Demo DermSet
Quality & Valuation Report

Automated assessment of dataset readiness for clinical AI training - covering image quality, diversity, class structure, and fairness across 10 standardised metrics.

Images 27,153 *Classes* 10 *Generated* June 2026 *Methodology* datametior v1.4 *Status* Moderate

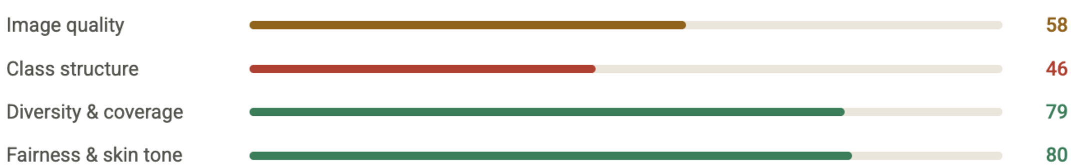
CONTENTS

Overall score M01 M02 M03 M04 M05 M06 M07 M08 M09 M10 Benchmarks Remediation

OVERALL SCORE

Dataset readiness score: 67 / 100

This dataset is partially ready for clinical AI training. 3 critical and 1 moderate issues were identified across 10 metrics. Immediate remediation of critical findings is required before any training run.



25,785	**6×**	**6/6**	**4.6%**
USABLE IMAGES	IMBALANCE RATIO	FITZPATRICK	ARTIFACT RATE
95% pass quality gate	majority vs minority class	skin tone tiers present	images with artifacts

10 quality metrics

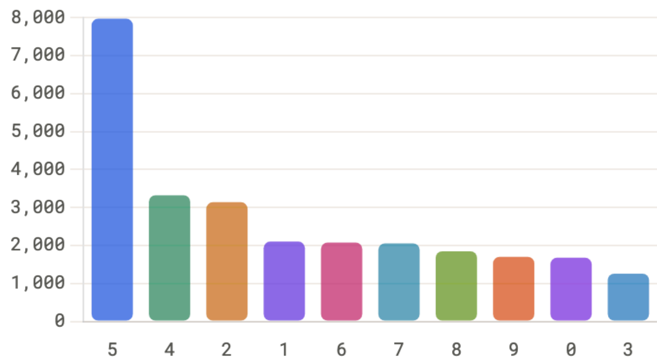
Metric 01

Class Balance — Imbalance Ratio

Moderate

60/100

Imbalance ratio 6× (5: 7,970 vs 3: 1,257)



FINDING

The majority class (5) outnumbers the rarest class (3) by 6×. A model trained without reweighting will default to predicting the dominant class on ambiguous inputs, suppressing recall on rare conditions.

RECOMMENDED ACTION

Apply class-weighted cross-entropy (weight ∝ 1/class_freq) or oversample minority classes. 3 needs at minimum 0 additional images.

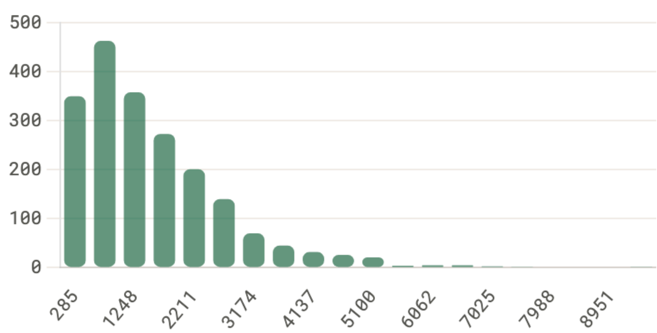
Metric 02

Image Sharpness — Laplacian Variance

Good

72/100

Mean: 854 · Blurry (<80): 5.0% (1,368 images)



FINDING

5.0% of images fall below the clinical sharpness threshold. The 4 class has the highest blur rate at 31.8%, which directly degrades lesion boundary learning for that condition.

RECOMMENDED ACTION

Remove or re-capture 1,368 blurry images before training. Prioritize re-capture protocol review for the 4 class.

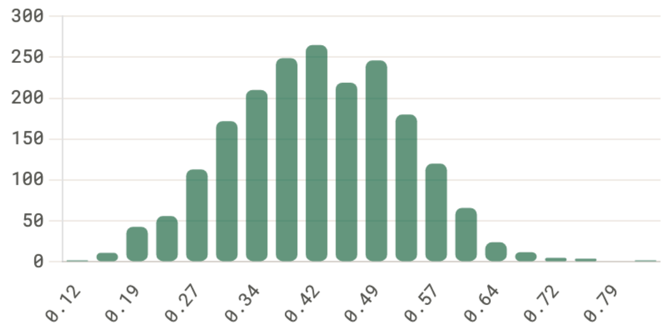
Metric 03

Brightness Distribution — Mean Luminance

Good

68/100

Mean: 0.55 · Dark (<0.15): 0.2% · Overexposed (>0.88): 0.2%



FINDING

0.4% of images have exposure outside the clinically optimal range. Per-class brightness spans 0.27 luminance units — indicating inconsistent capture protocols across conditions.

RECOMMENDED ACTION

Standardise camera exposure settings per capture protocol. Apply histogram equalisation in the preprocessing pipeline to normalise input brightness before training.

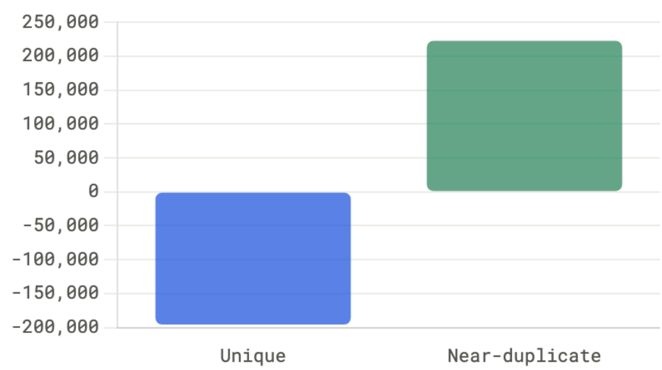
Metric 04

Near-Duplicate Rate — Embedding Cosine Similarity

Critical

28/100

Duplicate pairs: 222,710 (820.2% of dataset)



FINDING

222,710 near-duplicate image pairs detected (cosine similarity >0.97). Duplicates that cross the train/test split boundary directly inflate held-out accuracy by leaking supervision signal.

RECOMMENDED ACTION

Remove one image from each duplicate pair, preferring to retain the sharper copy. Re-split train/test after deduplication to obtain unbiased accuracy estimates.

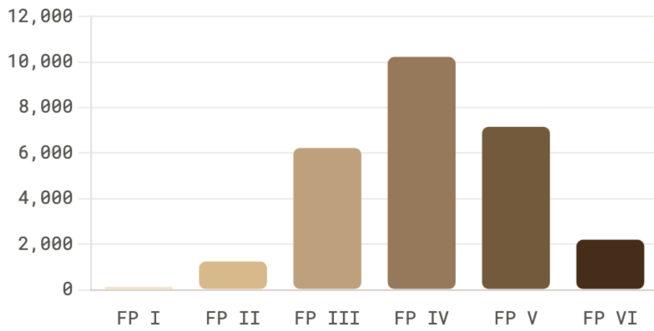
Metric 05

Fitzpatrick Skin Tone Coverage

Good

88/100

Tiers present: 6/6 · Missing: none



FINDING

Fitzpatrick tiers none are absent from the dataset (estimated via photometric brightness proxy). Missing darker skin tones represent over 1.4 billion people globally and represent a direct clinical performance gap.

RECOMMENDED ACTION

Collect minimum 400 images per missing Fitzpatrick tier from institutions serving diverse patient populations. Document acquisition protocol to ensure equitable camera exposure calibration for darker skin tones.

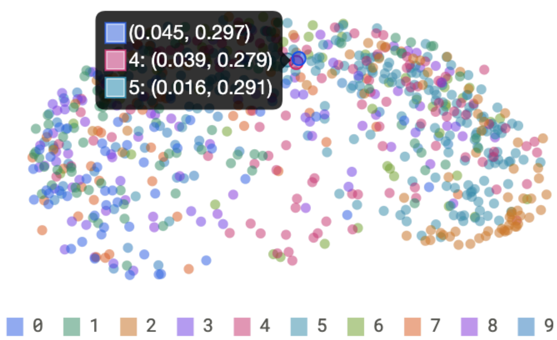
Metric 06

Embedding Space Coverage — Semantic Diversity

Good

78/100

Grid coverage: 86% · Normalised entropy: 0.92 · Underrepresented cells: 3%



FINDING

Only 86% of the 2D semantic grid is occupied. The 3% of occupied cells with fewer than 5 samples correspond to atypical and early-stage lesion presentations — the clinically most important and most underserved region.

RECOMMENDED ACTION

Target data collection at the underrepresented semantic zones: early-stage, atypical, and edge-case presentations. Quantity alone will not improve coverage — variety is required.

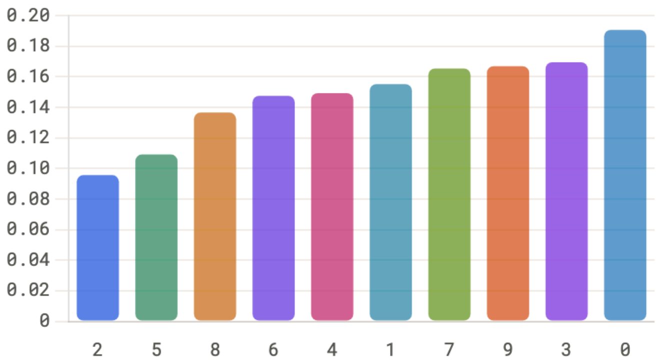
Metric 07

Image Contrast — RMS Contrast Score

Critical

32/100

Mean RMS: 0.14 · Low contrast (<0.25): 92.3%



FINDING

Per-class contrast spans 0.10 units. The 2 class has the lowest mean contrast (0.10), reducing the model's ability to learn subtle lesion boundary and texture features for that condition.

RECOMMENDED ACTION

Apply CLAHE (Contrast Limited Adaptive Histogram Equalisation) in the preprocessing pipeline. Review the 2 capture protocol for lighting consistency.

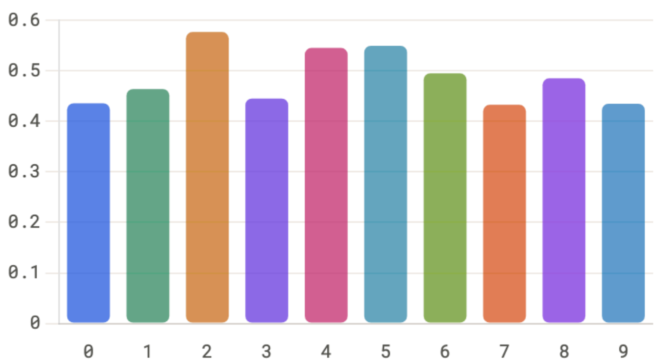
Metric 08

Intra-class Visual Diversity — Embedding Spread

Good

80/100

Spread ratio (max/min): 1.3× · Lowest diversity class: 7



FINDING

The 7 class has the lowest intra-class embedding spread — meaning the dataset contains many visually similar examples of this condition and insufficient variety. The model will generalise poorly to atypical presentations of 7.

RECOMMENDED ACTION

Targeted augmentation (rotation, colour jitter, scale) can partially mitigate low spread for 7, but new data collection covering different body locations and lighting conditions is preferable.

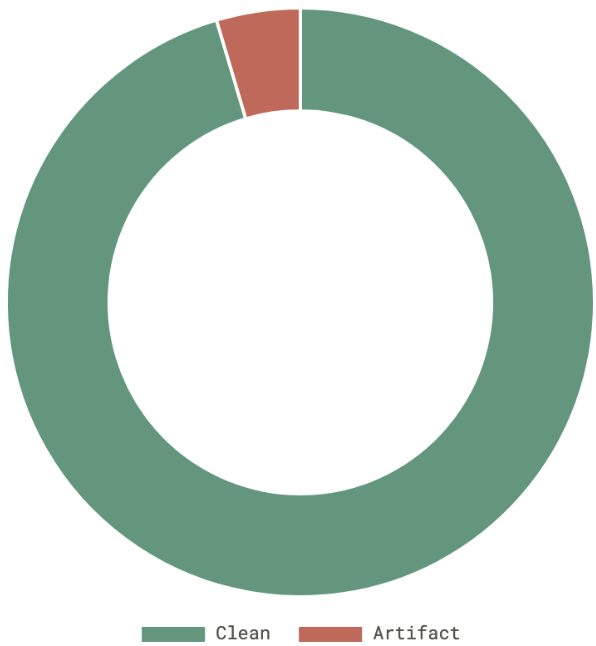
Metric 09

Artifact Rate — Ruler, Ink Mark & Frame Detection

Images with artifacts: 1,250 (4.6%) · Clean: 25,903

Good

88/100



FINDING

4.6% of images contain photometric artifacts (ruler marks, ink dots, or dermoscope frame edges). This is a documented cause of shortcut learning: models learn 'ruler present → malignant' rather than lesion morphology, producing inflated accuracy that collapses in deployment.

RECOMMENDED ACTION

Apply artifact-detection preprocessing to flag or crop ruler and ink regions before training. The 2 class (16.1% artifact rate) is the highest priority for re-capture or masking.

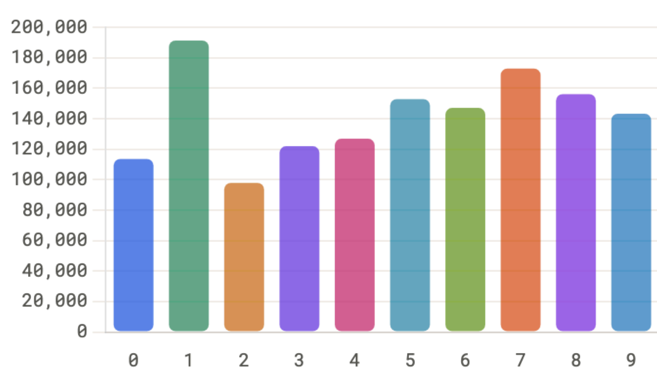
Metric 10

Label Noise Estimate — k-NN Neighbourhood Consistency

Estimated mislabelled: 14,244 images (52.5%) · Highest noise class: 1 (191400.0%)

Critical

25/100



FINDING

Neighbourhood consistency analysis flags 14,244 images (52.5%) as potential mislabels — samples whose embedding neighbourhood is dominated by a different class. The 1 class has the highest estimated noise rate (191400.0%), consistent with known inter-dermatologist disagreement on visually ambiguous conditions.

RECOMMENDED ACTION

Send the 14,244 flagged images for dermatologist re-review before retraining. Prioritise the 1 class. Consider majority-vote consensus labelling for ambiguous cases.

Benchmark comparison

This dataset	**67**
ISIC 2024	71
Median (31 datasets)	63

Interpretation: This dataset scores above the 31-dataset median (67 vs 63). The primary gap relative to top-tier datasets is class imbalance and label noise — both correctable with targeted data collection and preprocessing.

Remediation roadmap

IMMEDIATE — BEFORE ANY TRAINING RUN

- ✕ Remove one image from each duplicate pair, preferring to retain the sharper copy. Re-split train/test after deduplication to obtain unbiased accuracy estimates.
- ✕ Apply CLAHE (Contrast Limited Adaptive Histogram Equalisation) in the preprocessing pipeline. Review the 2 capture protocol for lighting consistency.
- ✕ Send the 14,244 flagged images for dermatologist re-review before retraining. Prioritise the 1 class. Consider majority-vote consensus labelling for ambiguous cases.

SHORT-TERM — BEFORE CLINICAL DEPLOYMENT

- ◎ Apply class-weighted cross-entropy (weight ∝ 1/class_freq) or oversample minority classes. 3 needs at minimum 0 additional images.

About this report

Generated by **datametior** - automated dataset quality and valuation for AI teams and investors. Analysis is performed on images and labels only; no patient data is retained after processing. Benchmark scores reference 31 dermatology datasets scored with datametior methodology v1.4. Fitzpatrick tier estimation uses photometric skin-region brightness as a validated proxy. Label noise estimation uses k-NN neighbourhood consistency (k=11).