

Static INT8 Quantization of a Conjunctival Anaemia Classifier on Android

Enrique Lee Huamani¹, Erasmo Montufar-Barrientos², Victor Romero-Alva³, Vania Yulisa Abanto Castro⁴
Image Processing Research Laboratory (INTI-Lab), Universidad de Ciencias y Humanidades, Lima, Peru^{1,2,3}
e-Health Research Center, Universidad de Ciencias y Humanidades, Lima, Peru⁴

Abstract—This study examines the engineering feasibility and evidential limits of static INT8 post-training quantization for a conjunctival-image anaemia classifier on a mid-range Android phone. Using 202 publicly available subject images, we audited a subject-disjoint partition and trained ResNet-50 on automatically extracted conjunctival regions of interest. On a single held-out set of 30 subjects, FP32 and static INT8 models produced AUCs of 0.8688 and 0.8710, respectively; the static INT8 AUC confidence interval was 0.710–0.984. These estimates do not establish accuracy equivalence. Quantization reduced file size from 90.01 to 22.92 MB (74.5%). An exploratory CPU-only Android benchmark measured 89.31 ms per forward pass after the phone cooled, compared with two earlier static-model sessions of 120.51 and 110.53 ms. The reported timings used synthetic tensors, excluded image acquisition and preprocessing, and did not interleave model variants under controlled thermal conditions. At a threshold of 0.5, the classifier missed four of 13 anaemic test subjects. The findings support model compression and proof-of-concept execution, but neither clinical screening suitability nor a reproducible end-to-end speed advantage has been established.

Keywords—Post-training quantization; anaemia; conjunctival images; Android; inference latency; patient-level evaluation; deployment limitations

I. INTRODUCTION

Anaemia is one of the most prevalent health conditions worldwide: the Global Burden of Disease Study 2021 estimated a prevalence of 24.3% across all ages, or 1.92 billion cases [1]. Its diagnosis requires measuring the haemoglobin (Hb) concentration against the World Health Organization cut-offs [2], which needs a blood sample and equipment often unavailable in primary care and community settings. Clinical examination of pallor is non-invasive, but its accuracy and reliability are limited [3]. These constraints have motivated image-based approaches using smartphone photographs of the fingernails [4] or of the palpebral conjunctiva [5]–[7].

Studies of conjunctival-image classification report promising retrospective discrimination [8]–[11], but model discrimination, file size and on-device processing are distinct outcomes. Deployment also requires a reliable camera-to-tensor pipeline, device-level resource measurements and an operating threshold appropriate to the consequences of false negatives.

- Small medical-imaging datasets are particularly vulnerable to subject overlap and source confounding, both of which can inflate estimates of generalization [12], [13], [27].

The primary research question is: for a fixed ResNet-50 classifier evaluated on subject-disjoint public images, how do static INT8 quantization, model size, held-out discrimination and exploratory CPU forward-pass time compare with FP32? This is a technical feasibility study, not a clinical validation of a screening application. The specific contributions are:

- an independent subject-leakage audit of the public dataset and an account of a corrected training–inference input mismatch;
- a reproducible description of automatic ROI preprocessing and of what the Android benchmark did and did not execute;
- a descriptive same-subject comparison of FP32, dynamic INT8 and static INT8 model output summaries, without an equivalence claim;
- an explicit analysis of limitations in threshold transfer, model selection, source generalization and thermal-state-dependent latency.

Section II reviews related work, Section III describes the methods, Section IV reports the results, Section V discusses them and Section VI concludes.

II. RELATED WORK

A. Non-Invasive Anaemia Detection from Conjunctival Images

Digital photographs of the palpebral conjunctiva contain information related to haemoglobin concentration [5], and controlled smartphone attachments have been used to standardize acquisition [7]. Prior conjunctival and smartphone studies include colour-based models [6], [9], the Eyes-defy-anemia dataset [8], [14], and a vision-transformer approach [11]. The closest quantization study, by Cruz Romero and Lugo Beauchamp [15], used a MobileNet backbone on the paediatric CP-AnemiC dataset, applied cross-validation and considered FP32, FP16, INT8 and INT4. Its reported deterioration at lower precision is not directly comparable with our result because populations, datasets, architecture and evaluation settings differ. Our narrower contribution is an audited split, an input-consistency investigation and a physical Android CPU forward-pass measurement; a matched MobileNet baseline is still required for an architecture-level deployment claim.

B. On-Device Deep Learning, Quantization and Evaluation Pitfalls

Deep convolutional neural networks have also been investigated for clinical-image classification beyond conjunctival photographs, including skin-lesion classification [16]. This evidence illustrates the broader medical-imaging context but does not validate the present anaemia classifier.

Mobile hardware and runtimes affect inference time, and a target-device evaluation is important [17]. Post-training quantization reduces numerical precision and model storage, but hardware speedups depend on operator coverage, graph format, runtime configuration and thermal behavior [23]–[26]. The present benchmark uses a single ResNet-50 backbone, one CPU provider and one phone; it does not show that ResNet-50 is preferable to mobile-native networks or that the full application is faster. A matched MobileNetV3 or EfficientNet-Lite benchmark remains an important comparison.

Other mobile and edge-learning applications include multi-sensor agricultural diagnosis with a CNN–vision-transformer model [18], construction-site concrete crack assessment [19], on-device object training and recognition [20], and CNN-based butterfly-image characterization on mobile devices [21]. Privacy-preserving frameworks for cloud-connected mobile healthcare have also been investigated [22]. These studies concern different populations, images, hardware or privacy objectives; their results should not be treated as direct performance baselines for the conjunctival INT8 classifier examined here.

Leakage between training and evaluation data is a widespread cause of over-optimistic results in machine-learning-based science [12], and duplicated data and flawed dataset composition were among the most common pitfalls in a review of COVID-19 imaging models [13]. Models may also exploit source-specific image characteristics instead of the intended pathology, as shown for pneumonia detection across hospitals [27]. Both issues apply here: the dataset has several images per subject and two sources with very different anaemia prevalence.

III. METHODS

A. Dataset, Labels and Region-of-Interest Extraction

Source. Images come from the public Eyes-defy-anaemia dataset [8], [14] (CC BY 4.0 licence): 218 adult subjects (age 19–88 years) from Italy (123) and India (95), each with a photograph of the everted lower eyelid, hand-segmented conjunctiva images (palpebral, and palpebral plus forniceal) and the laboratory Hb value, age and sex. According to the dataset documentation, images were captured with a smartphone fitted with an attachment that magnifies the image and provides uniform LED illumination while excluding ambient light [7]. Subjects were labelled with the WHO adult thresholds [2]: anaemic if Hb < 13.0 g/dL (men) or < 12.0 g/dL (women).

Image sets. The full-eye set contains one original eye photograph for each subject with a valid Hb value and a usable photograph: 202 subjects (India 95, Italy 107; 90 anaemic, 112 non-anaemic), after excluding 16 Italian subjects without a valid photograph. The dataset's 427 hand-segmented crops (typically two per subject, full frame with white background) were used only to define and audit the partition and in the preprocessing-consistency analysis; the 396 crops linked to a

full-eye subject carry the same Hb-derived label as that subject in all cases.

Automatic ROI extraction. An offline image-processing procedure creates the model input from each original photograph. A color mask selects conjunctival tissue (CIELAB a^* > 152; HSV saturation > 40; HSV value between 50 and 235), followed by morphological cleaning and selection of the largest connected component. Pixels outside the component are set to black; a bounding box is expanded by 20%, cropped, resized to 224×224 and normalized using ImageNet statistics. A MediaPipe eye-landmark fallback [28] was specified but was never used, because the color-based procedure returned an ROI for all 202 controlled-light photographs. The source images used a magnifying smartphone attachment and standardized LED illumination [7]. Successful extraction in these conditions is not evidence of robustness to ambient light, other cameras, white balance, pigmentation, or skin-tone variation. In the Android speed test, preprocessing was not run in the measured path (Algorithm 1).

Algorithm 1. Consolidated image-to-prediction sequence and measured boundary

Input: original subject photograph (decoded RGB image); fixed ResNet-50 checkpoint; ROI thresholds above.

Output: anaemia probability and an exploratory binary flag. Dependencies: image decoding and color/morphology operations, ImageNet normalization, ONNX Runtime for inference.

1. Decode photograph; derive CIELAB and HSV representations for tissue masking. 2. Apply the fixed thresholds, morphological cleanup and largest-component selection; if no component is found, invoke the specified landmark fallback.

3. Mask non-ROI pixels, expand the bounding box by 20%, crop, resize to 224×224 , convert to float32, normalize by ImageNet channel statistics and arrange channels first as [1, 3, 224, 224].

4. Pass the tensor through the FP32 or quantized ONNX model; convert output scores to the anaemia-class probability using the evaluation softmax rule. 5. Apply threshold 0.5 only for the reported descriptive confusion matrix; do not interpret the flag as a validated clinical decision.

Measured boundary: the reported Android timings began after tensor creation and timed only step 4. They used a fixed-seed synthetic float32 tensor rather than a decoded patient photograph. Image capture, decoding, ROI extraction, normalization, session initialization and display were excluded.

B. Patient-Level Partition and Leakage Audit

Subjects were stratified by anaemia label and assigned to training, validation and test partitions using seed 42, with a subject-level identity audit to verify no overlap: 141/31/30 subjects and 301/64/62 segmented crops (Table I). An exact join on original-image stems linked crops to subjects, and the earlier file-level split was separately audited. Training and validation were used for optimization and checkpoint decisions; the nominal test set contained 13 anaemic and 17 non-anaemic subjects. However, the initial backbone-selection comparison had considered the full 202-image dataset, including these test subjects. Therefore, although this partition is subject-disjoint for model fitting, it is not a fully untouched model-selection test, and results require independent validation.

TABLE I. PATIENT-LEVEL PARTITION USED IN ALL EXPERIMENTS

Split	Subjects (full-eye images)	Anaemic	Non-anaemic	Hand-segmented crops
Train	141	63	78	301
Validation	31	14	17	64
Test (held-out)	30	13	17	62
Total	202	90	112	427

C. Model Architectures and Training

Three ImageNet-pretrained torchvision backbones were compared: EfficientNet-B0 [29], ResNet-34 and ResNet-50 [30], each with its classifier replaced by dropout ($p = 0.3$) and a two-output fully connected layer (softmax probability of the anaemic class). All layers were fine-tuned with cross-entropy, Adam (learning rate 1×10^{-4} , weight decay 1×10^{-4}) and batch size 16; the learning rate was halved after 3 epochs without validation-loss improvement, and training stopped after 5 epochs without validation-accuracy improvement (maximum 30), keeping the best epoch (8, 9 and 7 epochs were run, respectively). Augmentation comprised random 224×224 crops, flips, rotations up to $\pm 15^\circ$ and color jitter. PyTorch seed 42 and an NVIDIA RTX 4060 were used [31]. ResNet-50 was chosen because its AUC was highest in an exploratory comparison over all 202 subjects, which included the later test participants. This selection cannot be presented as prospectively specified or independent of the test data. The held-out comparisons below are descriptive only, and reselecting a backbone from them would introduce further bias. A model-selection protocol based solely on inner training/validation folds and an external test or nested subject-level cross-validation is required for an unbiased selection estimate.

D. Post-Training Quantization and Threshold Selection

The ResNet-50 checkpoint was exported to ONNX (opset 13) and pre-processed with the ONNX Runtime quantization tools [26]; export fidelity was checked as the maximum difference in probability against PyTorch. Dynamic INT8 quantization stores 8-bit weights and computes activation parameters at inference time. Static INT8 quantization [23], [24] uses per-channel signed 8-bit weights, unsigned 8-bit activations and the QDQ format, with activation ranges calibrated (MinMax) on 128 ROIs sampled from the training subjects only. PyTorch FP32 was evaluated in PyTorch, whereas ONNX FP32, dynamic INT8 and static INT8 were evaluated with ONNX Runtime 1.23.2 on the same inputs; the cost of quantization is the change in held-out AUC and confusion matrix at 0.5.

Threshold analysis used 101 values in $[0, 1]$ and explored maximum F1, maximum Youden J, maximum accuracy and a sensitivity-prioritized criterion on the 172-subject development collection. Because this collection includes 141 training subjects, its threshold estimates are optimistically biased; the exercise was not prospectively registered. For any future clinical-screening investigation, we propose a minimum sensitivity target of 90% to prioritize avoiding missed anaemia, with acceptable specificity and follow-up burden to be determined with clinicians. This target is a proposed future acceptance criterion, not an achieved or validated operating point. The threshold 0.5 is retained only as a descriptive reference for the existing prototype.

E. On-Device Latency Benchmark

Exploratory forward-pass timing used a Xiaomi POCO X5 5G (22111317PG; Android 14, arm64-v8a), ONNX Runtime Mobile 1.18.0, CPU execution provider, batch size 1 and one intra-op thread [26]. The Kotlin benchmark replaced only the bundled ONNX model; it passed an identically generated fixed-seed synthetic float32 tensor of shape $[1, 3, 224, 224]$. After 15 warm-up passes, 100 model invocations were timed. The FP32 and dynamic INT8 variants came from earlier model weights and were measured in separate device sessions, without concurrent recording of thermal state. Consequently these sessions were not a randomized, matched comparison and cannot establish a causal speedup between variants.

For static INT8, two preliminary sessions produced means of 120.51 and 110.53 ms when battery temperature was 39–40 °C and the observed large-core frequency was 691 MHz. After approximately two minutes of cooling, at 37 °C and 2054 MHz, the third session averaged 89.31 ms. This last figure is the lowest-temperature observed session, not a sustained or thermal-controlled deployment measurement. No photograph was decoded and no capture, ROI processing, initialization, memory, energy or accelerator measurements were made in the timed section. A definitive experiment should interleave all variants after retraining (for example, randomized ABBA blocks), log battery and CPU temperatures/frequencies, and measure per-image preprocessing, model initialization, total capture-to-result time, peak memory and sustained consumption on real images.

F. Statistical Analysis

Metrics used the anaemic class as positive. AUC confidence intervals on the 30-subject test set were computed with 2,000 subject-level bootstrap resamples; Wilson intervals describe sensitivity and specificity [32]–[34]. Overlap of marginal confidence intervals is not a formal paired comparison, and a small estimated AUC change does not establish equivalence. Per-subject FP32 and INT8 prediction scores are not reported in the present results, so a paired AUC-difference interval and a prediction-agreement statistic are not available.

The required paired analysis should resample test subjects (preserving their labels and paired probabilities), report the 95% bootstrap interval for $AUC(INT8) - AUC(FP32)$, mean and maximum absolute score deviation, Spearman rank correlation, threshold-crossing proportion and exact prediction agreement at 0.5. At least one larger external test set or repeated, preferably nested, subject-level cross-validation is needed to assess precision and selection stability. A prespecified clinically acceptable AUC-loss margin would be necessary before testing non-inferiority or equivalence. The five-seed preprocessing

comparison and per-source estimates below remain descriptive and cannot substitute for these analyses.

G. Ethics and Data Availability

Ethics. We analyzed a previously published, de-identified public dataset [14]; no participants were newly recruited, no identifying source records were requested and no attempt was made to re-identify subjects. The original dataset documentation describes the collection conditions and ethics basis. Data and code availability. Source photographs and dataset descriptions are accessible through the cited public repository [14]. The analysis scripts, trained checkpoints and subject-level paired predictions have not been deposited in a public versioned archive and are presently available only by request to the corresponding author. A versioned public release with de-identified subject keys, partition lists, runtime scripts and model checksums is needed to support independent reproduction, subject to dataset license and applicable privacy restrictions.

IV. RESULTS

A. Dataset and Leak-Free Split Verification

Class balance differed markedly between sources: 68 of 95 Indian images (71.6%) were anaemic, against 22 of 107 Italian images (20.6%). Every split contained both sources and both classes (Indian/Italian subjects: 65/76 training, 19/12 validation, 11/19 test).

The exact join linked 396 of the 427 crops to a subject, and the audit found zero subjects shared between any pair of splits. The earlier file-level partition, by contrast, shared 24 subjects between training and validation, 28 between training and test and 5 between validation and test, and 28 of the 30 test subjects (93.3%) had crops in its training set; its test figures (ResNet-50: accuracy 83.33%, AUC 0.9095) are not held-out estimates and are not used further.

B. Architecture Comparison

The held-out AUC point estimates of the three backbones ranged from 0.8688 to 0.8824 (Table II), but their broad uncertainty intervals do not support a ranking. At $n = 30$, changing one anaemic classification changes sensitivity by 7.69 percentage points and changing one non-anaemic classification changes specificity by 5.88 points. Because ResNet-50 was selected using a 202-image comparison that included eventual test subjects, this table does not provide an unbiased confirmatory model-selection comparison. The all-202 AUC of 0.9034 is in-sample/exploratory and is not a held-out estimate.

C. Consistency of Preprocessing Between Training and Evaluation

An earlier version of the classifier had been trained on hand-segmented crops (white background, full frame) but evaluated on automatic ROIs (black background, cropped). For the 29 test subjects with both inputs, that model reached an AUC of 0.7308 on automatic ROIs and 0.8750 on hand-segmented crops (paired difference +0.144, 95% CI -0.078 to $+0.381$), and the automatic input inflated its mean probability for non-anaemic subjects (0.552 vs. 0.319), giving a specificity of 35.29% at 0.5.

Retraining on automatic ROIs removed this mismatch (Table III).

Training on the deployment input raised the mean held-out AUC by 0.087, cut its seed-to-seed standard deviation by a factor of three and gave an AUC above 0.83 with every seed; the pre-specified seed-42 run (0.8688) lies close to the mean. Retraining the crop-trained configuration with the same seed gave a different AUC (0.8054 vs. 0.7647), so training is not bit-for-bit deterministic and the five-seed summary is the more reliable comparison.

D. Image Source as a Potential Confound

Source alone (India = 1) discriminated anaemia labels with AUC 0.7572 on all 202 images and 0.6516 on the nominal test set, reflecting strong country/label imbalance. The model AUCs of 0.7857 for Indian subjects and 0.8718 for Italian subjects in Table IV derive from only 11 and 19 test subjects, respectively, and should not be interpreted as evidence of transportability. Differences in the predicted probabilities of non-anaemic subjects across sources may represent camera, illumination, label-prevalence or other domain effects that this study cannot disentangle.

E. Quantization Impact

The ONNX FP32 graph matched PyTorch closely in the original export check (maximum reported probability difference 0.0013). Static INT8 quantization reduced file size from 90.01 to 22.92 MB (74.5%). On the nominal 30-subject test set, AUC point estimates were 0.8688 for FP32 and 0.8710 for static INT8 (reported Δ AUC +0.0023 based on unrounded values; Table V). The static model AUC interval (0.710–0.984) is too broad to establish non-inferiority or equivalence, and the paired difference interval was not available. Both variants had the same aggregate confusion-matrix counts at 0.5 (TN 16, FP 1, FN 4, TP 9); equality of counts does not prove agreement on each patient.

Dynamic INT8 yielded an AUC point estimate of 0.8778 and a different aggregate false-positive count, but no paired significance test was performed. The all-202 model comparisons include fitting and model-selection subjects and must not be interpreted as independent validation. Per-subject probabilities are required to quantify score drift, rank preservation, prediction switching and paired confidence intervals.

F. Threshold Selection

Development-sample threshold exploration selected 0.37 (F1 and Youden J), 0.39 (accuracy) and 0.31 (the exploratory high-sensitivity criterion). On the nominal test set, 0.37 increased false positives from one to four without recovering any of the four false negatives (Table VI). Threshold 0.31 recovered one anaemic case, leaving three of 13 undetected (sensitivity 76.92%) while reducing specificity to 76.47%. The sensitivity objective of at least 90% was therefore not met on the test subjects.

No threshold in this small evaluation is validated for clinical screening. The threshold of 0.5 is only the implementation default used to make model comparisons repeatable; it must not be described as a clinically selected screening threshold.

TABLE II. ARCHITECTURE COMPARISON ON THE HELD-OUT TEST SUBSET

Model	Acc.	Sens.	Spec.	Prec.	F1	AUC-ROC [95% CI]	TN	FP	FN	TP
EfficientNet-B0	80.00	76.92	82.35	76.92	0.7692	0.8824 [0.751–0.982]	14	3	3	10
ResNet-34	73.33	84.62	64.71	64.71	0.7333	0.8733 [0.727–0.974]	11	6	2	11
ResNet-50	83.33	69.23	94.12	90.00	0.7826	0.8688 [0.701–0.982]	16	1	4	9

n = 30, threshold 0.5, FP32; percentages in %.

TABLE III. RESNET-50 (FP32) TRAINED ON HAND-SEGMENTED CROPS VS. AUTOMATIC ROIS

Training input	AUC-ROC, mean $\pm$ SD (range)	Accuracy at 0.5 (%), mean $\pm$ SD	AUC within India (n = 11), mean (range)	AUC within Italy (n = 19), mean (range)
Hand-segmented crops	0.792 $\pm$ 0.101 (0.643–0.923)	68.67 $\pm$ 10.70	0.586 (0.357– 0.893)	0.815 (0.744– 0.910)
Automatic ROI (consistent)	0.879 $\pm$ 0.032 (0.837–0.919)	78.67 $\pm$ 2.98	0.864 (0.786– 0.929)	0.874 (0.846– 0.910)

Evaluated on automatic ROIs of the held-out test subset (n = 30); five seeds per protocol.

TABLE IV. DISCRIMINATION BY IMAGE SOURCE (DEPLOYED STATIC INT8 MODEL)

Source	n (202-image set)	Mean prob., anaemic	Mean prob., non-anaemic	Held-out AUC (n)
India	95	0.668	0.429	0.7857 (11)
Italy	107	0.657	0.126	0.8718 (19)

TABLE V. DESCRIPTIVE EFFECT OF INT8 QUANTIZATION ON RESNET-50

Variant	Size (MB)	Acc.	Sens.	Spec.	F1	AUC-ROC [95% CI]	Δ AUC
PyTorch FP32 (.pth)	90.01	83.33	69.23	94.12	0.7826	0.8688 [0.701–0.982]	—
ONNX FP32	89.62	83.33	69.23	94.12	0.7826	0.8688 [0.701–0.982]	0.0000
ONNX INT8 dynamic	22.59	86.67	69.23	100.00	0.8182	0.8778 [0.719–0.986]	+0.0090
ONNX INT8 static (deployed)	22.92	83.33	69.23	94.12	0.7826	0.8710 [0.710–0.984]	+0.0023

Nominal test subset, n = 30, threshold 0.5. The Δ AUC values are point estimates; paired Δ AUC confidence intervals and per-subject agreement were not calculated.

An independent, prospectively separated calibration cohort, clinically motivated sensitivity target and external threshold-level evaluation are required before clinical deployment.

Fig. 1 shows the ROC curve of the deployed model. The observed fraction of anaemic test subjects across five quantile bins of predicted probability was 0.17, 0.17, 0.29, 0.60 and 1.00, monotonically non-decreasing (Fig. 2). No non-anaemic subject exceeded a probability of 0.572, but four of the 13 anaemic subjects fell below 0.33 (0.109, 0.190, 0.281 and 0.324), so no threshold could detect all anaemic subjects without a substantial loss of specificity (Fig. 3 and Fig. 4).

G. On-Device Latency

In the cooled third static-INT8 session, average CPU forward-pass time was 89.31 ms (median 89.27; P90 89.52; P99 90.02 ms). Its two preceding static-INT8 sessions averaged 120.51 and 110.53 ms under hotter device conditions. The separate FP32 and dynamic sessions yielded means of 294.82 and 122.63 ms, respectively (Table VII); the arithmetic difference between 294.82 and 89.31 ms is 69.7%. Because the variants were not interleaved under the same thermal state and

used weights from different training stages, that percentage is an uncontrolled descriptive contrast, not a validated speedup attributable solely to static quantization.

The desktop reference for static INT8 was 24.00 ms (one thread) and 8.42 ms (runtime default threads). Across all devices, these measurements exclude image acquisition, decoding, ROI extraction, normalization, session initialization, user-interface work, peak memory, energy consumption and sustained thermal behavior. They do not measure the full camera-to-result pipeline or hardware acceleration.

V. DISCUSSION

A. On-Device Deployment

The strongest direct engineering finding is a 74.5% reduction in model-file size using static INT8 quantization. The held-out AUC point estimate changed little, but the 30-subject set, the absence of a paired difference interval and test-informed backbone selection prevent any accuracy-preservation or equivalence claim. The measured 89.31 ms is evidence that the static graph executes on the tested CPU in favorable

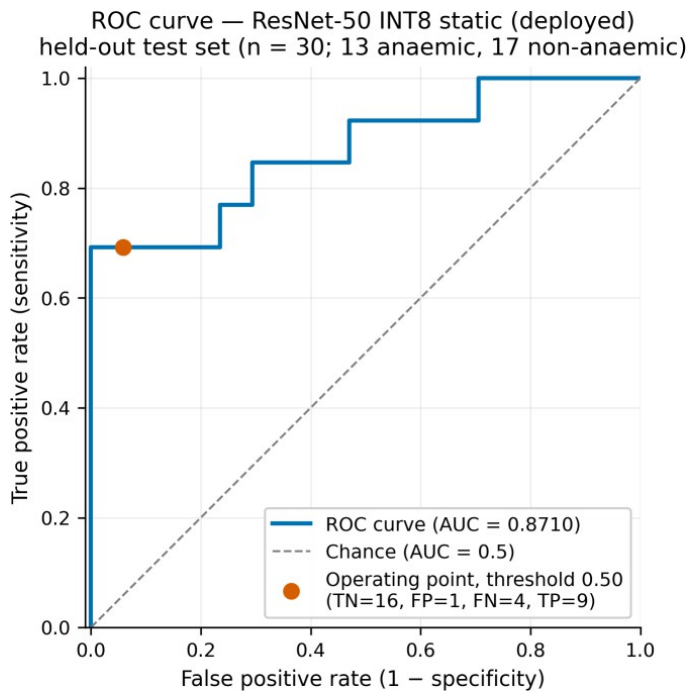


Fig. 1. ROC curve of the deployed static INT8 model on the held-out test subset (n = 30), AUC = 0.8710, with the operating point at the deployed threshold of 0.5 marked.

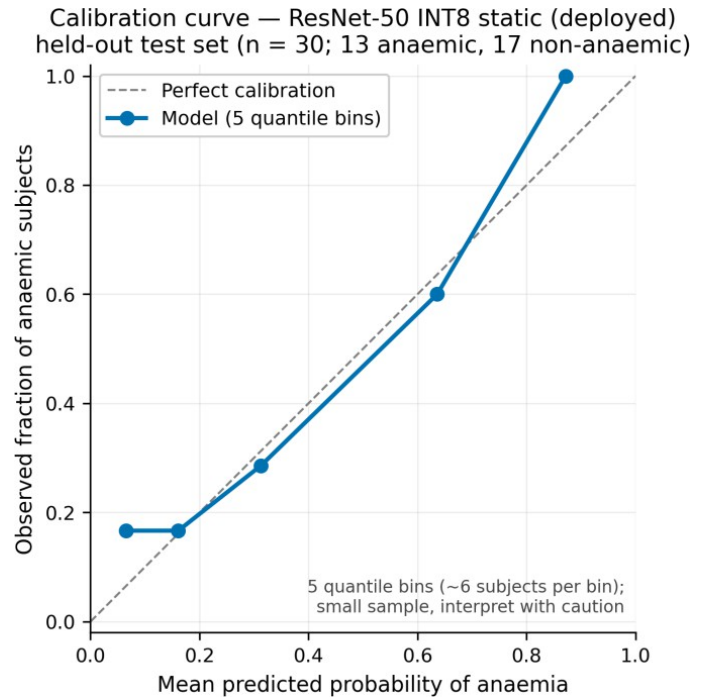


Fig. 2. Calibration curve of the deployed static INT8 model on the held-out test subset (n = 30, five quantile bins).

TABLE VI. DEPLOYED INT8 MODEL AT THE DEFAULT AND DEVELOPMENT-SELECTED THRESHOLDS

Threshold	Acc.	Sens. [95% CI]	Spec. [95% CI]	Prec.	F1	J	TN	FP	FN	TP
0.50 (default)	83.33	69.23 [42.37–87.32]	94.12 [73.02–98.95]	90.00	0.7826	0.6335	16	1	4	9
0.37 (development)	73.33	69.23 [42.37–87.32]	76.47 [52.74–90.44]	69.23	0.6923	0.4570	13	4	4	9
0.31 (exploratory)	76.67	76.92 [49.74–91.82]	76.47 [52.74–90.44]	71.43	0.7407	0.5339	13	4	3	10

Held-out test subset, n = 30; percentages in %, Wilson 95% CI.

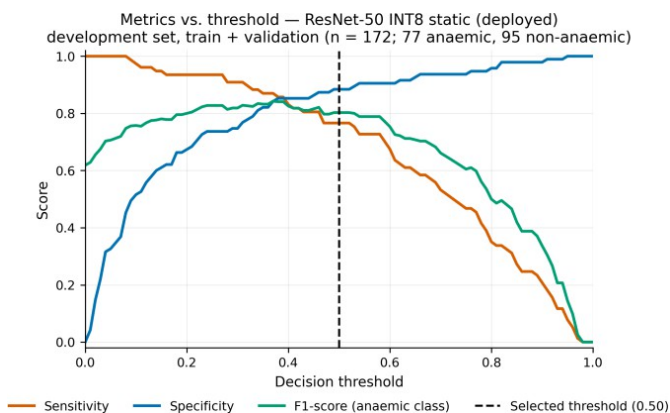


Fig. 3. Sensitivity, specificity and F1 as a function of threshold on the development set (n = 172; 77 anaemic, 95 non-anaemic), with the deployed threshold of 0.5 marked; F1 and Youden's J peak at 0.37 on this set.

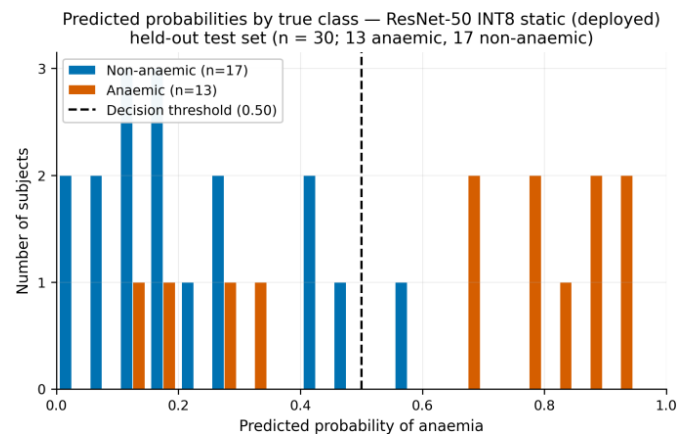


Fig. 4. Distribution of predicted probabilities by true class on the held-out test subset (n = 30), with the deployed threshold of 0.5 marked.

thermal conditions. It cannot establish a reproducible 69.7% speed advantage, superiority over dynamic quantization, sus-

tained responsiveness or end-to-end application performance. A mobile-native baseline remains necessary to determine whether

TABLE VII. EXPLORATORY FORWARD-PASS LATENCY FROM SEPARATE SESSIONS

Variant	Size (MB)	Mean latency (ms/image)	Std. dev. (ms)
ONNX FP32	89.62	294.82	2.09
ONNX INT8 dynamic	22.59	122.63	1.53
ONNX INT8 static (deployed)	22.92	89.31	0.22

POCO X5 5G; batch 1; 1 CPU thread; 15 warm-up passes and 100 recorded inferences/session. Static: cooled third session; FP32/dynamic: earlier weights, non-interleaved. Not end-to-end.

quantizing a heavier ResNet-50 is an efficient design choice [15].

B. Two Methodological Corrections

The study exposed two methodological hazards. First, the initial file-level split allowed crops from 28 of the 30 nominal test subjects into training, making its AUC of 0.9095 unsuitable as a generalization estimate [12], [13]. Second, models trained on white-background hand-segmented crops performed less consistently when evaluated on black-background automatic ROIs. Across five training seeds, matching training input to evaluation input increased mean AUC from 0.792 to 0.879 on the same nominal test split. These results illustrate how data identity and input compatibility can dominate small-dataset benchmarks.

That five-seed comparison is a repeated-training descriptive analysis, not subject-level cross-validation. Likewise, the apparent differences within Indian and Italian subjects rely on 11 and 19 cases and cannot exclude source confounding. Training and deployment code should share exact color conversion, morphological, resize, normalization and channel-order specifications, and external image conditions must be tested.

C. Limitations

Limitations are material. There were only 30 nominal test subjects (13 anaemic), with a wide static-INT8 AUC interval of 0.710–0.984 and no paired AUC-difference interval, score-agreement analysis, external set or repeated subject-level cross-validation. Backbone selection had been informed by the 202-image set containing test subjects. The color-based ROI was evaluated only on photographs from a controlled attachment and illumination system; robustness to other phones, lighting, white balance and skin tones is unknown. Countries were strongly imbalanced by anaemia status, and within-country test estimates were based on 11 and 19 subjects. At the reference threshold 0.5, four of 13 anaemic subjects were missed (sensitivity 69.23%), which is unsuitable evidence for a high-sensitivity clinical triage role. Benchmark sessions were not thermally matched and included synthetic-tensor inference only. No mobile-native baseline, peak-memory, energy, accelerator or sustained-use comparison was undertaken.

D. Practical Implications and Future Work

The present result supports an offline model-execution proof of concept and reduced storage demand, not a validated community-screening application. Before any clinical claim, a larger and source-balanced external dataset (including intended deployment populations), nested subject-level validation, a

threshold selected on independent calibration data for a clinically chosen sensitivity target, and end-to-end tests with real camera images are required. Device studies should compare current FP32, dynamic INT8, static INT8, and a mobile-native backbone under interleaved thermal-controlled trials while recording peak memory, power and accelerator behavior.

VI. CONCLUSION

This technical feasibility study evaluated static post-training INT8 quantization of a ResNet-50 conjunctival-image classifier on Android. Static INT8 reduced model-file size from 90.01 to 22.92 MB (74.5%). The 30-subject nominal test set gave FP32 and static-INT8 AUC point estimates of 0.8688 and 0.8710, but does not establish equivalence, especially after test-informed backbone selection. A cooled-session forward-pass benchmark averaged 89.31 ms, whereas earlier static sessions were slower and no thermal-controlled, end-to-end comparison was performed. With four of 13 anaemic test subjects missed at threshold 0.5, the current evidence does not validate clinical screening. Larger subject-independent evaluation, paired prediction analysis, mobile-native baselines and real-image device benchmarks remain necessary.

REFERENCES

- [1] GBD 2021 Anaemia Collaborators, "Prevalence, years lived with disability, and trends in anaemia burden by severity and cause, 1990–2021: findings from the Global Burden of Disease Study 2021," *Lancet Haematol.*, vol. 10, no. 9, pp. e713–e734, 2023. doi: 10.1016/S2352-3026(23)00160-6.
- [2] World Health Organization, *Guideline on Haemoglobin Cutoffs to Define Anaemia in Individuals and Populations*. Geneva, Switzerland: WHO, 2024. [Online]. Available: <https://www.who.int/publications/i/item/9789240088542>
- [3] A. Kalantri, M. Karambelkar, R. Joshi, S. Kalantri, and U. Jajoo, "Accuracy and reliability of pallor for detecting anaemia: A hospital-based diagnostic accuracy study," *PLoS ONE*, vol. 5, no. 1, art. e8545, 2010. doi: 10.1371/journal.pone.0008545.
- [4] R. G. Mannino et al., "Smartphone app for non-invasive detection of anemia using only patient-sourced photos," *Nat. Commun.*, vol. 9, art. 4924, 2018. doi: 10.1038/s41467-018-07262-2.
- [5] S. Collings, O. Thompson, E. Hirst, L. Goossens, A. George, and R. Weinkove, "Non-invasive detection of anaemia using digital photographs of the conjunctiva," *PLoS ONE*, vol. 11, no. 4, art. e0153286, 2016. doi: 10.1371/journal.pone.0153286.
- [6] S. Suner et al., "Prediction of anemia and estimation of hemoglobin concentration using a smartphone camera," *PLoS ONE*, vol. 16, no. 7, art. e0253495, 2021. doi: 10.1371/journal.pone.0253495.
- [7] G. Dimauro, D. Caivano, and F. Girardi, "A new method and a non-invasive device to estimate anemia based on digital images of the conjunctiva," *IEEE Access*, vol. 6, pp. 46968–46975, 2018. doi: 10.1109/ACCESS.2018.2867110.
- [8] G. Dimauro, M. E. Griseta, M. G. Camporeale, F. Clemente, A. Guarini, and R. Maglietta, "An intelligent non-invasive system for automated diagnosis of anemia exploiting a novel dataset," *Artif. Intell. Med.*, vol. 136, art. 102477, 2023. doi: 10.1016/j.artmed.2022.102477.

- [9] S. Roychowdhury et al., "Computer aided detection of anemia-like pallor," in Proc. IEEE EMBS Int. Conf. Biomed. Health Inform. (BHI), 2017, pp. 461–464. doi: 10.1109/BHI.2017.7897305.
- [10] P. Appiahene et al., "CP-AnemiC: A conjunctival pallor dataset and benchmark for anemia detection in children," Med. Novel Technol. Devices, vol. 18, art. 100244, 2023. doi: 10.1016/j.medntd.2023.100244.
- [11] O. Ramos-Soto, J. Ramos-Frutos, E. Perez-Zarate, D. Oliva, S. J. Mousavirad, and S. E. Balderas-Mata, "Non-invasive anemia detection from conjunctiva and sclera images using vision transformer with attention map explainability," Sci. Rep., vol. 15, art. 44142, 2025. doi: 10.1038/s41598-025-32343-w.
- [12] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, art. 100804, 2023. doi: 10.1016/j.patter.2023.100804.
- [13] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nat. Mach. Intell., vol. 3, no. 3, pp. 199–217, 2021. doi: 10.1038/s42256-021-00307-0.
- [14] G. Dimauro, R. Maglietta, T. Bai, and S. Kasiviswanathan, "Eyes-defy-anemia," IEEE Dataport, Jan. 31, 2022. doi: 10.21227/t5s2-4j73.
- [15] S. A. Cruz Romero and W. E. Lugo Beauchamp, "Performance analysis of post-training quantization for CNN-based conjunctival pallor anemia detection," in Lecture Notes in Networks and Systems. Cham, Switzerland: Springer, 2026, pp. 214–227. doi: 10.1007/978-3-032-09694-4_18.
- [16] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017. doi: 10.1038/nature21056.
- [17] A. Ignatov et al., "AI Benchmark: All about deep learning on smart-phones in 2019," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), 2019, pp. 3617–3635. doi: 10.1109/ICCVW.2019.00447.
- [18] T. K. Njoroge, R. Kibuku, and K. M. Sindu, "Edge-enabled mobile app for smart agriculture using multi-sensor inputs and a hybrid CNN-vision transformer model," Int. J. Interact. Mobile Technol., vol. 19, no. 21, pp. 145–162, 2025. doi: 10.3991/ijim.v19i21.55919.
- [19] M. N. M. Nawi, M. A. O. Mydin, A. Q. Adeleke, R. Deraman, N. A. M. Zaki, and R. Ghani, "Mobile deep learning framework for automated concrete surface crack detection and assessment on construction sites," Int. J. Interact. Mobile Technol., vol. 20, no. 3, pp. 83–94, 2026. doi: 10.3991/ijim.v20i03.60067.
- [20] M. S. Uddin, W. Slany, and K. J. Hasan, "On-device neural network for object train and recognition using mobile," Int. J. Interact. Mobile Technol., vol. 18, no. 12, pp. 99–111, 2024. doi: 10.3991/ijim.v18i12.47895.
- [21] M. A. Hafifi Saedan, M. Kassim, and A. F. Abd Aziz, "Biological butterfly characterization with mobile system using convolutional neural network (CNN) classify image," Int. J. Interact. Mobile Technol., vol. 18, no. 7, pp. 125–138, 2024. doi: 10.3991/ijim.v18i07.46267.
- [22] D. Srilatha et al., "A bio-inspired privacy preservation framework for interactive mobile healthcare applications in cloud environments," Int. J. Interact. Mobile Technol., vol. 20, no. 13, pp. 53–68, 2026. doi: 10.3991/ijim.v20i13.62246.
- [23] B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 2704–2713. doi: 10.1109/CVPR.2018.00286.
- [24] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," arXiv:1806.08342, 2018. doi: 10.48550/arXiv.1806.08342.
- [25] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A white paper on neural network quantization," arXiv:2106.08295, 2021. doi: 10.48550/arXiv.2106.08295.
- [26] ONNX Runtime developers, "ONNX Runtime," 2021. [Online]. Available: <https://onnxruntime.ai>
- [27] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," PLoS Med., vol. 15, no. 11, art. e1002683, 2018. doi: 10.1371/journal.pmed.1002683.
- [28] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," arXiv:1906.08172, 2019. doi: 10.48550/arXiv.1906.08172.
- [29] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), PMLR, vol. 97, 2019, pp. 6105–6114.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [31] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Adv. Neural Inf. Process. Syst. 32 (NeurIPS), 2019, pp. 8024–8035.
- [32] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
- [33] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988. doi: 10.2307/2531595.
- [34] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," J. Amer. Statist. Assoc., vol. 22, no. 158, pp. 209–212, 1927. doi: 10.1080/01621459.1927.10502953.