 30 positives)

Band	Images	Positives	Sensitivity [95% CI]	Specificity [95% CI]	AUROC	Precision ceiling
Fitzpatrick I–II	1,020	776	0.897 [0.869, 0.920]	0.287 [0.227, 0.355]	0.636	± 2.5 pp
Fitzpatrick III–IV	454	301	0.907 [0.868, 0.935]	0.366 [0.284, 0.457]	0.643	± 4.0 pp
Fitzpatrick V–VI	11	5	no claim	no claim	not reportable	± 31.3 pp

Secondary findings (descriptive; they do not alter the verdict)

- On this dataset the model's discrimination is weak in every band that can be measured (AUROC 0.64 in both), its probabilities are overconfident (expected calibration error 0.21–0.23; calibration slope 0.17–0.19, where 1.0 is ideal), and reaching 90% sensitivity requires flagging 63–71% of benign lesions. A single global temperature does not repair calibration across bands.
- 393 of 794 patients contributed more than one image; every interval is corrected for this (design effect 1.18 on detection).
- One skin-tone label source only, so label reliability could not be measured here. On the field's standard benchmark, two expert sources agree exactly on 47.9% of images.

Read this correctly

The model was developed on a different kind of image (dermoscopic) and population; PAD-UFES-20 is smartphone photography from Brazil. The figures above describe performance **on this evaluation set only**. They are not a characterisation of the model in its intended setting, not a clinical-performance claim, and not a regulatory determination. This is a statement of findings, not a certification, and confers no endorsement. The purpose of the sample is to show what an audit measures, how it states what the data can and cannot support, and that every number can be regenerated.

Method and reproduction

Pre-specified and published: RR-STEM Methods v1.1 (DOI 10.5281/zenodo.23149717); preprint DOI 10.21203/rs.3.rs-10670899/v1; preregistration OSF.IO/FMSNE. Two primary endpoints, Bonferroni-adjusted; all other findings descriptive; reported to TRIPOD+AI. The input file, results and a reproduction package regenerate all 336 reported values; a full 12-page report in the client format is available on request. Attribution: images and skin-tone labels from PAD-UFES-20 (Pacheco et al., 2020, CC BY 4.0).