Use case-focused metrics to evaluate machine learning for diseases involving parasite loads

Charles B. Delahunt
Noni Gachuhi
Matthew P. Horning
Global Health Labs, Bellevue, WA, 98007, USA

Abstract

Communal hill-climbing, via comparison of algorithm performances, can greatly accelerate ML research. However, it requires task-relevant metrics. For diseases involving parasite loads, e.g., malaria and neglected tropical diseases (NTDs) such as schistosomiasis, the metrics currently reported in ML papers (e.g., AUC, F1 score) are ill-suited to the clinical task. As a result, the hill-climbing system is not enabling progress towards solutions that address these dire illnesses.

Drawing on examples from malaria and NTDs, this paper highlights two gaps in current ML practice and proposes methods for improvement: (i) We describe aspects of ML development, and performance metrics in particular, that need to be firmly grounded in the clinical use case, and we offer methods for acquiring this domain knowledge. (ii) We describe in detail performance metrics to guide development of ML models for diseases involving parasite loads. We highlight the importance of a patient-level perspective, interpatient variability, false positive rates, limit of detection, and different types of error. We also discuss problems with ROC curves and AUC as commonly used in this context.

Keywords: Malaria, NTDs, metrics, sensitivity, specificity, limit of detection, false positive rate, ROC, AUC

1. Introduction

Communal hill-climbing, in which researchers compare performance of different algorithms using standardized metrics, is an established part of ML research culture, and has contributed to great advances on many tasks. A key prerequisite for the method is a set of task-relevant metrics [1].

Diseases involving parasite loads, that can be diagnosed by detecting parasites in microscopy images of a substrate (e.g., blood or filtered urine), are crucial global health challenges for which the hill-climbing system is stalled, due to widespread use of uninformative metrics. The resulting lack of communal progress towards deployable ML models prevents potential benefits of ML research from reaching the already underserved populations suffering from these diseases.

This paper seeks to enable the ML community to better serve these populations, by discussing appropriate metrics to guide ML research targeting illnesses such as (i) malaria (Giemsa-stained blood films); (ii) schistosomiasis, a neglected tropical disease or NTD (detection of helminth eggs from filtered stool or urine sample); (iii) filariasis and loiasis, both NTDs (Giemsa-stained blood films); (iv) more generally, any pathology where diagnosis is determined by the presence of a variable number of abnormal objects (e.g., pixels or cells in a histopathology slide).
This paper focuses on examples from microscopy diagnosis of malaria and helminths, because they clearly highlight the issues, and because they merit attention as serious, under-resourced health care problems [2; 3].

ML practitioners choose metrics to evaluate model performance by (i) what is customary, familiar, and convenient (e.g., object-level precision, recall, AUC, and F1 score); (ii) what has been done by previous authors; and (iii) what can generate the SOTA comparisons required for publication in the ML community. This creates a closed loop which perpetuates the use of certain metrics regardless of their effectiveness. When entrenched metrics do not assess algorithm performance in a clinically relevant way, it blocks progress towards deployable solutions to the clinical task.

For example, a recent review [4] describes serious problems in automated malaria detection papers: reported metrics are incomplete and not comparable between studies; metrics are object-based (not patient-based) and are thus not relevant to the clinical task; train and test sets contain objects from the same patient, which contradicts the patient-level focus; and datasets are too small. In addition, incorrect assumptions are built into algorithms: for example, diagnosis on thin blood films is common in ML papers, but contrary to clinical practice due to practical obstacles [5].

This paper seeks to accelerate the ML community’s progress towards deployable solutions, as follows: Section 2 details aspects of ML work that depend on a grasp of the clinical use-case (how the disease is diagnosed in the field), and offer tools to acquire this; and Section 3 describes effective ML metrics for diseases involving parasite loads.

2. The clinical use case
An ML solution fits into a larger, ML-independent, context. To be deployable, an algorithm must fit into an existing care structure* and meet or exceed existing clinical performance targets. So understanding these clinical constraints is a prerequisite for algorithm development (sometimes overlooked by ML practitioners [6]). This section discusses some crucial points to consider, tools for learning about use cases, and key issues relevant to these diseases.

2.1. Important domain specifics
Several domain-specific details are fundamental to effective algorithm development:

Basic facts about the clinical needs: For example, what are the proper uses of thick vs. thin blood films for malaria?

Performance metrics relevant in the field: Examples include patient-level sensitivity and specificity, and limit of detection (LoD). This knowledge enables ML researchers to tailor salient metrics to guide algorithm development, define objective functions, do internal assessment, and report results.

Performance specifications for deployment: Clinicians are unwilling to reduce patient care standards, so ML models must perform at least as well as current practice to be deployable. Field performance requirements are thus vital concerns, even if a particular model iteration does not attain them.

Domain-specific obstacles and shortcuts: Some difficult details need special treatment, and others allow for valuable shortcuts. For example, malaria parasites can exist at various depths of a thick blood film, so a single image plane will not capture all parasites in focus. On the plus side, the nuclei of

---

* We set aside the complex case of a disruptive technology potentially altering existing care protocols. In such cases, the protocols and the impact of any alterations should be analyzed.
white blood cells (WBCs) are always present in thin films, and they stain similarly to malaria parasite nuclei. So they give a color reference for the rare (or absent) parasites.

Shortcuts matter because generic methods applied as-is are unlikely to hit clinical performance requirements - a harder task than outdoing another generic method in a SOTA comparison.

**How annotations and training sets should be structured:** For example, malaria ring stages typically have both a round nucleus and a crescent-shaped cytoplasm; however, after drug treatment the rings often lack the cytoplasm, and appear in thin films as dark round dots, very similar to a common distractor type. So they have outsized impact on decision boundaries and require special care as to annotation and inclusion in training sets (see e.g., [7]).

### 2.2. Connecting to domain expertise

Avenues to acquire this vital domain expertise include (i) sourcing documentation and (ii) connecting with field experts.

**Documentation** of use cases and standards of care are published by various agencies, including: (i) the World Health Organization (WHO), e.g., [5; 8; 9]; (ii) ministries of health, e.g., [10; 11; 12]; (iii) funding agencies, e.g., the BMGF [3] and the Global Fund [13]; and (iv) various NGOs, e.g., WWARN [14].

Some ML papers (e.g., [15; 16]) cite relevant documents, but this is not (yet) common practice. The crucial point is that literature review should extend well beyond ML methods to focus on the clinical use-case itself, without an ML-centric filter.

**Field experts** - those who work in field clinics or who do field-based research - and **subject matter experts** - such as WHO personnel and long-time researchers in the space - are a priceless source of guidance and collaboration. The value of their experience and insight cannot be overstated.

Connecting with such experts is made easier by two things. First, people (on average) love to talk about their work. Second, field experts are often (again, on average) open to engaging with ML solutions and happy to co-author serious research.

Sources for contacts include: (i) published work, e.g., who is leading and authoring/co-authoring relevant studies; (ii) academic institutions with concentrations of research in the space; (iii) online interest groups, e.g., on LinkedIn (see Appendix); and (iv) conferences (non-ML-focused), their attendees, and proceedings, e.g., ASTMH [17]. Further examples are given in the Appendix.

Understanding the use-case needs, by documentation and especially (when possible) by collaborating with field experts, grounds the ML work in reality, a necessary condition for progress towards deployable solutions.

### 3. Salient metrics

This section outlines several metrics relevant to ML models that target diseases involving parasite loads.

#### 3.1. Patient level metrics

The importance of assessing algorithm performance at the patient level cannot be over-emphasized. The basic unit of clinical care is the patient*, so the most relevant metrics are defined at the patient level, not the object level. Performance assessed across pooled objects can be a useful intermediate step during ML development, but it is fundamentally unrealistic, because (i) it does not match the clinical task; (ii) it ignores interpatient variability; and (iii) it is dominated by high parasitemia samples. For example, consider four

---

* We set aside population-level diagnostics such as for Vitamin A deficiency [18].
malaria-positive patients, with
Patient 1: 50,000 parasites/µL (p/µL).
Patients 2, 3, 4: each 300 p/µL
Suppose the algorithm detects all parasites in {1}, and misses all parasites in {2,3,4}. Then the object-level sensitivity is 98%, while patient-level sensitivity is 25%.

Two metrics are particularly useful: false positive rate (FPR) and sensitivity. Each is calculated separately for each patient, using performance on objects within that patient’s sample. These are covered in 3.2 and 3.3, and underpin other metrics related to specificity, LoD, and quantitation (3.4, 3.5, and 3.7).

Interpatient variability (as in Figure 1) is a crucial issue that causes difficulty for ML, so it must be factored into algorithm evaluation. It is captured by the standard deviations of FPR and sensitivity (cf. 3.2, 3.3), to the degree that the dataset captures interpatient diversity.

A related issue is interclinic variability. For example, clinics can use different stain variants (e.g., Giemsa, Field, and JSB). Even clinics with nominally identical protocols can differ substantially, see e.g., [19] and a detailed example in [20]. Besides variations in presentation, different clinics may produce populations of samples with differently distributed FPRs and sensitivities. Implications of this on tuning algorithms are covered in 3.4.

3.2. FP Rate

False positive rate (FPR) is the number of distractors mislabeled as parasites per clinically relevant unit of substrate, hereafter cV, e.g., 1 µL of blood (malaria), 10 mL urine (Schistosoma haematobium), 1 gram stool (other Schisto), a specified number of cells in a histological sample, etc; but not “per image tile”, which generally has no clinical relevance*. Malaria ML papers with some FPR analysis include [21; 22; 23; 24]. We denote the vector of FPRs for the population of patients as \( F \).

FPR is not object-level specificity, which is a commonly reported but highly flawed measure in this context (see 3.9).

FPR can be calculated for any sample, but FPRs on positive samples may be erroneously boosted by mis- or unannotated parasites. Thus, the characteristics of the population’s FPR distribution are best determined on negative samples only.

Interpatient variability makes the standard deviation of FPR, \( \sigma(F) \), a crucial performance measure. The mean FPR, \( \mu(F) \) is not relevant because it can be subtracted out. However, since it tends to scale roughly with \( \sigma(F) \), it can give a hint as to the relative magnitude of \( \sigma(F) \) (see e.g., [23; 24]).

In datasets with insufficient numbers of patients, an FPR calculated over pooled objects has some value as a lower bound on \( F \).
In particular, it can be compared to the clinical LoD requirement. For example, a pooled-object FPR of 5,000/µL, vs. a required 100 p/µL LoD (malaria), is a clear sign of work still to do. Multiple splits of a set of pooled objects does not simulate \( \sigma(F) \), because each split will include the full patient diversity.

Aside: Samples with high FPRs are sometimes criticized as due to “poor sample preparation”. However, this is in the eye of the beholder: human clinicians readily and successfully diagnose “dirty” samples, while ML algorithms often fail. Thus, the need to “improve” sample prep is in fact a need to accommodate ML’s difficulty with handling highly variable sample artifacts. See [19], and a detailed example in [20].

* Image tiles can sometimes be translated into microscopy “Fields of View” used in protocols.
3.3. Sensitivity

Object level sensitivity (aka recall) is the fraction of parasites in a positive sample that are correctly labeled (e.g., by means of an object score threshold C):

$$\text{Sensitivity} = \frac{tp}{tp + fn}$$

where $tp = \text{true positives}$, i.e. parasites labeled correctly, and $fn = \text{false negatives}$, i.e. parasites labeled as distractors (or missed), in the examined volume $V$. There is no constraint on the size of $V$ or parasitemia, but sensitivities for patients with very few parasites are less reliable (cf. law of large numbers). Sensitivity is calculated for each patient separately. We denote the vector of the collected sensitivities for the population of positive patients as $S$. $S$ underpins metrics related to LoD and quantitation (3.5 and 3.7). For the rest of this subsection we consider patient-level sensitivity only (not $S$).

Patient-level sensitivity is crucial clinically, but is complex to interpret because it depends on two things:

(i) The particular parasitemia distribution of the tested set: Patients with low parasitemias (close to the LoD) are harder to identify. In malaria for example (LoD ≈ 100 p/µL), if all patients have parasitemias $> 1000$ p/µL, 100% sensitivity is (hopefully) trivial, while if all parasitemias are under 50 p/µL, very low sensitivity is likely.

(ii) The particular specificity: Sensitivity and specificity are paired, as seen in ROC curve operating points.

Thus, reporting patient-level sensitivity is uninformative without also reporting (i) the parasitemia distribution, and (ii) the associated specificity. A principled way to maximize patient-level sensitivity is given in 3.6.

3.4. Specificity

Specificity is the fraction of negative items (e.g., distractor objects or patients) that are correctly diagnosed as negative:

$$\text{Specificity} = \frac{tn}{tn + fp}$$

where $tn = \text{true negatives}$ (negative items in $V$ labeled correctly), and $fp = \text{false positives}$ (negative items labeled incorrectly).

Object-level specificity, even if calculated for each patient separately, has little usefulness and can be deceptive (see 3.9).

Patient-level specificity is highly salient. Clinical goals of high specificity include not overwhelming the health care system, avoiding excess treatments, and preventing misattribution. Thus, clinical use-cases generally require a high specificity (e.g. 90% for malaria diagnosis [8], 97.5% for schistosomiasis [9]).

Specificity is closely tied to FPR and can be readily tuned for an algorithm that labels objects: Suppose that objects have been detected then labeled by some method (e.g.,
a threshold \( C \) on object scores), that \( F \) is gaussian, and that patient diagnosis is determined by a threshold \( T \) on the number of positively-labeled objects per \( cV \) (i.e. a standard “detect, classify, count, then threshold” approach). To attain a target specificity \( K \), one can set

\[
T = \mu(F) + \alpha \sigma(F) \text{ where } \alpha \text{ is found (1)}
\]

via the (one-sided) error function and \( K \). Alternate formulations for the (likely) case of non-gaussian \( F \) are given in the Appendix.

Negative samples are easier to obtain and trivial to annotate (assuming accurate patient-level ground truth), and specificity depends only on negative samples. So \( T \) can ideally be tuned on a separate, dedicated validation set of negatives that capture a sufficient range of FPRs (both “dirty” and “clean” samples).

Different clinics can have widely different FPR distributions \( F \). Because \( \sigma(F) \) determines both specificity and LoD (cf. 3.5), different clinics may require different hyperparameters to hit the target patient specificity \( K \), leading to different LoDs. Thus, tuning an algorithm for deployment may involve multiple validation sets of negatives (by clinic), trading off low specificity at some clinics vs higher LoD at others.

### 3.5. Limit of detection (LoD)

Here, LoD roughly means the parasitemia at which the algorithm can consistently (e.g., 95% of cases) distinguish positive and negative cases. Examples: The required LoD for malaria microscopy is roughly 100 p/µL [5], i.e. 1 parasite per 50,000 Red Blood Cells (RBCs). For helminths, LoD is implicitly 1 egg (per 10 mL urine or 1 gram stool) [9].

LoD can be directly probed using holdout sets of low parasitemia positive samples. These are not as useful for training anyway, as they supply few parasite objects. However, it’s hard to acquire enough samples near the LoD.

Alternatively, an estimate of LoD can be calculated from \( F \) and \( S \), if a patient is ruled “positive” when \( N \geq T \), where \( N \) is the number of positively-labeled objects per \( cV \) (cf. 3.4). \( N = TP + FP \) in positive patients, and \( N = FP \) in negative patients. Here \( TP \) and \( FP \) denote counts per \( cV \), so \( TP = tp \frac{cV}{V} \) where \( tp \) is the number of parasites correctly labeled in \( V \) (similarly \( FP = fp \frac{cV}{V} \)).

To estimate LoD (as \( L \) parasites per \( cV \)), we can proceed as follows:

First set \( T \) high enough to ensure accuracy on negative samples (as in [23]): similarly to Eqn 1 (or variants in the Appendix), set

\[
T = \mu(F) + 1.65 \sigma(F) \quad (2)
\]

(i.e. set \( \alpha \) in Eqn 1 to 1.65 std devs above the mean, targeting 95% specificity).

For positive samples, the worst case is a very “clean” sample with low FPR, such as the 5th percentile of samples with \( FP = \mu(F) - 1.65 \sigma(F) \). Then we must depend mostly on detected parasites to ensure \( N \geq T \) for a positive diagnosis. Suppose for ease that the sample has average sensitivity = \( \mu(S) \). Then at LoD, \( TP = L \mu(S) \).

To correctly diagnose a positive sample at the LoD (just barely, so \( N = T \)), we need

\[
N = TP + FP = L \mu(S) + \mu(F) - 1.65 \sigma(F) = T = \mu(F) + 1.65 \sigma(F)
\]

\[
\Rightarrow L \mu(S) = 3.3 \sigma(F)
\]

So the estimated LoD (\( L \) per \( cV \)) has

\[
L = \frac{3.3 \sigma(F)}{\mu(S)} \quad (3)
\]

We have found this estimate to be a good (slightly optimistic) proxy for actual LoD when assessing algorithms during development. It has the practical advantage that
low parasitemia samples are unnecessary, because the vector $S$ can be well characterized by high parasitemia samples. It also allows useful comparison of algorithms, as it directly addresses a key clinical requirement and is anchored to the relevant unit $cV$.

A more nuanced (and pessimistic) proxy could account for $\sigma(S)$ by having a denominator $= \mu(S) - \beta \sigma(S)$ for some $\beta$.

### 3.6. Choosing operating points

Given a trained algorithm that uses the two hyperparameters $C$ and $T$, $\{C, T\}$ can be optimized in a principled way to maximize patient-level sensitivity, subject to the constraint of a fixed target specificity $K$:

- Set aside a validation set of negative samples. If there are sufficient positive samples to spare, include these also.
- For each $C$, calculate $F$ over the validation negatives, and $\mu(S)$ over the validation positives if available, or (less ideal but workable) over the training set positives.
- For each $C$, determine $T = T(C, K, F)$ which hits the target specificity $K$ on the validation negatives, as in 3.4.
- For each $C$, estimate LoD as in 3.5. Select the $C$ with the lowest LoD.
- Use this $\{C, T\}$ pair as algorithm hyperparameters to process test sets, and report patient-level specificity and sensitivity.

### 3.7. Quantitation

Quantitation sometimes has clinical importance. For example, accurate quantitation is needed to monitor for drug-resistant malaria strains by calculating clearance curves [25; 26; 27]. For helminths, quantitation targets are typically rough only (e.g., low, medium, high) [9]. Quantitation accuracy should be reported at the patient level due to high interpatient variability, e.g., by Bland-Altman plot. When parasitemias range over orders of magnitude (as in malaria and helminths), $R^2$ values are skewed by $L_2$ norm effects (fitting the log($P$) values can mitigate this).

We can estimate the parasitemia $\hat{P}$ for a given patient by (cf. [24])

$$\hat{P} = n \left( \frac{cV}{V} \right) - \hat{F} \hat{S},$$

where

$n = \text{number of alleged parasites found in } V,$
$\hat{F} = \text{expected FPR (e.g., } \mu(F)),$
$\hat{S} = \text{expected sensitivity (e.g., } \mu(S)),$
$cV = \text{clinically relevant volume of substrate},$
$V = \text{estimate of the volume examined}.$

Three types of error affect Eqn 4: irreducible Poisson, estimates of examined volume, and counts of alleged parasites.

*Irreducible Poisson error* is discussed in 3.8, and S.I. of [24].

*Examined volume error* (error in estimating $V$) impacts quantitation accuracy via the $\frac{cV}{V}$ term of Eqn. 4. For example, thick film blood volume $V$ is typically estimated by counting WBCs [5]. Any error in the WBC count causes proportional quantitation error. This error type can be compartmentalized as follows: (i) manually count WBCs on a test set, ensuring oracle $V$ estimates; and (ii) separately report the patient-level error statistics of the WBC counter.

*Parasite counting errors* stem from patient-level variations in sensitivity and FPR: The number of alleged parasites per $cV$ in the sample is $(tp + fp) \frac{cV}{V} = TP + FP$.

The difference between $TP$ and $SP$ (the expected number of correctly labeled true parasites, where $P$ is the true parasite count per $cV$) is due to deviation of the sample’s sensitivity from the expected $\hat{S}$. Similarly, the difference between $FP$ and $\hat{F}$ (the expected FPR) is due to the deviation of this sample’s FPR from expected. $\sigma(S)$ and $\sigma(F)$ quantify these deviations over the population. A figure of merit to assess parasite counting error, derived and discussed in
[24], is thus
\[
\frac{\sigma(S)}{\mu(S)} + \frac{\sigma(F)}{\mu(S)} \frac{1}{P}
\]  
(5)

While the FPR term is usually hardest to control, it also shrinks as $1/P$, so for large $P$ the sensitivity term dominates.

We note that ground truth parasitemia estimates are also subject to these error types, which complicates assessment of a model’s quantitation accuracy.

3.8. Effect of Poisson statistics

For a fixed nominal parasitemia $P$, Poisson statistics for rare events give variation in the number of parasites found in a given sample with volume $V$. This has two impacts:

(i) For diagnosis, a low LoD requires that a large volume $V$ be examined to ensure that at least a couple true parasites are present. Otherwise, a statistically predictable subset of samples from positive patients will contain 0 parasites, reducing patient-level sensitivity from the start.

(ii) For quantitation, a sufficiently high volume $V$ (depending on $P$) must be examined to control irreducible error (S.I. of [24]).

In both cases automated systems can hold a strong advantage via scanning higher volumes than human technicians, who often by necessity work in a high Poisson error regime (S.I. of [24]).

When reporting results on datasets of small size, authors should understand how Poisson variability limits their estimates of algorithm performance. For example, some thin blood film datasets contain in total less than the minimum amount needed to diagnose one patient.

3.9. Problems with ROCs and AUCs

Object-level ROC curves, and the associated Area Under Curve (AUC), are routinely reported by ML research papers involving parasite detection. However, these are misleading indicators of performance in this context (they can be useful for internal algorithm work).

First, they do not address the clinical need for patient-centric care. In particular, they ignore the crucial matter of patient-level variability ($\sigma(F)$).

Second, there is often a large imbalance between distractors and positive objects, especially in samples near clinical LoD. A common situation is a model that diagnoses malaria on thin films by labeling individual RBCs as infected or not. 5 million RBCs/$\mu$L vs. 100 p/$\mu$L at LoD gives 50,000 negative objects for each positive object, so a 0.999 AUC can coexist with an average of 50 FPs per parasite at LoD (a very low SNR). Since one detected parasite and one FP object have equal impact on diagnosis (as determined by exceeding a threshold $T$), FP noise will swamp the diagnostic signal of detected parasites. In such cases with large class imbalance (say $D:1$), the leftmost $\frac{1}{D}$ vertical sliver of the ROC curve, with $y$-axis rescaled to be full width, reflects a more meaningful (and more sobering) ROC, because this expanded sliver visually weights TP counts and FP counts equally, as shown in Figure 2.

![Figure 2: For a 20:1 distractor-to-parasite ratio, stretching the left vertical sliver gives a more meaningful ROC curve. Diagonal red lines show operating points that give equal numbers of TPs and FPs.](image)
Third, the object-level ROC curve is ill-defined, because it depends on how distractors are defined. For example, when using thick films to diagnose malaria, “distractor” can mean only the most difficult objects that closely resemble parasites; or any dark blob; or even every pixel in an image. Figure 3 shows an example in which considering only “difficult” distractors (top) results in a low AUC, while considering additional, mostly “easy” distractors (bottom) gives a higher AUC with no change in actual performance as measured by FPR.

Patient-level ROCs can give a useful sense of algorithm behavior near the clinical performance requirements. However, clinically irrelevant operating points dilute their AUC as a performance metric.

4. Discussion

The illnesses addressed here are targets amenable to ML methods, and successful development of deployable ML solutions would yield tremendous benefits for currently underserved populations.

Unfortunately, communal hill-climbing has not yet been effectively set up for these targets due to use of incorrect comparison metrics. As a result, the synergistic power of the ML community is not being applied with full force.

Individual ML research teams can improve the situation by grounding their ML work in an understanding of the use case, and tailoring metrics to the clinical needs. We have described such metrics here: variation in FPR, per-patient sensitivity, LoD, patient-level sensitivity and specificity, and a figure of merit for quantitation.

Peer reviewers play a special role in determining the success or failure of hill-climbing: (i) reviewers can assess algorithms and performance results according to whether they incorporate the requirements of the clinical use case. (ii) when authors present new metrics, well-grounded in the use-case, this can be more valuable than a comparison based on customary but inferior metrics. By recognizing when this is the case, reviewers can disrupt the cycle that perpetuates a counter-productive status quo.

With attention to the clinical use case and deliberate choice of metrics, the ML community can better position itself to deliver concrete benefit to the populations suffering the dire effects of the illnesses considered here.
Acknowledgments

Funding provided by Global Health Labs, Inc. (www.ghlabs.org)

References

[1] L. Maier-Hein, A. Reinke, and et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv, 2022. URL https://arxiv.org/abs/2206.01653.

[2] World malaria report 2019. World Health Organization, Geneva, Switzerland, 2019.

[3] Bill and Melinda Gates Foundation. neglected tropical diseases. URL https://www.gatesfoundation.org/our-work/programs/global-health/neglected-tropical-diseases.

[4] M. Poostchi, K. Silamut, R. Maude, S. Jaeger, and G. Thoma. Image analysis and machine learning for detecting malaria. Translational Research, 2018.

[5] Basic Malaria Microscopy: Tutor’s guide. World Health Organization, Geneva, Switzerland, 2010.

[6] D. Koller and Y. Bengio. A fireside chat with Daphne Koller at ICLR, 2018. URL https://www.youtube.com/watch?v=N4mdV1CIpvI.

[7] A. Ng. Practical limitations of today’s deep learning in healthcare. ML4H conference, 2020. URL https://nips.cc/virtual/2020/19127.

[8] Malaria Microscopy Quality Assurance Manual V2. World Health Organization, Geneva, Switzerland, 2016.

[9] Diagnostic target product profiles for monitoring, evaluation and surveillance of schistosomiasis control programmes. World Health Organization, Geneva, Switzerland, 2021.

[10] Manual de Procedimientos de Laboratorio Para el Diagnostico de Malaria. Ministerio de Salud, Peru, Lima, Peru, 2003.

[11] Strategic plan for control of leishmaniasis 2021-2025. Ministry of Health, Kenya, Nairobi, Kenya, 2020. URL www.health.go.ke.

[12] Center for Disease Control. Treatment of malaria, guidelines for clinicians. URL https://www.cdc.gov/malaria/diagnosis_treatment/clinicians2.html.

[13] The Global Fund. URL https://www.theglobalfund.org/en/.

[14] WorldWide Antimalarial Resistance Network. URL https://www.wwarn.org/.

[15] M.P. Horning, C.B. Delahunt, C.M. Bachman, and et al. Performance of a fully-automated system on a WHO malaria microscopy evaluation slide set. Malaria J, 2021.

[16] P. Oyibo, S. Jujjavarapu, J.C. Diehl, and et al. Schistoscope: An automated microscope with artificial intelligence for detection of schistosoma haemato-bium eggs in resource-limited settings. Micromachines, 2022. URL https://www.mdpi.com/2072-666X/13/5/643.

[17] American Society of Tropical Medicine and Hygiene. URL https://www.astmh.org/annual-meeting.

[18] Serum retinol concentrations for determining the prevalence of vitamin A deficiency in populations. World Health Organization, Geneva, Switzerland, 2011. (WHO/NMH/NHD/MNM/11.3).
[19] D. Das, R. Vongpromed, M. Dhorda, and et al. Field evaluation of the diagnostic performance of easyscan go: a digital malaria microscopy device based on machine-learning. Malaria J, 2022.

[20] K. Torres, C.M. Bachman, V.D. Gamboa, and et al. Automated microscopy for routine malaria diagnosis: a field comparison on Giemsa-stained blood films in Peru. Malaria J, 2018.

[21] N. Linder and et al. A malaria diagnostic tool based on computer vision screening and visualization of plasmodium falciparum candidate areas in digitized blood smears. PLoS One, 2014.

[22] P. Manescu, M.J. Shaw, M. Elmi, and et al. Expert-level automated malaria diagnosis on routine blood films with deep neural networks. American Journal of Hematology, 2020.

[23] C. Mehanian, M. Jaiswal, C. Delahunt, C. Thompson, M. Horning, L. Hu, and et al. Computer-automated malaria diagnosis and quantitation using convolutional neural networks. ICCV, 2017.

[24] C.B. Delahunt, M.S. Jaiswal, C. Mehanian, and et al. Fully-automated patient-level malaria assessment on field-prepared thin blood film microscopy images. IEEE GHTC, 2019. URL https://arxiv.org/abs/1908.01901. (with S.I.).

[25] E.A. Ashley, M. Dhorda, R.M. Fairhurst, C. Amaratunga, and et al. Spread of artemisinin resistance in Plasmodium falciparum malaria. New England J of Medicine, 2014.

[26] N. White. The parasite clearance curve. Malaria J, 2011.

[27] Malaria Microscopy Standard Operating Procedure MM-SOP-09: Malaria Parasite Counting. World Health Organization, Geneva, Switzerland, 2016.

Appendix A.

Connecting with domain expertise

Other examples of resources include:

The Presidents Malaria Initiative, https://www.pmi.gov/resources/.

The Malaria Consortium https://www.malariaconsortium.org/projects/projects-database/

PSI (Population Services International), https://www.psi.org/practice-area/malaria/.

LinkedIn, Neglected Tropical Diseases Interest Group, https://www.linkedin.com/groups/2060379/

Appendix B.

Setting a threshold for diagnosis

The methods for setting $T$ in 3.4 and for estimating LoD in 3.5 both assume that the FPR vector $F$ is gaussian. In our experience this may not be the case. Rather, the FPR distribution is often asymmetrical, with mostly low-FPR samples and a few high-FPR samples. This can be handled by modifying the methods in 3.4 and 3.5:

1. For $\mu(F)$, use the median of $F$ instead of the mean of $F$. Similarly, if the vector $S$ is non-gaussian, the median can be used instead of the mean for $\mu(S)$.

2. For $\sigma(F)$, use one-sided std devs, which are calculated on a symmetric distribution created by reflecting the points to the right (or left) of the median across the median as centerpoint. This gives a large right std dev $\sigma_R(F)$ and a small left std dev $\sigma_L(F)$.
Then the new version of Eqn 1 is

\[ T = \text{median}(F) + \alpha \sigma_R(F) \quad (6) \]

The new version of Eqn 3 is

\[ L = \frac{1.65 \left( \sigma_L(F) + \sigma_R(F) \right)}{\mu(S)} \quad (7) \]

Two other methods of calculating \( T \) from \( F \) may be useful:

(i) Set \( T \) based on the \( K^{th} \) percentile of \( F \);
(ii) manually choose \( T \) based on a scatterplot of the \( FP \) counts in the validation negative samples.

As before, the detected objects are assumed to be already classified. If a threshold \( C \) on object scores is used to classify objects, then first \( T \) must be calculated for each \( C \), before choosing the best \( \{C, T\} \) pair, as outlined in 3.6.

The manual method of choosing \( \{C, T\} \) takes time, but it can yield the best results in a field deployment because it is most closely tailored to the empirical FPR distribution.