Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Model-dependent accuracy inflation in machine learning evaluation on public sickle cell imaging datasets

Philomina Jenifer Antonysamy
SRM Institute of Science and Technology
Priti Rishi
pritir@srmist.edu.in

SRM Institute of Science and Technology

Research Article

Keywords: Sickle cell disease, Data leakage, Source-held-out validation, Near-duplicate detection, Label error, Optimism correction, Trustworthy AI, Screening

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11151372/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Model-dependent accuracy inflation in machine learning evaluation on public sickle cell imaging datasets

A. Philomina Jenifer and **Priti Rishi**

Department of Electronics and Communication Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai 600 026, Tamil Nadu, India

pa9233@srmist.edu.in · pritir@srmist.edu.in (corresponding author)

Abstract

Background Automated sickle cell detection is reported at high accuracy on a small number of shared public image collections, yet whether those collections support the accuracies reported on them has not been measured. This study quantifies how much detection performance is attributable to how the datasets were built rather than to the erythrocytes.

Methods Four public corpora were analysed. Repeated images were recovered by dihedral, shift-tolerant correlation verified at two resolutions, with the grouping threshold anchored to a blinded two-rater reference study of 60 image pairs. Inflation was the loss in area under the curve when sources were held out, using hand-crafted features with two learners and frozen and fine-tuned convolutional networks. File traces were tested with an accuracy-based and a balanced confounding index against permutation nulls; drawn annotations were located by differencing the two released versions of one corpus; and bootstrap optimism correction was compared with the source-held estimate.

Results Similarity separated true repeats from distinct fields with an AUC of 0.978; above 0.55 every consensus pair was a repeat, whereas a threshold calibrated on synthetic copies recovered only 66% of genuine repeats. With hand-crafted features, holding sources out reduced AUC by 0.094 on a release with twelve duplicated control fields, by 0.025 to 0.084 on a smartphone corpus photographing each field about twice, and by 0.092 under participant-level holdout. A fine-tuned network lost 0.021 on the smartphone corpus, its source-held error 1.8 times the naive error, but nothing on the duplicated release. The accuracy-based index read zero for traces marking a third of the majority class, which the balanced index recovered. A resolution trace was visible to features and network yet unused, whereas bounding boxes in one corpus's labelled release raised AUC by 0.058. Bootstrap optimism correction overstated source-held performance, and one field was filed under both classes.

Conclusions Inflation depends on the model and must be measured for each model reported. Participant-level dependence cannot be verified by image-level source recovery, and most releases distribute no identifiers. Six checks, each calibrated to the corpus itself, are proposed as an acceptance test for imaging tools entering screening programmes.

Keywords Sickle cell disease, Data leakage, Source-held-out validation, Near-duplicate detection, Label error, Optimism correction, Trustworthy AI, Screening

1 Introduction

Sickle cell disease affects an estimated 300,000 newborns each year, concentrated in sub-Saharan Africa and South Asia [1], and outcomes depend on early identification and sustained care [2]. India's National Sickle Cell Anaemia Elimination Mission launched in 2023, aims to screen about 70 million people up to 40 years of age with validated point-of-care tests, alongside management and

counselling [3, 4]. Automated analysis of blood films is repeatedly proposed as a scalable aid in that setting [5–7].

The evidence for such tools rests on a few shared public collections, and the accuracies reported on them are high — including a network reported by the present authors on one of the corpora examined here [8]. An earlier audit by the same authors found that class membership in an assembled corpus including one of these collections could be recovered from how its images were acquired [9]. The reported accuracies are almost always obtained by splitting images at random. A random split assumes that images are independent, and in medical imaging that assumption routinely fails: models learn the hospital, scanner or patient rather than the disease [10–13], and leakage between training and test data inflates reported performance across fields [14–16], including through near-duplicate images shared between training and test sets [17]; in routine health records, the pattern of missing data can itself become a model's strongest predictor [18]. Reporting standards now require authors to state how evaluation data were kept independent [19–21]. They cannot, however, tell a reader how to establish independence when a release carries no record of which images came from the same source.

Three kinds of dependence can break a random split in a blood-film corpus. The same image may be saved twice. The same field may be photographed again, with different exposure, focus or zoom, producing an image that no hash will match. And different fields from the same person share that person's blood. Each lets a classifier score well by recognising what it has already seen. None has been measured in the public sickle cell record.

This study measures them. Its objectives were:

1. to recover which images in public sickle cell corpora share a source, by a procedure validated against blinded human judgement;
2. to quantify how much detection accuracy is lost when sources are held out, across four corpora, for hand-crafted features and for frozen and fine-tuned convolutional networks;
3. to test whether file-level traces and drawn annotations present in the corpora are available to a classifier, and whether they are actually used;
4. to determine whether standard bootstrap optimism correction recovers the performance measured when sources are held out; and
5. to formalise the resulting checks as an acceptance test for imaging tools entering screening programmes.

Each objective corresponds to a section of the Results: source recovery (Sect. 3.1), inflation and its stability (Sects. 3.2–3.3), file traces and their use (Sects. 3.4–3.6), optimism correction (Sect. 3.7) and the acceptance test (Sect. 3.8). Together they treat the trustworthiness of an AI evaluation as a question that can be answered from the data, which matters most where such tools are meant to serve health systems at scale.

2 Methods

2.1 Data sources

Four public corpora were analysed (Table 1). Each defines a different task, so they are not pooled; each is analysed on its own terms.

Corpus	Content	Task	Identifiers
RedTell	Brightfield fields paired with calcium fluorescence	SCD vs control	none

Corpus	Content	Task	Identifiers
	[22]; 24 SCD, 37 control fields used here; 18 thalassemia fields not used		
Tushabe	Smartphone photographs through a microscope eyepiece [23]; 422 sickle, 147 normal	sickle present vs absent	none
erythrocytesIDB	Individual cell crops from SCD patients [24]; 211 elongated, 202 circular	elongated vs circular	none
erythroSight	Differential phase contrast tiles with per-cell morphology measurements [25]; 1,224 tiles from 3 SCD and 3 normal participants; 3 sickle cell trait participants used only in Check 5	SCD vs normal	participant

Table 1. The four corpora. Only erythroSight distributes participant identifiers. Every erythrocytesIDB donor has sickle cell disease, so its task is shape classification rather than disease detection. RedTell's thalassemia fields and erythroSight's trait participants were excluded from the two-class analyses.

Tushabe is distributed by its creators in two versions [23], with and without bounding boxes drawn around the sickle cells. The copy analysed throughout comprises the deposit's negative folder and unlabelled positive folder, here named Normal and Sickle; the labelled positive folder was used only in Sect. 3.5. The creators report three photographs of each smear, taken on four phones through two microscopes, from 140 participants [23].

2.2 Recovering sources

Every image was converted to greyscale, cropped to its central 60% to remove the eyepiece vi-gnette, high-pass filtered by subtracting a Gaussian blur of one-eighth the image width, and weighted by a Hanning window, so that a shared bright disc or illumination gradient could not drive the comparison. Similarity between two images was the maximum normalised cross-correlation over the eight flips and rotations of the square and over translations of up to 15% of the image, computed by phase correlation [26, 27]. For RedTell and erythrocytesIDB, the ten best candidate matches of each image were found at 64 × 64 pixels and each was verified at 128 × 128, where gen-uine repeats keep their cells aligned and texturally similar fields fall apart. Two images were linked only if each was among the other's three closest verified matches, so that a single chance match could not chain unrelated images together, and sources were the connected components of the re-sulting graph. For Tushabe, whose human-anchored threshold is defined at 128 pixels, similarity was computed directly at 128 × 128 for every pair, with the same mutual linking.

2.3 Human reference study

To set the linking threshold from human judgement rather than from synthetic copies, 60 Tushabe image pairs were drawn at random within six similarity bins spanning 0.30 to 1.00, ten per bin, and presented in random order at full size with no similarity score shown. Each pair was labelled as the same field, a different field, or unsure, the criterion being the same cells in the same arrangement, allowing for shift, rotation, zoom and exposure. The first author labelled all pairs and served as the reference standard. A second, independent rating was produced by an AI model (Claude, Anthropic) viewing the same images before either the first author's labels or the scores were seen; its labels

were used only to measure agreement, by Cohen's kappa [28], and are an additional assessment rather than an independent human inter-rater validation. One pair (pair 8) was excluded, leaving 59: it had been described in detail as a worked example while the enlarged sheets were prepared, so it was no longer blind. The threshold was set from these judgements of image identity alone: the raters saw neither the similarity scores nor any diagnostic label, and no classification result entered the choice. Labelling was completed before the final inflation runs. Because the threshold was nev-ertheless chosen on the same corpus to which it is applied, the sensitivity of every conclusion to that choice is reported in Sect. 3.3. Because pairs were sampled within bins, the study measures how well similarity separates repeats from distinct fields, not how common repeats are in the corpus, and pairs below 0.30 were not sampled. Agreement on definite answers and discrimination were com-puted on the pairs both raters decided; pairs with any "unsure" answer were excluded from them.

2.4 Inflation

Accuracy inflation, the term used throughout, was defined as the difference between the area under the curve under a random image-level split and under a split that keeps every image of a source in the same fold. Inflation is therefore an observed difference between two evaluation designs on the same data, not an estimate of a causal effect, and no significance test is applied to it. Both used five-fold stratified cross-validation, the second in its grouped form [29], repeated over twelve to twenty different fold assignments so that the held-out estimate carries its own spread. Two feature sets were used — grey-level histogram, local binary pattern and orientation-gradient features, and a colour-histogram and thumbnail set — each with a random forest of 150 trees and an L2-regularised logistic regression on standardised features. The same extractors and learners were applied to every image corpus, so that differences between corpora reflect the data rather than the model. For ery-throSight, the source was the participant, and the features were per-tile medians of nine cell-mor-phology measurements distributed with the collection rather than image features; with three partic-ipants per class, grouped cross-validation used three folds, and fold assignments that left a test fold with a single class were discarded. Two convolutional networks were evaluated on RedTell and Tushabe with the same sources [30]: embeddings from an ImageNet-pretrained ResNet-50 at 448 × 448 pixels with an L2-regularised logistic regression over twelve fold assignments, and an Ima-geNet-pretrained ResNet-18 fine-tuned end to end at 320 × 320 pixels for ten epochs (AdamW, learning rate 10^{-4}, cosine schedule, class-weighted loss, random dihedral and brightness–contrast augmentation, test-time averaging over four rotations) over three fold assignments. Folds, epochs and augmentation were identical under both splits, so that only the division of images differed.

2.5 File traces

Whether a file property alone recovers the class was measured by a depth-one decision tree on the property, in two forms. The accuracy form, (A − B)/(100 − B) with B the majority-class rate, fol-lows the authors' earlier corpus-level index [9]. The balanced form, (BA − 0.5)/0.5 with BA the balanced accuracy, weights both classes equally. The balanced form was compared with its null dis-tribution under 200 label permutations, with p = (k + 1)/201 for k permuted values at least as large as the observed one. The four file properties are correlated and were treated as exploratory; no cor-rection for multiple comparisons was applied.

2.6 Whether a trace is used

A trace present in the files may or may not be used by a classifier. For a flagged trace, detection was re-estimated within the strata it defines, and the classifier was asked whether it could separate the strata from features alone.

2.7 Drawn annotations

Every image in one Tushabe copy was matched to its counterpart in the other by correlation of 64-pixel thumbnails, and the pair was differenced at a common resolution. Pixels dark in one copy and not in the other were retained, and straight horizontal and vertical segments among them were isolated by morphological opening. Validation on a controlled case showed that a thin box drawn on one copy is detected in that copy only, and that a copy that is merely re-compressed is not flagged. Pairs left unmatched were resolved by eye and by correlation over the eight flips and rotations. A separate control, using the same feature types with a 500-tree random forest over ten repeated five-fold splits, compared the two copies on all images, on the subset of sickle images 1,000 pixels wide together with all normals, and after re-encoding the unlabelled images at JPEG quality 95 to ap-proximate the file size of the labelled copy.

2.8 Optimism correction

Harrell's bootstrap optimism correction [31–33] was compared with the source-held estimate, both by resampling images and by resampling sources. To keep predictors few relative to images, as in clinical prediction, the model for this comparison was ten principal components followed by logistic regression; bootstrap correction fails when predictors outnumber observations, and that failure had to be kept distinct from this one.

2.9 Validation of the instruments

Copies of known images were planted under ten transformations — exact, flip and rotation, JPEG compression, resizing, brightness change, noise, and crops and shifts — to measure how many the source recovery finds. A second archive was simulated by re-encoding a growing share of one class, to measure when the confounding index detects it.

2.10 Use of AI tools

An AI model (Claude, Anthropic; accessed through claude.ai in September 2026) was used under the authors' oversight for two purposes. First, it provided an independent second rating in the ref-erence study, used only to measure agreement, with the first author's labels as the reference stan-dard (Sect. 2.3); it viewed the same enlarged pair images under the same written labelling guide, and its labels were recorded before the first author's labels, the similarity scores or the class folders were opened. Second, it supported the writing of analysis code and the editing of the text. No con-fidential or identifiable data was shared with any external service; the images used are public. All figures were produced by the released code from the analysis outputs, and no generative model was used to create or alter any image. All code was run and all outputs, analyses, citations and conclu-sions were checked by the authors, who take full responsibility for the content.

3 Results

3.1 Source recovery against human judgement

Of the 59 scored pairs, both raters gave a definite answer on 52; the rest received at least one "unsure". The two raters agreed on every one of those 52 ($\kappa = 1.000$); every disagreement was one rater answering "unsure", and the three-category κ was 0.802. Against the consensus, similarity separated repeats from distinct fields with an AUC of 0.978 (95% bootstrap interval 0.933 to 1.000). No pair between 0.30 and 0.45 was a repeat, three quarters of pairs between 0.45 and 0.55 were, and every

pair at 0.55 or above was (35 of 35; Wilson 95% interval 0.90 to 1.00), so the data are consistent with a true rate of 0.90 or more above that threshold (Table 2, Fig. 1a).

Similarity at 128 px	Pairs	Same field	Proportion	Reading
0.30 to 0.45	9	0	0.00	distinct
0.45 – 0.55	8	6	0.75	mostly repeats
0.55 – 0.65	8	8	1.00	repeats
0.65 – 0.75	9	9	1.00	repeats
0.75 – 0.90	8	8	1.00	repeats
0.90 and above	10	10	1.00	repeats

Table 2. Probability that a pair is the same field, by similarity, against the consensus of both raters (52 decided pairs). With 8 to 10 pairs per bin, the Wilson 95% interval for a proportion of 1.00 has a lower limit of about 0.70, and that for 0.75 spans 0.41 to 0.93.

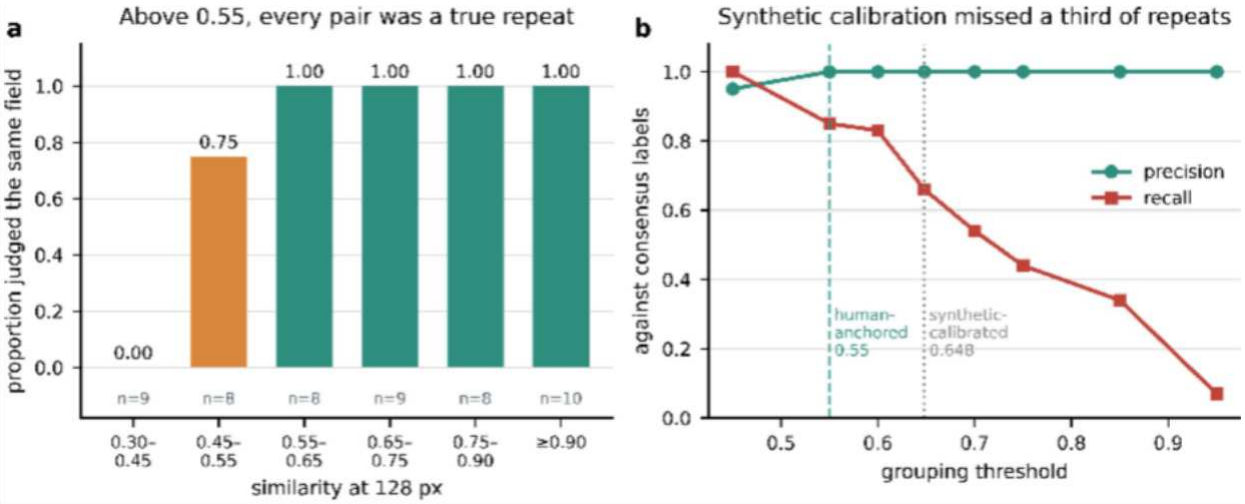


Fig. 1. Human reference study. (a) Proportion of pairs judged the same field in each similarity bin. (b) Precision and recall of source recovery against threshold. The threshold calibrated on synthetic copies (0.648) recovers 66% of genuine repeats; the human-anchored threshold (0.55) recovers 85%, both at a precision of 1.00.

The threshold that had been calibrated on synthetic copies was 0.648. It kept precision at 1.00 but recovered only 66% of genuine repeats, against 85% at 0.55 (Fig. 1b). Genuine re-captures differ from one another in exposure, focus, zoom and vignette more than flipped or re-compressed copies do, so they score lower; a threshold set on synthetic copies therefore underestimates how far real repeats drift. On RedTell, where the repeats are byte-identical, the procedure recovered exactly the twelve duplicated control fields and the same 49 sources at every threshold from 0.65 to 0.95.

3.2 Inflation across the corpora

Holding sources out lowered detection on every corpus not at ceiling (Table 3, Fig. 2). On RedTell the loss was 0.094 with grey features and a random forest, and 0.067 to 0.111 across the four combinations. On Tushabe, with the human-anchored threshold and mutual linking, 569 images collapsed to 280 independent fields — about two photographs per field — and the loss was 0.025 to 0.084. The creators report three photographs per smear [23], so the number of independent fields is likely lower still. Two pairs of normal images are byte-identical (141.jpg and 144.jpg; 142.jpg and 143.jpg), as noted in the earlier audit [9], and 4.jpg and 91.jpg are one sickle field saved at two sizes, which the grouping placed in separate sources. On erythroSight, where participants could be held out directly, the random forest lost 0.092. The erythrocytesIDB task was at ceiling under both

splits; because source recovery is unreliable on isolated cell crops (Sect. 4.1), its source-held value is reported for completeness and not interpreted.

The same comparison with convolutional networks separated the two kinds of repeat. On Tushabe, where fields were re-photographed, inflation persisted for every model: 0.010 with frozen ResNet-50 embeddings (naive 0.964, source-held 0.954) and 0.021 with a fine-tuned ResNet-18 (0.973 against 0.952), positive in each of its three paired repeats (0.015 to 0.030). Near the ceiling a small difference in AUC is a large one in error: for the fine-tuned network, one minus the AUC rose from 0.027 to 0.048, so the source-held error was 1.8 times the naive one. On RedTell, whose repeats are byte-identical copies of control fields, neither network lost anything (−0.009 frozen, −0.007 fine-tuned; naive 0.918 and 0.972), although the fine-tuned naive estimate ranged from 0.932 to 1.000 across repeats. Duplicated images therefore inflated the hand-crafted models only, whereas re-photographed fields inflated every model tested.

Corpus	Naive	Source-held	Source-held range	Difference	Across features and learners
RedTell	0.824	0.730	0.681 to 0.792	+0.094	+0.067 to +0.111
Tushabe	0.905	0.880	0.874 to 0.889	+0.025	+0.025 to +0.084
erythroSight	1.000	0.908	0.816 to 0.997	+0.092	+0.011 to +0.092
erythrocytesIDB	1.000	1.000	0.999 to 1.000	0.000	ceiling; uninformative

Table 3. Detection AUC under a random split and a source-held split, grey features with a random forest; for erythroSight, cell-morphology measurements with participant holdout. The range is the 2.5th to 97.5th percentile of the source-held estimate over repeated fold assignments; it describes that variability and is not a confidence interval. Differences are reported as observed quantities and were not subjected to significance testing. Tushabe uses the human-anchored threshold with mutual linking, the most conservative setting. The last column gives the range of the difference over the four feature–learner combinations (two learners for erythroSight). The erythrocytesIDB row is shown for completeness only.

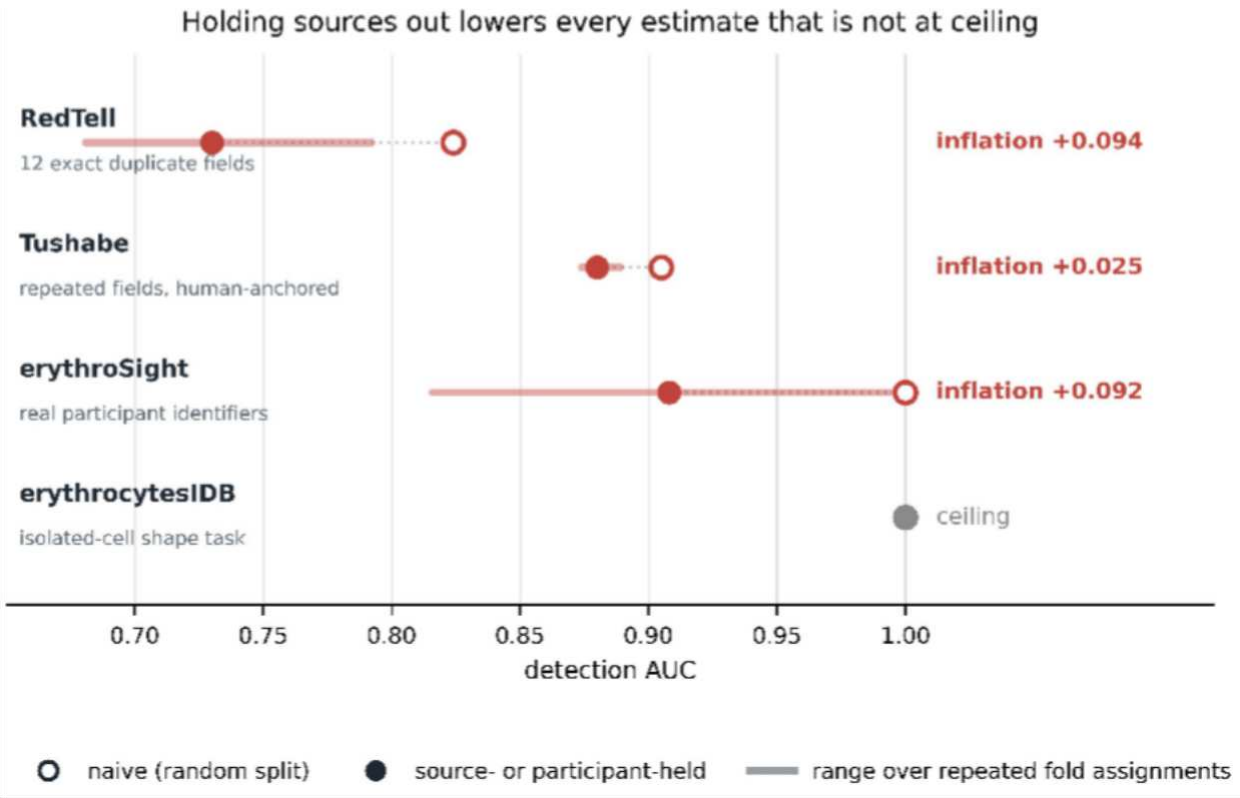


Fig. 2. Inflation across the four corpora. Open circles give the random-split AUC, filled circles the source- or participant-held AUC, and bars the range over repeated fold assignments.

There was no consistent learner effect. Logistic regression lost more than the random forest on Tushabe, less on erythroSight, and about the same on RedTell. Flexible learners can be more exposed to dependence, but the data do not show that they always are.

These estimates cover only the dependence that pixels can reveal. Different fields from one person show different cells, so the image-level comparison used here cannot link them (Fig. 3). Only erythroSight distributes identifiers, and there participant-level holdout cost 0.092 — as much as the image-level duplicates in RedTell. On Tushabe, whose 569 images come from 140 participants [23], the filenames run continuously from 1 in each class folder, capture order could not be recov-ered from them, and the participant-level share of inflation cannot be measured.

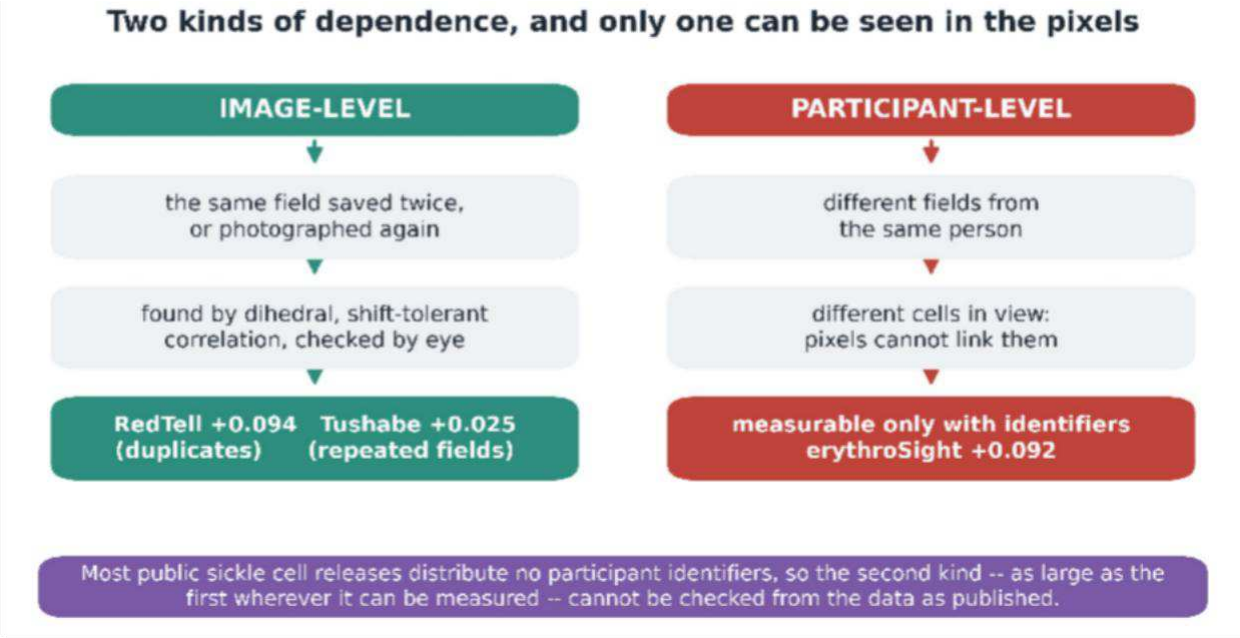


Fig. 3. Two kinds of dependence. Image-level repeats can be recovered and their cost measured; participant-level dependence can be measured only where identifiers are distributed.

3.3 Stability of the estimate

Across thresholds from 0.45 to 0.648, the number of independent Tushabe fields moved only from 249 to 297, and inflation with grey features and a forest stayed between 0.019 and 0.038 (Fig. 4). The result does not depend on where the threshold is set within that range.

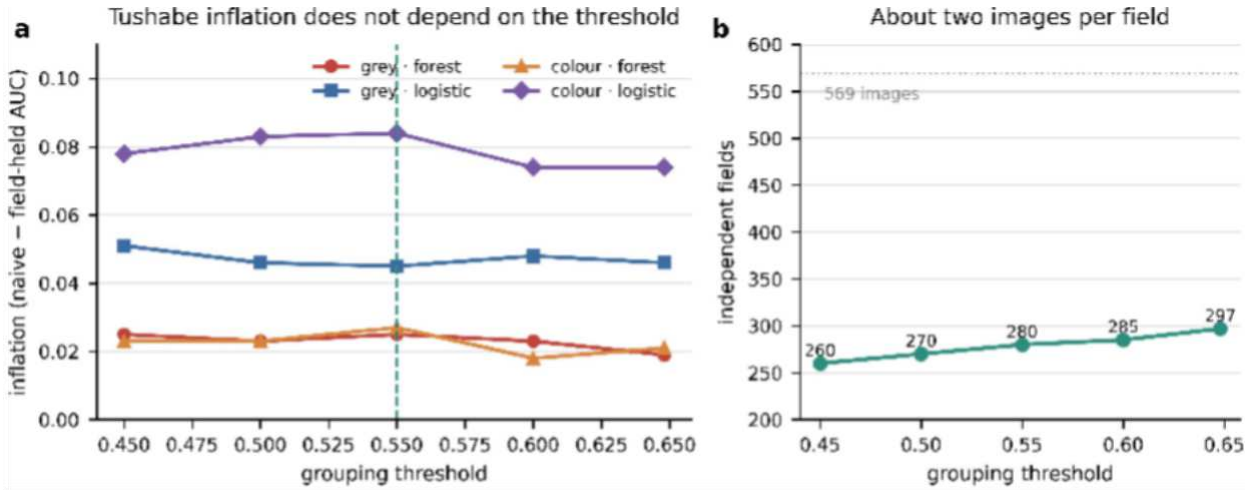


Fig. 4. Tushabe. (a) Inflation for each feature set and learner against grouping threshold. (b) Independent fields against threshold; the dashed line in (a) marks the human-anchored threshold.

3.4 File traces and the balanced index

On Tushabe, file size, width, height and pixel count all gave an accuracy-based index of zero, and balanced indices of 0.322 to 0.354 ($p \leq 0.005$, the smallest value attainable with 200 permutations; Fig. 5). Sickle images are the majority class, at 74%, so a trace that picks out more sickle images cannot beat always guessing sickle, and the accuracy form cannot register it. The balanced form equals the fraction of a class that a trace identifies: in a simulation built from Tushabe's counts it read 0.323 against a marked fraction of 136/422 = 0.322, and on RedTell, with uniform files, the index equalled the re-encoded share exactly (Supplementary Fig. S1). The folders also differ in stain: 57% of sickle and 38% of normal images were Leishman-stained [23], an imbalance shown in the earlier audit to be recoverable from colour statistics [9]; stain is not recorded per file and was not tested here.

Two flags were not acquisition traces, and are reported as such. Tushabe's filename ordinal gave 0.642 only because each class folder is numbered from 1 and the sickle folder holds more files; an ordinal is an acquisition trace only when numbering is global, as in RedTell, where it gave 1.000. erythrocytesIDB's file size gave 0.292 because an elongated cell compresses to a different size than a round one — biology reaching the file, not acquisition.

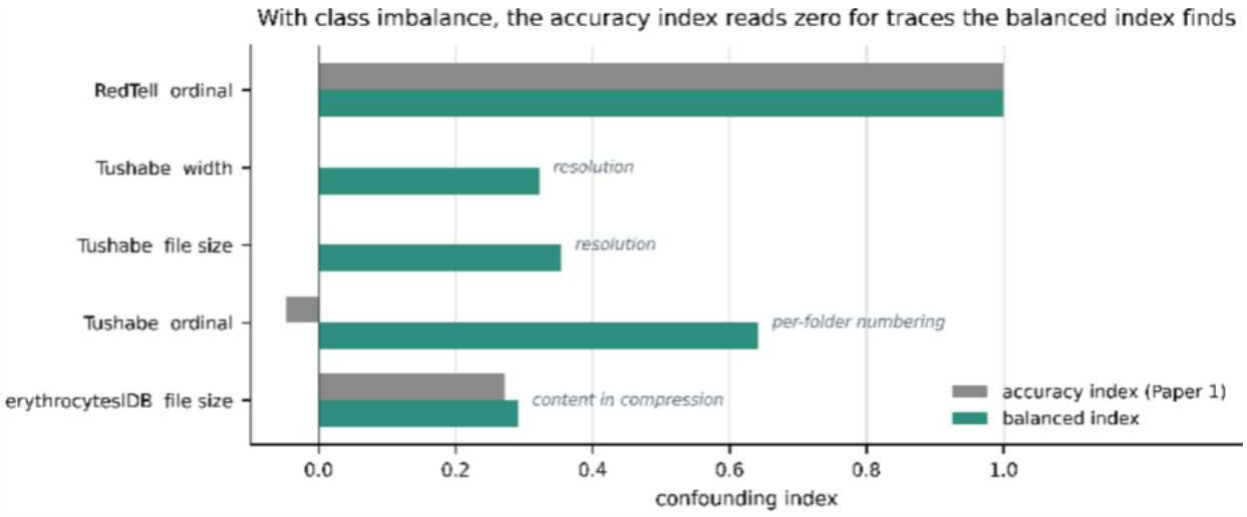


Fig. 5. Accuracy-based and balanced confounding index for the traces that carry class information. Italic notes give the reason for each.

3.5 A trace can be present and unused

All 147 normal Tushabe images are 1,000 pixels wide, whereas of the 422 sickle images 283 are 1,000 pixels wide, 136 are at native phone resolution (1,749 to 4,160 pixels) and three are smaller; no normal image is at native resolution. The creators state that every image was resized to a width of 1,000 pixels [23]; the labelled folder conforms, but the unlabelled folder analysed here does not. The features could see the difference — native and 1,000-pixel sickle images were separable at an AUC of 0.866 — but the classifier did not use it as a shortcut. Native-resolution sickle images were harder to detect than 1,000-pixel ones (0.860 against 0.938), and removing them raised AUC from 0.905 to 0.944 (Fig. 6a–b) — a hidden stratum within the positive class, of the kind that aggregate accuracy conceals [34]. The frozen network behaved the same way: it separated native from 1,000-pixel sickle images almost perfectly (AUC 0.993), yet native images were again the harder ones (0.957 against 0.974), and removing them raised AUC from 0.964 to 0.977.

The labelled folder behaved differently. An earlier audit by the present authors found that assembling a corpus from it raised a reference classifier's accuracy, but global image statistics could not locate the cause [9]. The unlabelled folder analysed throughout has a mean sickle file size of 415 kB and the labelled folder 177 kB, and all 147 normal images were byte-identical between the two. Every one of the 422 sickle images in the labelled folder carried black rectangles drawn around individual sickle cells. Of these, 417 were paired with their unlabelled counterparts automatically; three more were paired by eye and by rotation-tolerant correlation, one having been rotated by 90° (93.jpg) and two cropped (91.jpg, 176.jpg). The remaining two, 35.jpg and 155.jpg, were byte-iden-tical in both folders and carried boxes in both, so the unlabelled folder is not entirely free of anno-tation. On the labelled folder detection rose from 0.905 to 0.963, a gain of 0.058 that comes from the drawn boxes rather than from the cells (Fig. 6c). In the separate control (Sect. 2.7) the gain did not depend on image size or compression: restricted to the 283 sickle images at 1,000 pixels and all normals, the boxes removed 89% of the remaining error against 94% on the full set, whereas re-en-coding the unlabelled images to the file size of the labelled folder left detection unchanged (AUC 0.972 before and after). Marks of this kind have been shown to drive classifiers in dermatology [35]; here they were located without any model of what a mark looks like, because two versions of the same release were available to compare.

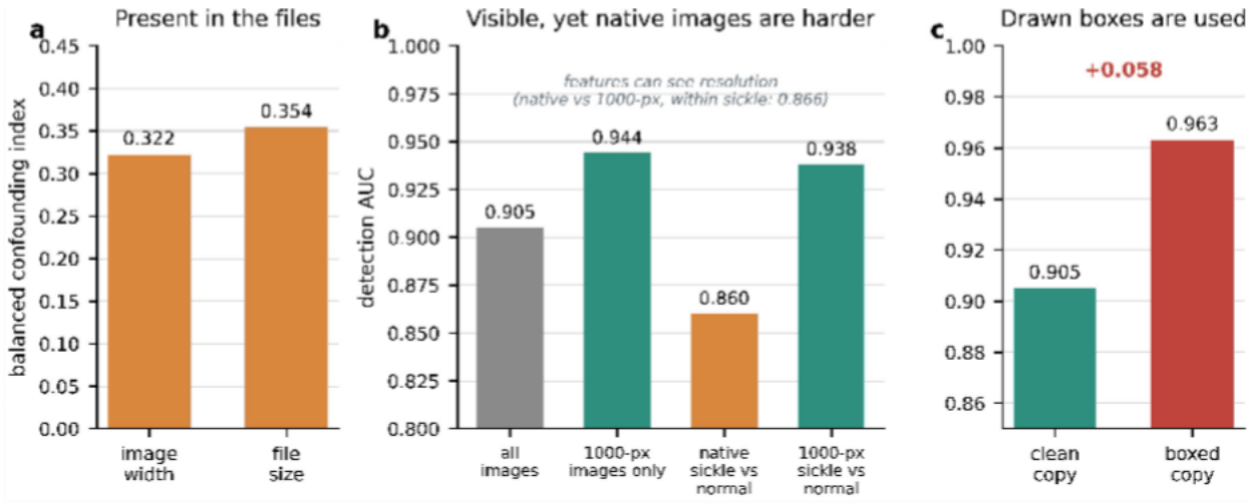


Fig. 6. Availability and use. (a) The resolution trace is present in the files. (b) It is visible to the features but native-resolution images are harder, not easier. (c) Boxes drawn on the sickle images of the labelled folder are used, raising AUC by 0.058.

3.6 A field filed under both classes

Three groups of repeated fields spanned both classes. In one, confirmed by eye, a single field appears as Normal/146.jpg and as Sickle/208.jpg and Sickle/366.jpg: the same cells in the same arrangement, differing only in brightness and vignette (Fig. 7). Two distinctive cells lie 218, 221 and 216 pixels apart in the three photographs, a spread of under 3%. Label errors of this kind are common even in widely used benchmarks [36], but here the conflict was found through the repeated fields rather than through a trained model. Because a label records whether at least one sickle cell was identified in that photograph [23], a conflict can arise if a sickle cell is visible in one exposure and not in another; the data cannot say which label is correct, and the conflict is reported as such. The other two groups are candidates that have not yet been inspected.

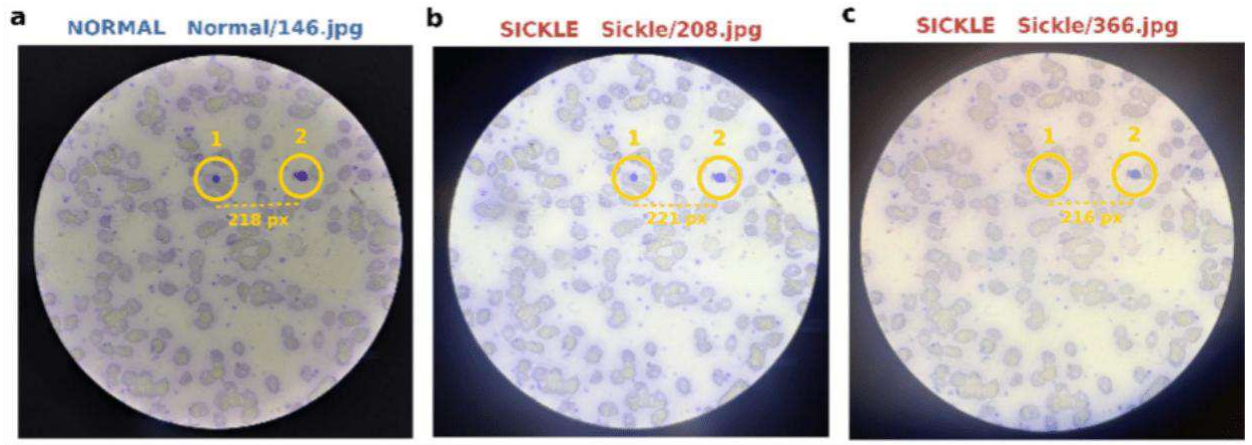


Fig. 7. One field filed under both classes in the published Tushabe dataset. (a) Normal/146.jpg; (b) Sickle/208.jpg; (c) Sickle/366.jpg. The two circled landmark cells occupy the same positions in all three photographs and lie 218, 221 and 216 pixels apart; the images differ only in exposure and vignette. Photographs are shown as distributed, each resized to a common height with no other adjustment; magnification is not recorded in the source dataset, so no scale bar can be given.

3.7 Standard optimism correction

Harrell's bootstrap correction gave 0.723 on RedTell against a source-held estimate of 0.674, and 0.715 on Tushabe against 0.697 (Fig. 8). These figures come from the ten-component logistic-regression model of Sect. 2.8 and are not comparable with the random-forest estimates in Table 3. Resampling sources rather than images moved the estimates only slightly, to 0.713 and 0.702. Boot-

strap correction always scores each model against the full original set, which contains the very images — and their repeats — that the model was trained on; no resampling scheme removes that, and only holding sources out does.

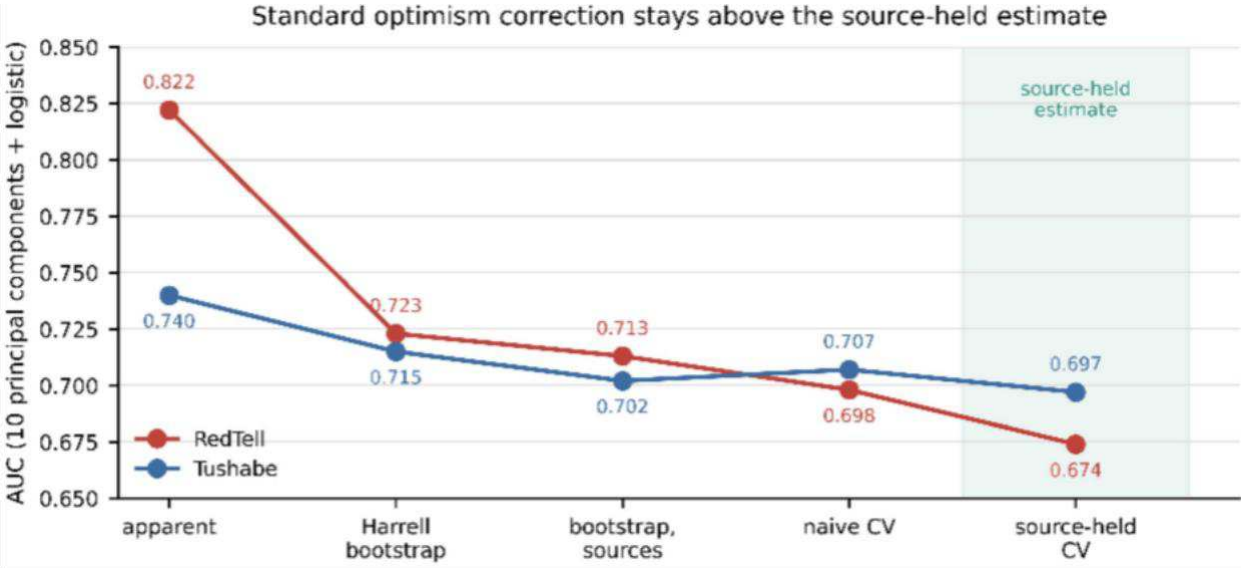


Fig. 8. Optimism correction against the source-held estimate, for ten principal components with logistic regression. The model differs from that of Table 3, so the values are not comparable with it.

3.8 An acceptance test

The checks were assembled into a proposed six-step test for an imaging tool before independent validation for a screening district (Fig. 9): file traces, repeats, the source-held gap, whether any flagged trace is used, the direction of effects across three groups where a third group is imaged in the same collection, and label conflicts among repeated sources. Fixed cut-offs were rejected by the validation: on RedTell, whose files are uniform, a re-encoded archive was detected once it made up a quarter of one class, whereas on Tushabe, photographed on four phones [23], the same archive stayed hidden until it made up the whole class (Supplementary Fig. S1). Each check is therefore judged against its own corpus rather than a fixed cut-off: file traces against a permutation null, repeats against the human-anchored threshold, and the source-held gap against its spread over fold assignments. In the form used here, Check 3 reports the source-held figure whenever the naive AUC lies above the upper end of the source-held range, and Check 4 treats a trace as used when the images it marks are classified more accurately than the rest, or when removing it lowers AUC by more than that spread. Check 5 was demonstrated on erythroSight, with its three sickle cell trait par-ticipants as the third group. A model trained on SCD and normal participants alone placed all three trait participants with the normals (mean predicted probability of SCD 0.000 to 0.041, against 0.994 to 1.000 for SCD), and trait shape measurements lay between normal and SCD, close to normal (eccentricity, circularity, aspect ratio and curl at 0.06 to 0.18 of the distance), although area and major axis lay slightly beyond normal. This is the direction expected of a disease signal rather than an acquisition trace, since trait carriers have largely normal red cells. The SCD and normal scores are in-sample, and three participants per group allow only a descriptive reading.

Check	What it tests	Evidence from this study
1 File traces	balanced confounding index on size, dimensions and filename order, against a permutation null	Tushabe width 0.322 and file size 0.354, both read 0.000 by the accuracy index; RedTell ordinal 1.000
2 Repeats	shift-tolerant correlation, threshold set from pairs labelled by eye	Tushabe 569 images → 280 fields, two byte-identical normal pairs; RedTell 12 duplicate controls; rater

Check	What it tests	Evidence from this study
		agreement $\kappa = 1.000$
3 Source-held gap	naive against source-held AUC over repeated fold assignments	RedTell +0.094; Tushabe +0.025 to +0.084; erythroSight +0.092; fine-tuned ResNet-18 on Tushabe +0.021
4 Use of a trace	stratified evaluation of any flagged trace	resolution present but unused (AUC 0.905 → 0.944 when removed); drawn boxes used (+0.058)
5 Three-group direction (conditional)	direction of each effect across disease, a third group and control, where one is imaged in the same collection	erythroSight with sickle cell trait as third group: trait scored with normals (P(SCD) ≤ 0.041); RedTell thalassemia fields available, not analysed
6 Label conflicts	repeated sources spanning both classes	one field filed as Normal and Sickle, confirmed by eye; two further candidates

Table 4. The six checks, what each tests, and the evidence for it in this study. Each check is judged against its own corpus rather than a universal threshold.

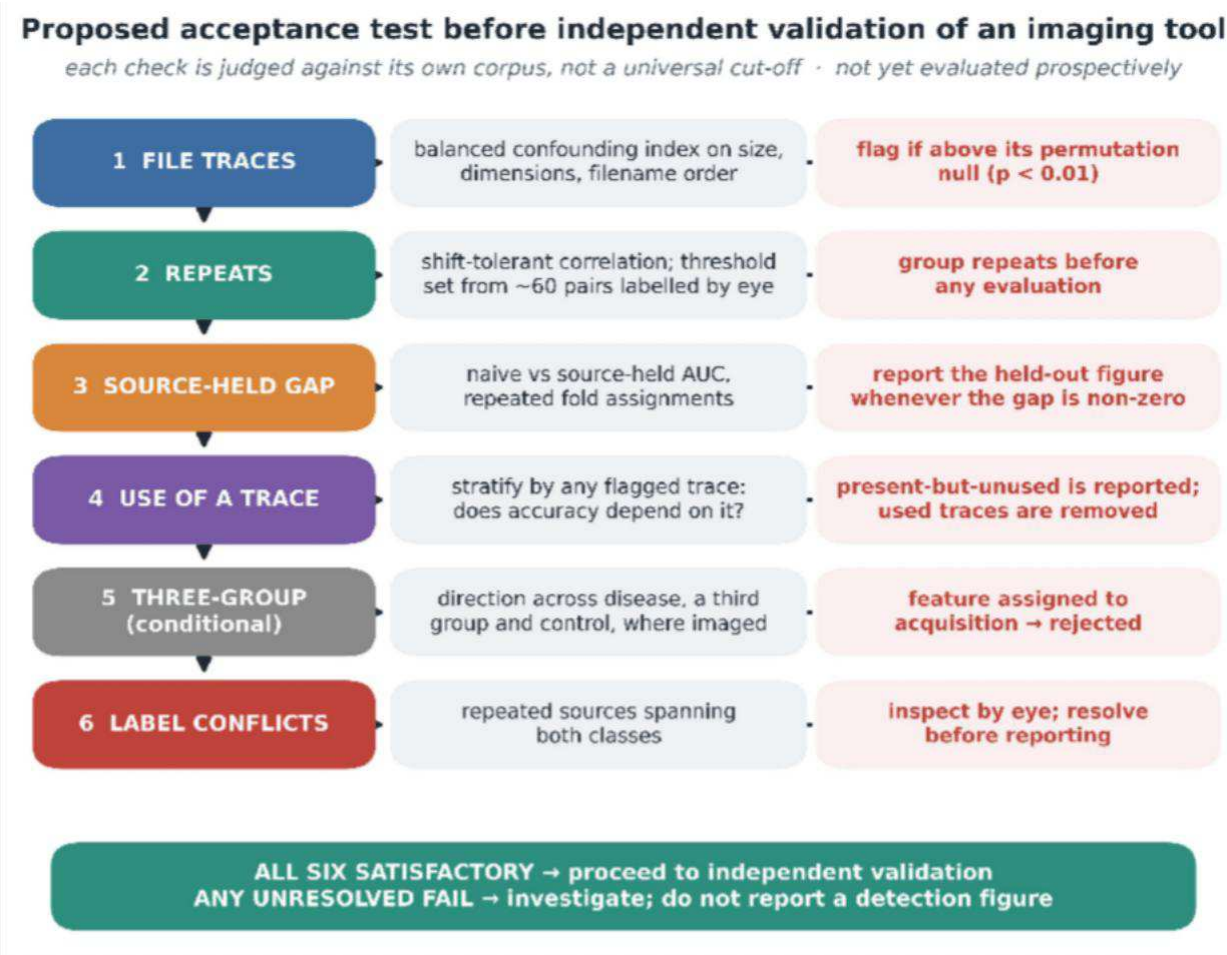


Fig. 9. The proposed acceptance test. Each check is judged against its own corpus rather than a universal threshold. Passing all six leads to independent validation, not directly to deployment, since the test has not been evaluated prospectively.

4 Discussion

Every public sickle cell corpus examined here that is not at ceiling reports a higher AUC under a random split than when its sources are kept apart, by 0.025 to 0.094 in area under the curve with grey features and a random forest, and by up to 0.111 across feature sets and learners. The size of

the difference depends on the model. Re-photographed fields still inflated a fine-tuned network by 0.021 on Tushabe, leaving its source-held error 1.8 times the naive one, whereas the byte-identical duplicates that inflated hand-crafted models on RedTell by 0.094 left pretrained networks unaf-fected. Inflation is therefore a property of a corpus and a model together, and must be measured for the model being reported. For a field in which reported accuracies cluster in the high nineties, that is the difference between a result that holds and one that does not. The differences are consistent with three forms of source dependence — duplicated images, re-photographed fields and shared partic-ipants — of which only the first two can be seen in the images; their individual contributions cannot be separated uniformly across the four corpora. No single mechanism is held to account for the dif-ferences reported here, and their size should not be assumed to transfer to other corpora, which dif-fer in how they were assembled.

That asymmetry is the main finding. Where each could be measured, the dependence visible in pix-els and the dependence visible only through identifiers were of the same order (0.094 on RedTell images, 0.092 on erythroSight morphology measurements), although the two come from different corpora and features and the second rests on six participants, and the second kind is invisible in the corpora that lack identifiers, which is almost all of them. Reporting standards ask authors to de-scribe how independence was secured [19, 21], but a user of these corpora cannot do so from the files. Distributing a participant identifier with each image would cost nothing and would make the question answerable. Data-quality assessment before model development has been argued for rou-tine health records [18]; the present results extend the case to imaging, where the defects that matter are repeated sources and drawn marks rather than missing fields.

Three further results bear on how such checks should be done. Thresholds for “the same image” must be set against real repeats, not synthetic ones: planted copies drift less than genuine re-cap-tures, and calibrating on them missed a third of the repeats here. Confounding measures based on accuracy are blind under class imbalance, and a balanced form should replace them. And a trace present in the files is not necessarily used: the resolution trace was visible to hand-crafted features and to a network, and unused by both, the drawn boxes were used, and only a stratified test distin-guished the two. Presence and use are different questions, and a confound check must ask both.

The practical consequence is for programmes that deploy such tools. Screening under the National Sickle Cell Anaemia Elimination Mission relies on point-of-care tests [3, 4], and imaging tools are being proposed as scalable aids alongside them [5–7]. The acceptance test proposed here needs only the corpus, standard computing resources and some sixty image pairs labelled by eye. It flags each of the problems reported in this paper, although it was assembled from them and has not yet been tested prospectively.

4.1 Limitations

- Participant-level inflation was measured on one corpus, with three participants per group, and its interval is wide.
- Source recovery is a lower bound: at the human-anchored threshold it recovered 85% of re-peats, and crops of 10% or more were largely missed (Supplementary Fig. S2).
- On isolated cell crops the procedure failed: erythrocytesIDB’s 413 distinct cells were merged into 338 sources, every inspected link being false. The method is suited to fields, not to single cells, and the erythrocytesIDB source-held estimate is therefore not interpreted.
- The network comparison used one architecture of each kind, three repeats of fine-tuning over ten epochs without a hyperparameter search, and two corpora. RedTell’s 61 images are

too few for a firm conclusion about networks, and larger or differently trained networks may memorise fields more or less than those tested.

- The reference study used 59 pairs sampled within similarity bins, which estimates the probability of a repeat at each similarity but not the prevalence of repeats in the corpus. Its second rater was an AI model, used for agreement only, and its conclusions rest on 52 decided pairs, so the interval around each proportion is wide (Table 2).
- The grouping threshold was calibrated on the corpus to which it is applied. It was set from judgements of image identity alone, and the estimate varies by less than 0.02 across thresholds from 0.45 to 0.648, but it has not been validated on independent data and should not be taken as universal; an externally calibrated threshold would be stronger.
- Two of the three label conflicts have not been confirmed by eye, and the acceptance test is a proposal: it has not been used in a district, and passing it has not been shown to predict field performance.

5 Conclusions

Across four public sickle cell imaging corpora, detection AUC under a random split exceeded AUC with sources held apart by up to 0.11 for hand-crafted models and by 0.021 for a fine-tuned net-work, a loss consistent with duplicated images, re-photographed fields and shared participants. Re-photographed fields inflated every model tested, duplicated images only the hand-crafted ones, so inflation must be measured for the model being reported. The part that comes from shared participants cannot be measured in corpora that distribute no identifiers. Thresholds for detecting repeats must be anchored to human judgement, confounding indices must be balanced, and a trace found in the files must be tested for use before it is blamed. The labelled release of one corpus carries drawn annotations worth 0.058 in AUC, and one field in that corpus is filed under both classes. Six checks, each calibrated to the corpus itself, are proposed as an acceptance test for imaging tools entering screening programmes, so that AI deployed in health systems is judged on what it measures rather than on how its data were assembled.

Abbreviations

AI Artificial intelligence

AUC Area under the receiver operating characteristic curve

CLAIM Checklist for Artificial Intelligence in Medical Imaging

CV Cross-validation

NSCEM National Sickle Cell Anaemia Elimination Mission

PROBAST+AI Prediction model Risk Of Bias ASsessment Tool, updated for AI

SCD Sickle cell disease

TRIPOD+AI Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, updated for AI

Acknowledgements

The authors thank the creators of the RedTell, Tushabe, erythrocytesIDB and erythroSight collections for making their data publicly available.

Author contributions

APJ conceptualised and designed the study, developed the analysis pipeline, performed the analyses and the reference labelling, and wrote the original draft. PR supervised the study, validated the findings and reviewed and edited the manuscript. Both authors read and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability

The four image collections analysed are publicly available from their original deposits; none is redistributed here, and each remains subject to the licence under which it was released. RedTell [22]: Zenodo, https://doi.org/10.5281/zenodo.7801430 (version 3, CC BY 4.0; anemia data archive, brightfield images), downloaded September 2025. Tushabe [23]: Kaggle, https://www.kaggle.com/datasets/florencetushabe/sickle-cell-disease-dataset; the analysed copy comprises the deposit's negative (Clear) folder and unlabelled positive folder, and the copy compared in Sect. 3.5 its labelled positive folder. erythrocytesIDB [24]: http://erythrocytesidb.uib.es/ (erythrocytesIDB1 subset), downloaded March 2026. erythroSight [25]: Federated Research Data Repository, https://doi.org/10.20383/103.0916 (per-cell morphology tables for the SCD, trait and normal participants), downloaded July 2026. The outputs behind every table and figure are archived with the code (see Code availability).

Code availability

All custom analysis code is openly available and archived with a permanent identifier at Zenodo (https://doi.org/10.5281/zenodo.22909501), with the development history at https://github.com/Jenifer87/scd-accuracy-inflation. The repository contains every script used to produce the results and figures reported here, together with the labels from the reference study (Sect. 2.3).

Declarations

Ethics approval and consent to participate. This study analysed publicly available, anonymised datasets and collected no new human data. The ethical oversight and consent arrangements for each source dataset are described in its original publication or repository documentation: Tushabe was collected with approval from the Clarke International University Research and Ethics Committee, Uganda (MRRH-2023-263) [23]; RedTell under protocols approved by the Medical Ethical Re-search Board of the University Medical Center Utrecht (15/426M) and the Ethical Committee of Clinical Investigations of Hospital Clinic, Spain [22]; and erythroSight under approved research protocols with consent for public deposit of de-identified images [25].

Clinical trial number. Not applicable. This study is not a clinical trial; it is a secondary analysis of publicly available, de-identified imaging datasets.

Consent for publication. Not applicable. The images are entirely unidentifiable and no individual details are reported.

Competing interests. The authors declare that they have no competing interests.

Supplementary figures

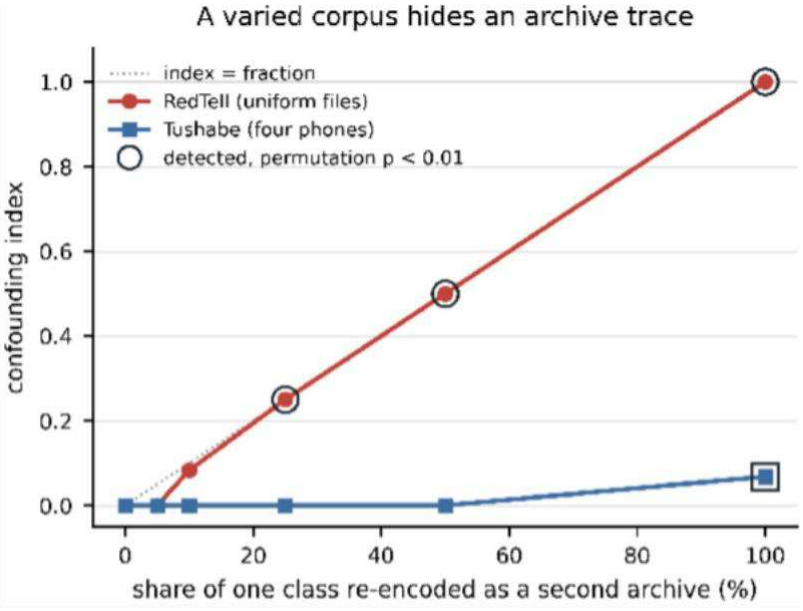


Fig. S1. Confounding index against the share of one class re-encoded as a second archive. On RedTell, with uniform files, the index equals the share and is detected from 25%; on Tushabe, photographed on four phones, the archive stays hidden until it makes up the whole class.

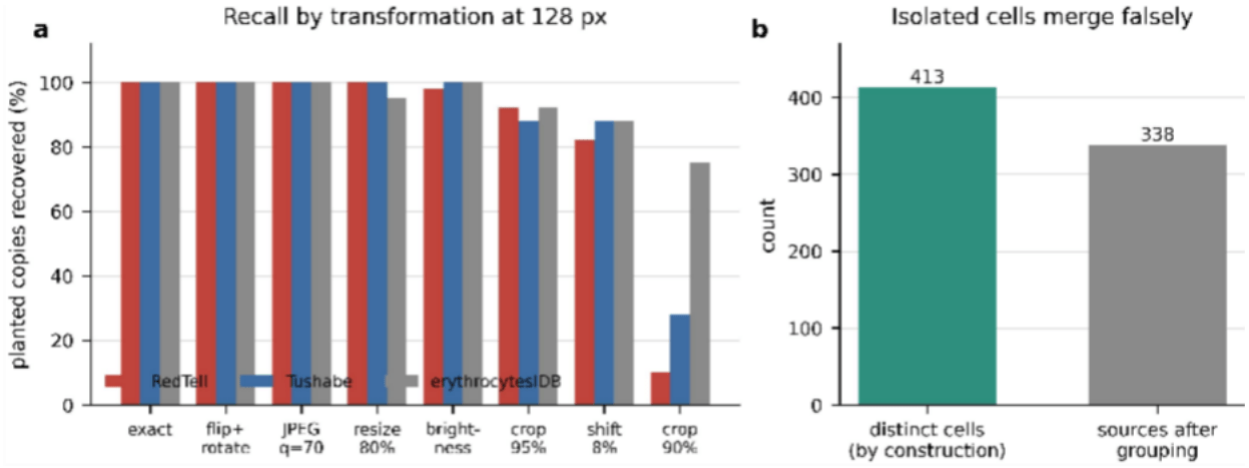


Fig. S2. Source recovery. (a) Planted copies recovered at 128 px, by transformation. (b) On isolated cells, distinct by construction, the procedure merges falsely.

References

1. Piel FB, Hay SI, Gupta S, Weatherall DJ, Williams TN. Global burden of sickle cell anaemia in children under five, 2010–2050: modelling based on demographics, excess mortality, and interventions. PLoS Med. 2013;10(7):e1001484.
2. Ware RE, de Montalembert M, Tshilolo L, Abboud MR. Sickle cell disease. Lancet. 2017;390(10091):311–23.
3. Ministry of Health and Family Welfare, Government of India. Operational guidelines for Na-tional Sickle Cell Anaemia Elimination Mission. New Delhi: MoHFW; 2023. https://sickle.nhm.gov.in/.
4. Steele C, Sinski A, Asibey J, et al. Point-of-care screening for sickle cell disease in low-resource settings: a multi-center evaluation of HemoTypeSC, a novel rapid test. Am J Hematol. 2019;94(1):39–45.
5. de Haan K, Koydemir HC, Rivenson Y, et al. Automated screening of sickle cells using a smartphone-based microscope and deep learning. NPJ Digit Med. 2020;3:76.

6. Manescu P, Bendkowski C, Claveau R, Elmi M, Brown BJ, Pawar V, Shaw MJ, Fernandez-Reyes D. A weakly supervised deep learning approach for detecting malaria and sickle cells in blood films. In: Martel AL, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science, vol 12265. Cham: Springer; 2020. p. 226–35.
7. Das PK, Dash A, Meher S. ACDSSNet: atrous convolution-based deep semantic segmentation network for efficient detection of sickle cell anemia. IEEE J Biomed Health Inform. 2024;28(10):5676–84.
8. Philomina Jenifer A, Nagarajan P. OptiSCDNet: an optimized network for sickle cell morphol-ogy analysis with integrated binary and severity classification. In: Proceedings of the Interna-tional Conference on NexGen Networks and Cybernetics (IC2NC). 2025. p. 838–42.
9. Jenifer AP, Rishi P. A training-free audit of acquisition confounding in medical image corpora, with a case study on public sickle cell blood films: data and code [dataset]. Zenodo. 2026. https://doi.org/10.5281/zenodo.22701097.
10. Zech JR, Badgeley MA, Liu M, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683.
11. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med. 2019;2:31.
12. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun. 2021;12:4423.
13. de Jong ED, Marcus E, Teuwen J. Current pathology foundation models are unrobust to medical center differences. arXiv:2501.18055. 2025.
14. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using ma-chine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3:199–217.
15. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med. 2022;5:48.
16. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804.
17. Barz B, Denzler J. Do we train on test data? Purging CIFAR of near-duplicates. J Imaging. 2020;6(6):41.
18. Phiri J, Zimba A, Njovu C, et al. A multi-facility data quality assessment of electronic health records for trustworthy AI in HIV public health in Zambia. Discov Artif Intell. 2026;6:1208. https://doi.org/10.1007/s44163-026-02293-x.
19. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
20. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024;6(4):e240300.
21. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505.
22. Sadafi A, Bordukova M, Makhro A, Navab N, Bogdanova A, Marr C. RedTell: an AI tool for interpretable analysis of red blood cell morphology. Front Physiol. 2023;14:1058720. https://doi.org/10.3389/fphys.2023.1058720.
23. Tushabe F, Mwesige S, Vicent K, Othieno E, Nsiimire E, Mutabazi P, Musani S, Areu D. A dataset of microscopic images of sickle and normal red blood cells. Acta Sci Microbiol. 2024;7(12):22–29. https://doi.org/10.31080/ASMI.2024.07.1453.
24. González-Hidalgo M, Guerrero-Peña FA, Herold-García S, Jaume-i-Capó A, Marrero-Fernández PD. Red blood cell cluster separation from digital images for use in sickle cell disease. IEEE J Biomed Health Inform. 2015;19(4):1514–25.

25. Shrestha P, Li H, Prakash M, Merkeley H, Kapoor V, Pande R, Stoeber B. erythroSight: open-access image dataset of blood cells from individuals with sickle cell disease and/or beta-thalassemia or without known hemaglobinopathies, captured using automated microscope Octopi [dataset]. Federated Research Data Repository; 2024. https://doi.org/10.20383/103.0916.
26. Kuglin CD, Hines DC. The phase correlation image alignment method. In: Proceedings of the IEEE International Conference on Cybernetics and Society. 1975. p. 163–5.
27. Reddy BS, Chatterji BN. An FFT-based technique for translation, rotation, and scale-invariant image registration. IEEE Trans Image Process. 1996;5(8):1266–71.
28. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
29. Roberts DR, Bahn V, Ciuti S, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913–29.
30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. p. 770–8. https://doi.org/10.1109/CVPR.2016.90.
31. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc. 1997;92(438):548–60.
32. Harrell FE. Regression modeling strategies. 2nd ed. Cham: Springer; 2015.
33. Steyerberg EW. Clinical prediction models. 2nd ed. Cham: Springer; 2019.
34. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. 2020. p. 151–9.
35. Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermo-scopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol. 2019;155(10):1135–41.
36. Northcutt CG, Athalye A, Mueller J. Pervasive label errors in test sets destabilize machine learning benchmarks. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. 2021.