This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

TRIDENT: Safety-Gated, On-Device Triage for Paediatric Teledermatology on Pigmented Skin

Submission #25, MIRASOL 2026

Oluwaferanmi Oladepo Nunsi Shiaki

This document accompanies the main manuscript with (S1) a completed TRIPOD+AI reporting checklist, (S2) the full numerical detail behind the JPEG degradation-disentanglement sweep summarised in Sec. 3.6, (S3) a note on what is and is not preserved from the Gaussian-blur generalisation check, and (S4) the on-device quantization ablation that motivated shipping FP16 rather than INT8. All numbers below are taken directly from experiment logs already computed for the main paper; nothing here is a new experiment.

S1. TRIPOD+AI Reporting Checklist

We report against the 27-item TRIPOD+AI checklist for studies developing or evaluating a multivariable prediction model (Collins et al., *BMJ*, 2024). Table S1 maps each item to where it is addressed in the main text, or states plainly where it is not reported. Several items are marked *deferred to source dataset* because TRIDENT trains and evaluates on the pre-existing, publicly described PASSION benchmark [PASSION2024] rather than newly collected patient data; we did not restate PASSION's own data-collection, consent, or ethics details, which are documented in that dataset's original publication.

Item	Topic	Status / location in main text
1	Title	Identifies model development/evaluation, target population (paediatric SSA), and predicted outcome (triage action).
2	Abstract	Covered; see Abstract.
3a	Background/rationale	Sec. 1, para. 1 (dermatologist scarcity; WhatsApp degradation).
3b	Target population / care pathway	Sec. 1, paras. 1–2 (CHW → WhatsApp → AI → retake/treat/refer).
3c	Known health inequalities	Partial — pigmented-skin focus and the referral-burden audit (Sec. 3.4) address a skin-tone/geography axis; not framed as background epidemiology.
4	Objectives	Contributions list, Sec. 1.
5a	Data sources	Sec. 2.1 (PASSION, single source for train/val/test via official split).
5b	Collection dates	Not reported here — deferred to source dataset [PASSION2024].
6a	Setting / number of centres	Partial — four countries named; centre-level counts not given.
6b	Eligibility criteria	Not reported here — deferred to source dataset.
6c	Treatments received	Not applicable — single-image triage task, not a treatment-outcome study.
7	Data preprocessing / QC	Partial — Secs. 2.2–2.3 describe the degradation and augmentation pipeline; not stratified by subgroup.
8a	Outcome definition	Sec. 2.1: four disease classes plus binary impetiginization; diagnostic (single image), no time horizon.

Item	Topic	Status / location in main text
8b	Assessor qualifications	Not reported here — deferred to source dataset.
8c	Outcome-assessment blinding	Not reported here — deferred to source dataset.
9a	Predictor pre-selection	Not applicable — end-to-end learning from pixels, no hand-selected predictors.
9b	Predictor definition/measurement	Partial — consumer-device image acquisition described generally (Sec. 1); no device-level detail.
9c	Subjective predictor measurement	Not applicable.
10	Sample-size justification	Not reported — the full available PASSION split is used; no prospective power calculation.
11	Missing data	Not reported.
12a	Data partition	Sec. 2.1 (official split plus stratified 15% validation slice).
12b	Predictor handling	Sec. 2.3 (ImageNet normalisation, augmentation policy).
12c	Model, hyperparameters, internal validation	Secs. 2.3–2.4 (architecture, optimiser, schedule, early stopping on val macro-F1).
12d	Heterogeneity across runs	Three-seed ensemble; val macro-F1 reported with s.d. (Sec. 3.1).
12e	Performance measures/plots	F1, balanced accuracy, ECE, Brier, AUROC, risk–coverage curve (Sec. 3, Fig. 2).
12f	Model updating/recalibration	Temperature scaling, Sec. 2.4.
12g	How predictions are calculated	Triage rule given as Eq. 1; network is the cited EfficientNet-B0 plus an ONNX export (Sec. 4); full training code not linked in the manuscript.
13	Class imbalance	Sec. 2.4 (inverse-frequency class weights; focal loss for impetiginization).
14	Fairness approaches	Partial — the referral-burden audit (Sec. 3.4) is explicitly framed as deployment monitoring, not a fairness intervention (stated in Contributions).
15	Model output	Eq. 1: discrete {retake, treat, refer} action plus calibrated probabilities.
16	Dev. vs. evaluation data differences	Partial — same source distribution; the Tanzania fold is flagged as underpowered (Limitation i).
17	Ethical approval	Not restated in this manuscript — deferred to source dataset's own approval [PASSION2024].
18a	Funding	Acknowledgements.
18b	Conflicts of interest	Acknowledgements ("we declare no competing interests").
18c	Protocol	Not applicable — no pre-registered protocol.
18d	Registration	Not applicable — not a registered study.
18e	Data sharing	Partial — PASSION is a public benchmark [PASSION2024]; no new patient data released by this work.
18f	Code sharing	Not stated in the manuscript.
19	Patient/public involvement	Not reported.
20a	Participant flow	Partial — test-set composition given (Sec. 3.1: 986 images, 329 subjects, 198 impetiginized); no flow diagram.
20b	Characteristics by source	Partial — country-level referral rates (Sec. 3.4); no full demographic table.
20c	Predictor-distribution comparison	Not reported.
21	N per analysis	Reported per experiment (e.g. 986/329 for the main test set; 300-image subset for the disentanglement sweeps, S2 below).
22	Full model specification	Partial — architecture and triage rule specified; full weights/training code not linked in the manuscript.

Item	Topic	Status / location in main text
23a	Performance with confidence intervals	Partial — bootstrap 95% CIs given for impetiginization AUROC/F1 (Sec. 3.3); not for every headline metric.
23b	Cluster (country) heterogeneity	Sec. 3.4 (referral-burden audit).
24	Model-updating results	Not applicable — no post-hoc updating beyond the calibration already covered in 12f.
25	Interpretation incl. fairness	Partial — Secs. 3.6 and 4 interpret results; fairness framing is limited to the audit's explicit caveats.
26	Limitations	Limitations (i)–(v), Sec. 3.
27a	Handling poor/unavailable input	The quality head and retake action are the mechanism for this (Secs. 2.4, 3.2).
27b	User-interaction requirements	Partial — CHW workflow described qualitatively (Sec. 4); no formal interaction specification.
27c	Future research / applicability	Outlook (Sec. 4) and Limitation (v).

S2. Extended Detail: JPEG Degradation-Disentanglement Sweep

Sec. 3.6 of the main text summarises a controlled sweep on a 300-image test subset (three-seed ensemble) used to separate what the quality head $\hat{q}$ and the uncertainty signal u each respond to. This is a smaller, separate subset from the full-test-set sweep in Table 2 of the main text, so the exact mean $\hat{q}$ values differ slightly between the two; both show the same monotonic trend.

Table 2: Mean $\hat{q}$ and mean u by JPEG quality factor, 300-image disentanglement subset (Sec. 3.6).

QF	100 (orig.)	90	70	50	30	20	10
Mean $\hat{q}$	0.054	0.054	0.069	0.098	0.238	0.456	0.663
Mean u	0.256	0.255	0.254	0.256	0.256	0.264	0.286

Table 3: Correlation with QF and with each other, full precision (rounded in main text to $r = -0.74$, -0.04, and 0.08 respectively).

Pair	Pearson r	p	Spearman ρ	p
$\hat{q}$ vs. QF	-0.737	< 0.001	-0.750	< 0.001
u vs. QF	-0.040	0.067	-0.046	0.035
$\hat{q}$ vs. u (pooled across QF)	0.080	< 0.001	-0.014	0.512

Orthogonality also holds on undegraded data. A finding not reported in the main text: on the *original*, undegraded validation and test images, $\hat{q}$ and u show essentially no linear relationship (val Pearson $r = 0.010$, $p = 0.804$; test Pearson $r = -0.012$, $p = 0.696$), though Spearman's ρ detects a small but statistically significant monotonic association on both splits (val $\rho = -0.124$, $p = 0.002$; test $\rho = -0.198$, $p < 0.001$). Given the sample sizes involved, the significant p-values here reflect a real but practically small monotonic trend, not a strong dependency — consistent with the main text's claim that the two signals are near-orthogonal, and suggesting that trend is not an artefact of the synthetic JPEG sweep specifically.

S3. Gaussian-Blur Generalisation Check: What Is and Is Not Preserved

Sec. 3.6 also reports a Gaussian-blur sweep (radius $\in \{1, 2, 4, 6, 8, 12\}$px) showing that $\hat{q}$ does not track blur severity ($r = 0.11$, versus $r = -0.74$ for JPEG; flipped-vs.-non-flipped $p \geq 0.26$ at every radius tested), while u's elevation on flipped images transfers to this unseen corruption family ($p = 0.009$ at radius 2, $p < 0.0001$ by radius 6). Those two anchor statistics are exactly as reported in the main text. The full per-radius breakdown (mean $\hat{q}/u$ and flip counts at each of the six radii) was computed on the HPC cluster via `scripts/hpc/eval_blur_diagnostic.py` but was not saved back to this repository as a machine-readable artifact, so we do not reconstruct a fuller table here rather than approximate it. The script is included in the code release and reproduces the full sweep from the released checkpoints.

S4. On-Device Quantization Ablation

Sec. 4 of the main text reports the deployed FP16 ONNX model at 8.1 MB, 73 ms median / 76 ms p95 latency on a Samsung Galaxy A34. Table S4 gives the ablation that motivated FP16 over INT8: static INT8 quantization of the same (single, seed-42, pre-ensemble) checkpoint collapses macro-F1 from 0.678 to 0.414 and worsens calibration (ECE 0.099 $\rightarrow$ 0.170), which we attribute to EfficientNet-B0's heterogeneous squeeze-and-excitation block activation ranges being poorly served by a single fixed INT8 quantization scale. FP16 retains accuracy close to FP32 at roughly half the model size, which is why it is the version benchmarked on-device in the main text.

Table 4: Precision-format ablation, single seed-42 checkpoint (pre-ensemble). Latency is single-thread CPU on a development machine via `onnxruntime` — not directly comparable to the on-device ARM (Galaxy A34) benchmark reported in the main text for the deployed FP16 model.

Precision	Macro-F1	ECE	ms/img (dev. CPU)
FP32	0.678	0.099	150.3
INT8 (static)	0.414	0.170	130.0