frontiers | Frontiers in Malaria

TYPE Methods
PUBLISHED 17 April 2024
DOI 10.3389/fmala.2024.1250220

OPEN ACCESS

EDITED BY
Anita Ghansah,
Noguchi Memorial Institute for Medical Research, Ghana

REVIEWED BY
Geoffrey H. Siwo,
University of Michigan, United States
Sonal Kale,
National Institute of Allergy and Infectious Diseases (NIH), United States

*CORRESPONDENCE
Charles B. Delahunt
charles.delahunt@ghlabs.org

RECEIVED 29 June 2023
ACCEPTED 18 March 2024
PUBLISHED 17 April 2024

CITATION
Delahunt CB, Gachuhi N and Horning MP (2024) Metrics to guide development of machine learning algorithms for malaria diagnosis.
Front. Malar. 2:1250220.
doi: 10.3389/fmala.2024.1250220

Metrics to guide development of machine learning algorithms for malaria diagnosis

Charles B. Delahunt*, Noni Gachuhi and Matthew P. Horning

Global Health Labs, Bellevue, WA, United States

Automated malaria diagnosis is a difficult but high-value target for machine learning (ML), and effective algorithms could save many thousands of children's lives. However, current ML efforts largely neglect crucial use case constraints and are thus not clinically useful. Two factors in particular are crucial to developing algorithms translatable to clinical field settings: (i) clear understanding of the clinical needs that ML solutions must accommodate; and (ii) task-relevant metrics for guiding and evaluating ML models. Neglect of these factors has seriously hampered past ML work on malaria, because the resulting algorithms do not align with clinical needs. In this paper we address these two issues in the context of automated malaria diagnosis via microscopy on Giemsa-stained blood films. The intended audience are ML researchers as well as anyone evaluating the performance of ML models for malaria. First, we describe why domain expertise is crucial to effectively apply ML to malaria, and list technical documents and other resources that provide this domain knowledge. Second, we detail performance metrics tailored to the clinical requirements of malaria diagnosis, to guide development of ML models and evaluate model performance through the lens of clinical needs (versus a generic ML lens). 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 reasons why ROC curves, AUC, and F1, as commonly used in ML work, are poorly suited to this context. These findings also apply to other diseases involving parasite loads, including neglected tropical diseases (NTDs) such as schistosomiasis.

KEYWORDS

malaria, NTDs, schistosomiasis, metrics, machine learning, sensitivity, specificity, limit of detection

1 Introduction

Malaria and some neglected tropical diseases (e.g., schistosomiasis) involve parasite loads that can be detected in microscopy images of a substrate (e.g., blood or filtered urine). They are thus amenable, though difficult, targets for automated diagnosis via machine learning (ML) methods. These diseases are also very high-value ML targets: They are

serious global health challenges affecting hundreds of millions of people, especially children, in underserved populations (WHO, 2019; BMGF, 2023; WHO, 2023). Because microscopy on Giemsa-stained blood films is a widespread clinical diagnostic, effective automated ML systems have great potential benefit since they can naturally fit into clinical workflow and enable hard-pressed clinics to treat more patients. In addition, drug resistance sentinel sites have a heavy demand for parasite quantitation on Giemsa-stained blood films, a use case well-suited to automated microscopy.

However, ML methods developed for malaria diagnosis[1] using Giemsa-stained blood films have so far largely failed to translate to useful deployment, for several reasons.

(i) The task is difficult: malaria parasites are small and closely resemble certain artifact types; field blood films are highly variable in stain color, types and numbers of artifacts, and parasite appearance; digital microscopy images vary in color, quality, and resolution; images are often full of distractor objects; and the low limits of detection required for clinical use result in low signal-to-noise ratios (e.g., one parasite per 30 large fields of view).

(ii) ML development has typically proceeded in a heavily ML-centric mindset, without careful attention to (or even knowledge of) the domain specifics, use cases, and clinical requirements of malaria. This yields algorithms that, almost by design, fail to meet clinical needs and cannot be built upon (see Figure 1).

(iii) ML development can only optimize what is measured, so a crucial prerequisite for successful development is a set of task-relevant metrics (Maier-Hein et al., 2022; Reinke and Tizabi, 2024). These tailored metrics have largely been lacking for malaria, for which ML development has instead been guided by generic and ill-suited ML metrics such as object-level ROC curves.

This paper seeks to accelerate the ML community's progress toward translatable solutions for malaria diagnosis, by describing tools and techniques which we have found to be essential for development of clinically effective ML algorithms. The intended audience are ML researchers as well as anyone evaluating the performance of ML models for malaria. It captures lessons learned by our group over a decade of applying ML to malaria diagnosis. The resulting algorithms (Delahunt et al., 2014a; Mehanian et al., 2017; Delahunt et al., 2019) are, to our knowledge, the most effective and also the most extensively field tested in clinical trials (Torres et al., 2018; Vongpromek et al., 2019; Horning et al., 2021; Das et al., 2022; Rees-Channer et al., 2023) yet built for fully automated diagnosis of malaria on Giemsa-stained blood films in clinical settings (we note that in medicine the gold standard of evidence is the third-party clinical trial, not the ML-style comparison). These field trials showed that our algorithms, though state-of-the-art, still fall short of the clinical demands, and highlight the need for more robust algorithms to truly impact this category of malaria diagnosis.

The paper is structured as follows: Section 2 details aspects of ML work that depend on a grasp of the clinical use-case (e.g., how the disease is diagnosed in the field), lists malaria documents especially relevant to ML work, and discusses other domain knowledge resources. Section 3 first describes serious problems with commonly-used ML metrics, then describes ML metrics tailored specifically to malaria and NTDs that can be applied during development of ML algorithms to optimize and evaluate their clinical effectiveness. We focus throughout on malaria, but sometimes mention NTDs (e.g., schistosomiasis) because the same principles and methods apply.

2 The clinical use case

To be clinically useful an ML solution must fit into a larger, ML-independent context. It must interlock with other pieces that are shaped by clinician needs, site requirements, protocols currently in use, patient needs, business environment, etc (Wiens et al., 2019). This strong constraint to mesh with non-ML considerations is often overlooked by ML practitioners, leading to algorithms that are elegant (from an ML perspective) but ill-suited for use (from a clinical perspective) (Koller and Bengio, 2018).

In particular, a clinically useful ML algorithm must fit into an existing care structure and meet or exceed existing clinical performance targets. So understanding these clinical constraints is a basic prerequisite for algorithm development. (We set aside the complex case of a disruptive technology potentially altering existing care protocols. Such cases of course require careful analysis.)

This section discusses some crucial points to consider, and lists resources for learning about malaria use cases.

2.1 Important domain specifics

Several domain-specific details are fundamental to effective algorithm development:

2.1.1 Basic facts about the clinical needs

For example, what are the proper uses of thick vs. thin blood films for malaria?

2.1.2 Performance metrics relevant in the clinic

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 (like those we give in Section 3), define objective functions, do internal assessment, and report algorithm results meaningfully.

2.1.3 Performance specifications

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,

1 Defined here as detecting, quantitating, and identifying the species of *Plasmodium* parasites in peripheral blood (CDC, 2023).

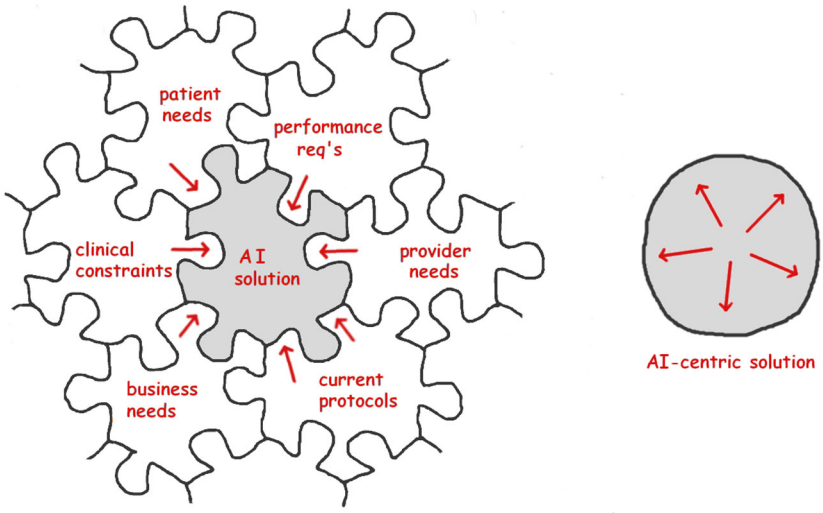

FIGURE 1
Left: Effective ML (AI) solutions must interlock with domain requirements and will be shaped by non-ML pressures arising from the use case. Right: Solutions developed with an ML-centric perspective, neglecting the use case, will be mainly shaped by purely ML concerns and will thus fail to match clinical needs ("interlocking" metaphor due to Dr. Scott McClelland; jigsaw outline from https://draradech.github.io/jigsaw/jigsaw-hex.html).

even if a particular model iteration does not attain them (since the work can then be built upon or extended).

2.1.4 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 white blood cells (WBCs) are plentiful in thick films and stain similarly to malaria parasite nuclei, so they can serve as a ready-made color reference for the rare (or absent) parasites. Shortcuts matter because generic methods applied as-is are unlikely to hit clinical performance requirements, which is a much harder task than simply outdoing another generic method in a ML-style comparison.

2.1.5 Structuring annotations and training sets

Annotations and training data are central to ML success, and must be tailored to the task. For example, malaria ring forms (the youngest parasite stage) typically have both a round nucleus and a crescent-shaped cytoplasm (examples in Figure 2). However, after drug treatment the rings often lack visible cytoplasm, appearing in thick films as dark round dots which are very similar to a common distractor type. As a result, they have outsized impact on decision boundaries and require special care as to annotation and inclusion in training sets.

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

2.2 Documentation

Effective ML solutions need to design in accommodations to non-ML (e.g., clinical) constraints. Therefore, literature review to inform ML work should extend well beyond ML methods and focus on the clinical use-case itself, without an ML-centric filter. Documentation of use cases and standards of care are published by various agencies, including the World Health Organization

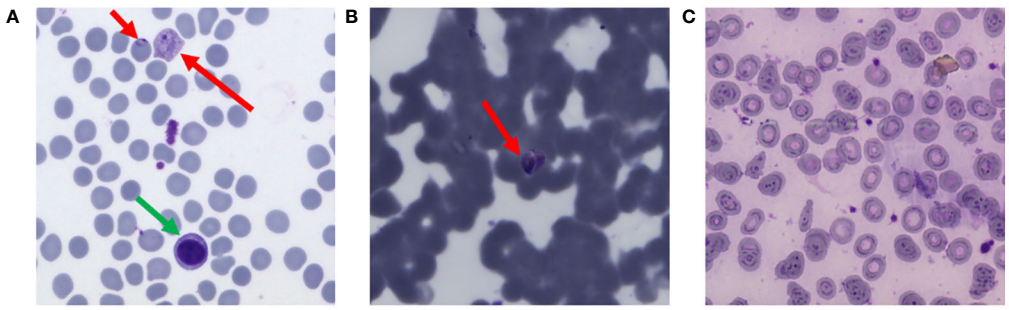

FIGURE 2
Examples of interpatient variability, thin blood films. Red arrows point to parasites, green arrow is a white blood cell. **(A)** Typical "ideal" blood film. **(B)** Poor quality. **(C)** Malaria-negative, with numerous stain and debris artifacts. By permission from Delahunt et al. (2019).

(WHO), ministries of health, and non-government organizations (e.g., the Bill and Melinda Gates Foundation, the Global Fund, and the Worldwide Antimalarial Resistance Network).

Below we list some references that are especially relevant to ML researchers designing algorithms for automated malaria diagnosis using Giemsa-stained blood films.

2.2.1 Appropriate evidence for ML

- The WHO has issued guidelines on how to generate meaningful evidence for ML-based medical tools (WHO, 2021a, especially section 1). This document is important for ML as applied to any medical use case. Crucially, evidence of algorithm performance during development must be firmly grounded in the clinical use case. This requirement underpins the metrics described below in 3.2–3.11.

2.2.2 Protocols for malaria microscopy

Various groups have published diagnosis protocols which detail the clinical task.

- WHO's guidelines are an essential resource (WHO, 2010) and (WHO, 2016a) (see especially SOPs 8 and 9 for diagnosis and quantitation).
- Ministries of Health also have useful protocols, e.g., Peru (Ministerio de Salud, 2003) and USA (CDC, 2023a).
- WWARN and the WHO have developed protocols tailored to research contexts (e.g., drug resistance sentinel sites) (WHO, 2016b).

2.2.3 Evaluation tests

- The WHO has developed a system to evaluate malaria microscopists. This uses a set of 56 blood slides with carefully specified parasitemias and species (WHO, 2016c, section 6). The "WHO 56" evaluation reflects the tasks and accuracies required in the clinic and is thus a valuable and challenging test for ML algorithms. Its difficulty gives an appreciation of the skills of human field microscopists. The defined competency levels offer clear and clinically meaningful performance targets for ML algorithms. Note that the "WHO 56" differs slightly from the previous version (the "WHO 55") found in (WHO, 2009).
- A similar but distinct evaluation set of blood slides, tailored to research rather than clinical contexts, is detailed in (WHO, 2016b).
- Peru's quality control protocols implicitly describe performance requirements (Ministerio de Salud, 2003, sections 7.2 and 9.2).

2.2.4 Neglected tropical diseases

- The WHO has defined target product profiles, including sensitivity and specificity requirements, that are relevant to automated ML systems targeting schistosomiasis (WHO, 2002; WHO, 2021b).

2.2.5 Other performance specifications

- The above documents also provide detail concerning other general product requirements relevant to any ML solution that aims for translation to clinics. These issues include time-to-result, throughput, electricity/battery constraints, price, and (implicitly) computational constraints.

2.2.6 ML publications

- Some ML papers (e.g., Horning et al., 2021; Oyibo et al., 2023) cite non-ML documents relevant to use case, but this is not (yet) common practice. So ML-based literature search is insufficient.

2.3 Domain experts

Domain experts are a vital source of guidance and collaboration. They include field experts, i.e. 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 (these groups overlap). The value of their experience and insight to effective algorithm development cannot be overstated.

As an example, our group's entire ML program for malaria diagnosis has depended absolutely upon expert input from a technical advisory panel, as well as on continued contacts and advice from field clinics. To the degree that our work has succeeded, this expert input has been the key ingredient (along with the closely entwined matter of data collection and curation). We would argue that ML development can only progress toward clinically useful algorithms when domain expertise is somehow integrated into the team (recent examples include Yang et al., 2020; Manescu et al., 2020a, b; Kassim et al., 2021; Poostchi et al., 2018a; Yu et al., 2023; and, for schistosomiasis, Armstrong et al., 2022; Oyibo et al., 2022, 2023).

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; and (iv) non-ML conferences, their attendees, and proceedings, e.g., the American Society of Tropical Medicine and Hygiene.

3 Salient metrics for ML work

Salient metrics are essential to ML work, both to guide development and to report results meaningfully. Unfortunately,

the metrics routinely applied to ML work on malaria (e.g., object-level precision, recall, AUC, and F1 score) have disqualifying drawbacks in the malaria context.

A 2018 review of automated malaria detection papers (Poostchi et al., 2018b) described serious problems (which still persist): 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. We note that in addition, incorrect assumptions are built into algorithms: for example, diagnosis on thin blood films is common in ML papers, despite being contrary to clinical practice due to practical obstacles (Long, 2015; WHO, 2016a; though see recent work on thin film spreaders in Noul, 2023; Nowak et al., 2023).

In this section, we first (3.1) discuss some problems with commonly-used ML metrics and argue that these should not be used to report ML results for malaria.

We then describe in detail (3.2–3.11) some alternative metrics which have high clinical relevance for the malaria use case. These metrics are effective tools both to guide ML development and to report meaningful ML performance results not only for malaria, but also other diseases involving parasite loads such as malaria, NTDs, or 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).

A full list of mathematical notation is given in Table 1.

TABLE 1 List of notation, using malaria as the reference context.

General terms			
LoD	Limit of Detection	NTD	Neglected Tropical Disease
RBC	Red Blood Cell	WBC	White Blood Cell
P	a parasitemia, in p/μL	#	"number of"
μL	microliter	p/μL	parasites per microliter
V	estimated volume of blood examined	cV	clinically-relevant volume (1 μL of blood)
fp	# false positive objects in V	FP	# false positive objects per cV for one patient
tp	# true positive objects in V	TP	# true positive objects per cV for one patient
n	# suspected parasites in V = tp + fp	N	# suspected parasites in cV = TP + FP
fn	# false negative objects in V	tn	# true negative objects in V
Terms used in metric definitions			
FPR	False Positive Rate: FP of one patient	$\boldsymbol{F}$	vector of patients' FPRs
$\sigma(\boldsymbol{F})$	standard deviation of $\boldsymbol{F}$	$\mu(\boldsymbol{F})$	mean of $\boldsymbol{F}$
σ_L ($\boldsymbol{F}$)	lefthanded standard deviation of $\boldsymbol{F}$	σ_R ($\boldsymbol{F}$)	righthanded standard deviation of $\boldsymbol{F}$
S	object-level sensitivity of one patient	$\boldsymbol{S}$	vector of patients' S's
$\sigma(\boldsymbol{S})$	standard deviation of S	$\mu(\boldsymbol{S})$	mean of S
$\hat{F}$	expected FPR, e.g., $\mu(\boldsymbol{F})$	$\hat{S}$	expected sensitivity S, e.g., $\mu(\boldsymbol{S})$
C	threshold on object classifier scores	$\boldsymbol{T}$	threshold on # suspected parasites per cV
K	desired patient-level specificity	α, β	scalars
$\hat{P}$	model's estimate of a true P	$\boldsymbol{L}$	a model's LoD, in p/cV (i.e. p/μL)

3.1 Problems with ROCs, AUCs, and precision

ML practitioners choose metrics to evaluate model performance by (i) what is customary, familiar, and convenient; (ii) what has been done by previous authors; (iii) what can generate the "state of the art" (SOTA) comparisons required for publication in the ML community; and (iv) what is acceptable to ML reviewers. This creates a closed loop which perpetuates the use of certain metrics without regard to their effectiveness. When entrenched metrics do not assess algorithm performance in a clinically relevant way, it blocks progress toward deployable solutions.

Several commonly-used ML metrics, including object-level ROC curves, AUC, object precision, and F1 score, appear frequently in the ML malaria literature. However, in the malaria context these are flawed measures of performance, for reasons given below. They also do not meet standards of evidence per (WHO, 2021a). They should therefore be avoided when reporting results (though they can be useful intermediate measures for internal algorithm work).

3.1.1 Object-level ROC curves and AUC

Object-level ROC curves, and the associated Area Under Curve (AUC), are routinely reported by ML research papers involving parasite detection. ROC curves plot sensitivity (fraction of parasites detected) vs one minus specificity (fraction of distractors misclassified as parasites). However, they have three key weaknesses in this context (except perhaps as intermediate measures for internal algorithm work).

First, they do not address the clinical need for patient-centric care. In particular, they ignore the crucial matter of high inter-patient variability of object-level accuracy (this variability is discussed in 3.4 and 3.3).

Second, real samples often have a large imbalance between distractors and positive objects, especially at parasitemias near clinical limit of detection (LoD). A common situation is a model that seeks to diagnoses malaria on thin films by labeling individual red blood cells (RBCs) as infected or not. Since 1 μL of blood contains roughly 5 million RBCs, a parasitemia of 100 p/μL gives 50,000 negative objects for each positive object. So a 0.999 AUC can coexist with an average of 50 False Positive objects *per parasite* (a very poor SNR). Since one detected parasite and one False Positive object have equal impact on diagnosis (if using the standard method described in 3.6 of exceeding a threshold count of suspected

parasites in the sample), False Positive noise will swamp the diagnostic signal of detected parasites.

In such cases with large class imbalance (say D:1), the leftmost $\frac{1}{D}^{\text{th}}$ 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 detected parasite (True Positive) counts and False Positive counts equally, as shown in Figure 3.

Third, the object-level ROC curve depends heavily on how distractors are defined because this determines the distractor pool. For example, when using thick films to diagnose malaria, "distractor" can mean (i) only the most difficult objects that closely resemble parasites; or (ii) any dark blob; or even (iii) every pixel in an image. Figure 4 shows an example in which including only "difficult" distractors (top) results in a low AUC, while including additional, mostly "easy" distractors (bottom) gives a higher AUC with no change in actual performance as measured by the number of False Positives per detected parasite.

More informative than the object-level ROC is the Free ROC (FROC), which plots object-level sensitivity vs. the number of False Positives per unit volume of blood (see 3.3). FROCs for object level are useful for development work: they clarify where gains can be made by favorably trading off object-level sensitivity for lower False Positive rates. When datasets lack sufficient numbers of patients, FROCs on pooled objects can provide some insight into algorithm performance, with the caveat that they ignore patient-level variability.

3.1.2 Patient-level ROCs

Patient-level ROCs can give a useful sense of algorithm behavior near the clinical performance requirements and are well worth reporting when sufficient data exists to plot them. However, there are two caveats. First, the only salient portion of a patient-level ROC is the region near clinically relevant operating points (e.g., specificity 90%). Second, because sensitivity is parasitemia-dependent (3.4), the ROC is dependent also. Thus, a given algorithm may have much higher AUROC on a population with primarily high parasitemias than on one with lower parasitemias.

3.1.3 Precision

Object-level Precision is the ratio of detected parasites over all detected objects, and often appears as an ML metric. This metric, as used, tends to badly underestimate the effects of parasite-to-distractor imbalances at the low LoDs required for clinical use, as follows.

In ML papers, precision is often calculated on datasets with the clinically unrealistic situation of roughly balanced parasite and distractor counts, either because the numbers of objects have been artificially balanced or because the positive samples had high parasitemias (i.e. many parasites per volume V). Since False Positive counts roughly scale with volume V, high parasitemia samples yield much more balanced True Positive : False Positive $\frac{tp}{tp+fp}$ ratios, which tend to give precisions which do not generalize to low parasitemia samples.

For example, a precision of 0.99 calculated on samples with $P \approx$ 10,000 p/μL corresponds to 100 False Positives per μL (assuming perfect sensitivity). At the required LoD of 100 p/μL, these same 100 False Positives correspond to 100 parasites, giving precision = 0.5, a much less attractive result.

The related metric *F1*, the harmonic mean of precision and object-level sensitivity (also problematic, as noted in 3.4), is a similarly misleading metric for reporting algorithm results, and in addition has no clinical utility.

The rest of this section (3.2–3.11) discusses metrics that better reflect malaria's clinical use case.

3.2 Patient level metrics

The importance of assessing algorithm performance *at the patient level* cannot be over-emphasized. The basic unit of clinical

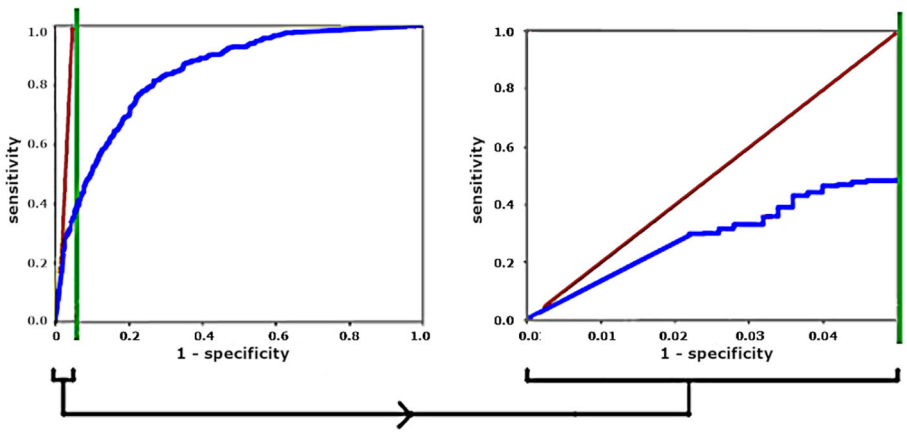

FIGURE 3
For a 20:1 distractor-to-parasite ratio, stretching the left vertical sliver gives a more meaningful ROC curve. Diagonal red lines show operating points that give equal numbers of True Positives and False Positives.

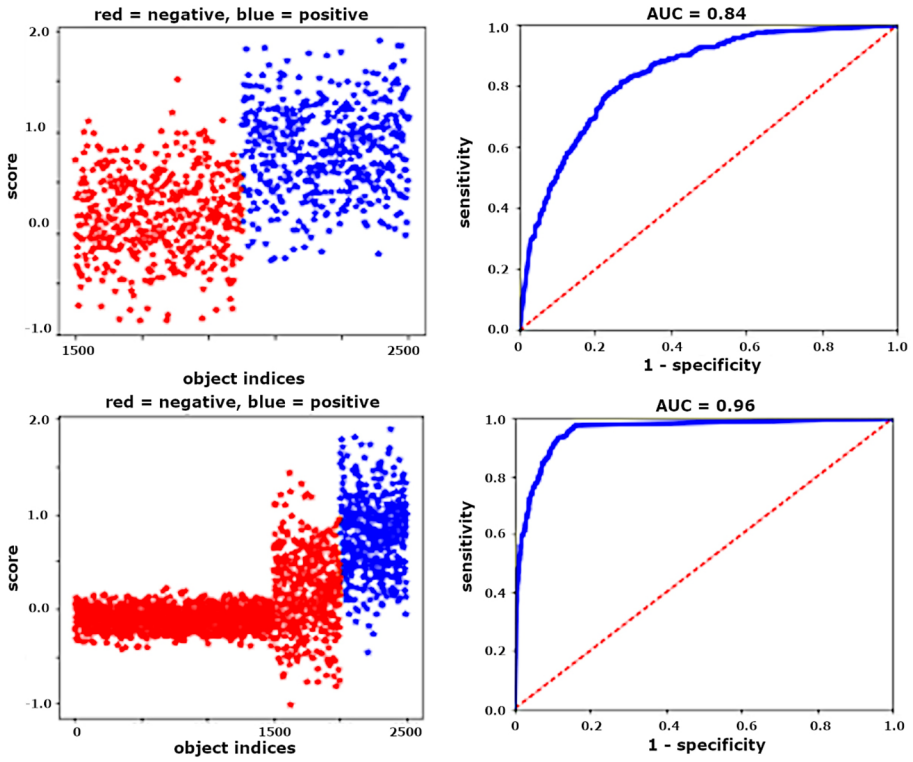

FIGURE 4
For unchanged algorithm and FPR, ROCs are artificially improved by increasing the number of easy distractor objects. Left: object scores (positives are blue, negatives are red). Right: associated ROC curves.

care is the patient[2], 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 malaria-positive patients, with

Patient 1: 50,000 parasites/μL (p/μL).

Patients 2, 3, 4: each 300 p/μL.

Suppose the algorithm detects almost all parasites in {1}, and misses all parasites in {2,3,4} (a realistic scenario due to interslide variability). Then the object-level sensitivity is 98%, while patient-level sensitivity is 25%.

We have found that two metrics, each defined on a per-patient basis, are particularly useful: *false positive rate* (FPR) and *sensitivity*. Each is calculated separately for each patient, using algorithm accuracy on objects within that patient's sample. These are covered in 3.3 and 3.4, and underpin other metrics related to specificity (3.5), LoD (3.6), and quantitation (3.9).

Interpatient variability (as in Figure 2) poses great difficulty for ML, so it must be factored into algorithm evaluation. It is captured by the standard deviations of FPR and sensitivity (cf. 3.3, 3.4), 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) which yield different color ranges. Even clinics with nominally identical protocols can differ substantially (see e.g., Das et al., 2022 and a detailed example in Torres et al., 2018). Besides variations in presentation, different clinics may produce populations of samples with differently distributed FPRs and sensitivities. Implications of this for tuning algorithms are covered in 3.5.

3.3 False positive 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 NTDs), a specified number of cells in a histological sample, etc; but *not* "per image tile", which generally has no clinical relevance (though image tiles can often be translated into the microscopy "Fields of View" used in protocols). Malaria ML papers with some FPR analysis include Linder et al. (2014), Mehanian et al. (2017), Delahunt et al. (2019), and Manescu et al. (2020a). Crucially, FPR is calculated separately for each patient. 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 3.1).

While FPR can be calculated for any sample, FPRs on positive samples may be erroneously boosted by mis- or unannotated

2 We set aside population-level diagnostics such as for Vitamin A deficiency WHO (2011).

parasites. Thus, the population's FPR distribution is best characterized using negative samples only.

Interpatient variability makes the standard deviation of FPR, $\sigma(\boldsymbol{F})$, a crucial performance measure. The mean FPR $\mu(\boldsymbol{F})$ is less relevant because it can be subtracted out, as shown in 3.6 and 3.9. However, since it tends to scale roughly with $\sigma(\boldsymbol{F})$, it can give a hint as to the relative magnitude of $\sigma(\boldsymbol{F})$ (see e.g., Mehanian et al., 2017; Delahunt et al., 2019).

In datasets with insufficient numbers of patients, an FPR calculated over pooled objects has some value as a lower bound on $\boldsymbol{F}$. 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 that work is still needed. Multiple splits of a set of pooled objects does not simulate $\sigma(\boldsymbol{F})$, because each split will include the full patient diversity.

Aside: Samples with high FPRs are sometimes criticized as being due to "poor sample preparation". However, except for extreme cases this is in the eye of the beholder: human clinicians readily and successfully diagnose "dirty" samples on which ML algorithms fail. Thus, the need to improve sample prep is to large degree a need to accommodate ML methods' struggles with handling highly variable sample presentations. See Das et al. (2022), and a detailed example in Torres et al. (2018).

3.4 Sensitivity

Sensitivity (aka recall) is the fraction of positive items in a set that are correctly labeled: Sensitivity = $\frac{tp}{tp+fn}$, where tp = true positives, i.e. positive items labeled correctly, and fn = false negatives, i.e. positive items labeled as negative or missed. The "items" can be parasites (object-level) or malaria-positive patients (patient-level).

3.4.1 Pooled object sensitivity

Sensitivity over a pooled set of parasites from multiple patients has some value as an *intermediate* assessment metric during ML development (e.g., as a loss function for gradient descent training), if it is analyzed carefully to avoid problems such as imbalanced parasitemias distorting the object pool (cf. the example given in 3.2).

3.4.2 Per-patient object sensitivity

A clinically realistic and useful version of object-level sensitivity measures each patient separately:

Per patient object-level sensitivity S is the fraction of parasites in the examined volume V of a positive sample that are correctly labeled (e.g., by means of an object score threshold C): $S = \frac{tp}{tp+fn}$ where tp = parasites labeled correctly, and fn = parasites labeled as distractors (or missed). There is no constraint on the size of V or parasitemia, but sensitivities for patients with few parasites are less reliable (cf. the law of large numbers). Each patient's object-level sensitivity is calculated separately. We denote the vector of sensitivities for the (malaria-positive) population as $\boldsymbol{S}$. $\boldsymbol{S}$ underpins metrics related to LoD (3.6) and quantitation (3.9).

3.4.3 Patient-level sensitivity

Patient-level sensitivity is sensitivity in the usual clinical sense of the fraction of positive patients correctly diagnosed (not S). It is of course a vital metric 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 (where 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 and move in opposite directions, as seen in ROC curves.

Thus, reporting patient-level sensitivity is uninformative and even misleading unless one also reports (i) the parasitemia distribution, and (ii) the associated specificity on negative samples. The WHO competency levels are an important example: These levels crucially assume the parasitemia distribution of the WHO 56 diagnosis slide set, viz 20 negative slides and 20 positive slides with parasitemias between 80 and 200 p/μL (WHO, 2016c). WHO competency level ratings do not apply to results on distributions with higher parasitemia samples.

A principled way to maximize patient-level sensitivity is given in 3.7.

3.4.4 Effect of species on sensitivity

Algorithm sensitivity results should be broken down by species as well as by parasitemia, because malaria species has strong impact on patient-level sensitivity. This is due to the unique synchronization and sequestration behaviors of *P. falciparum* (Garnham, 1966):

(i) In *falciparum* the large, distinctive late stage forms sequester out of the peripheral blood, leaving only the smaller ring forms that are harder to detect and disambiguate from distractor objects (especially in thick films). As a result, in our experience non-*falciparum* infections (i.e. *vivax, ovale, malariae, knowlesi*) are much easier to detect in blood films (given equal parasitemias), which allows an algorithm to have lower LoD and higher patient-level sensitivity (Torres et al., 2018; Delahunt et al., 2019; Horning et al., 2021; Das et al., 2022; Rees-Channer et al., 2023).

(ii) *falciparum* parasites tend to synchronize in peripheral blood, with the presenting parasites forming a narrow age distribution. This strongly impacts diagnostic methods that target the biomarker hemozoin: non-*falciparum* samples can be very sensitively detected due to the reliable presence of late-stage, high hemozoin parasites (Arndt et al., 2021), but even high parasitemia *falciparum* samples can lack detectable hemozoin due to synchronized populations of early stage ring forms (Jamjoom, 1988; Rebelo et al., 2011; Delahunt et al., 2014b), resulting in drastically different sensitivities by species. (Hemozoin appears to be a sensitive biomarker for *falciparum* in cultured blood because synchronization is absent.)

This is a high-stakes issue because *falciparum* is much more often fatal than non-*falciparum* species.

3.5 Specificity

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

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

3.5.1 Object-level specificity

Object-level specificity, even if calculated for each patient separately, has little usefulness and can be highly deceptive (see 3.1).

3.5.2 Patient-level specificity

Patient-level specificity, i.e. in the usual clinical sense, 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 (WHO 2016c), 97.5% for schistosomiasis (WHO, 2021b)).

Specificity is closely tied to FPR (3.3) 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 $\boldsymbol{F}$ (from 3.3) 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

$$T = \mu(\boldsymbol{F}) + \alpha\,\sigma(\boldsymbol{F}) \tag{1}$$

where α is found via the (one-sided) error function and K. Alternate formulations for the case of nongaussian $\boldsymbol{F}$ are given in 3.8.

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).

Note that different clinics can have widely different FPR distributions $\boldsymbol{F}$. Because $\sigma(\boldsymbol{F})$ determines both specificity (Equation 1) and LoD (3.6), 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), with clinic-dependent tradeoffs between specificity and higher LoD.

3.6 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. Based on the WHO evaluation criteria (WHO, 2016c), the required LoD for malaria microscopy is roughly 100 p/μL, i.e. 1 parasite per 50,000 red blood cells (RBCs) or 80 white blood cells (WBCs). However, expert microscopists routinely achieve LoDs ≈50 p/μL (e.g., Vilela, pers. comm.; Bell, pers. comm.), and the lower LoD is of course clinically desirable. For helminths, LoD is implicitly 1 egg (per 10 mL urine or 1 gram stool) (WHO, 2002, 2021b). Standard *Loa loa* diagnosis by blood microscopy has an effective LoD > 200 mf/mL when $V = 10\mu L$ (Mischlinger et al., 2021).

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, this is impractical because it's hard to acquire enough field-prepared positive blood films near the LoD (a work-around is to ablate parasites in the image set of a sample with parasitemia above the LoD, to lower its visible parasitemia).

We can calculate a useful estimate of LoD from $\boldsymbol{F}$ and $\boldsymbol{S}$ as follows:

Denote the putative LoD as L parasites per cV, and suppose that a patient is diagnosed as "positive" when $N \geq T$, where N is the number of positively-labeled objects per cV. Note that $N = TP + FP$ in positive patients, and $N = FP$ in negative patients, where 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 $FP = fp\frac{cV}{V}$).

- Make T high enough to ensure to enforce 95% specificity on negative samples as described in (Mehanian et al., 2017) by setting α to 1.65 std devs in Equation 1:

$$T = \mu(\boldsymbol{F}) + 1.65\sigma(\boldsymbol{F}) \tag{2}$$

- Then for positive samples the worst case is a very "clean" sample with low FPR, such as the 5$^{\text{th}}$ percentile of samples with $FP = \mu(\boldsymbol{F}) - 1.65\sigma(\boldsymbol{F})$. In this case 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 = $\mu(\boldsymbol{S})$. Then a sample at LoD has $TP = L\mu(\boldsymbol{S})$.
- To diagnose this positive sample correctly (but just barely, i.e. $N = T$), we need

$$N = TP + FP = L\,\mu(\boldsymbol{S}) + \mu(\boldsymbol{F}) - 1.65\sigma(\boldsymbol{F})$$

$$= T = \mu(\boldsymbol{F}) + 1.65\sigma(\boldsymbol{F})$$

$$\Rightarrow L\,\mu(\boldsymbol{S}) = 3.3\sigma(\boldsymbol{F})$$

So the estimated LoD (L per cV) has

$$L = \frac{3.3\sigma(\boldsymbol{F})}{\mu(\boldsymbol{S})} \tag{3}$$

- Optionally, +1 can be added to the numerator (i.e. require $N = T+1$) to prevent unpredictable behavior should both $\sigma(\boldsymbol{F})$ and $\mu(\boldsymbol{S})$ approach 0:

$$L = \frac{3.3\sigma(\boldsymbol{F}) + 1}{\mu(\boldsymbol{S})} \tag{4}$$

In our algorithm development, we have found this estimate to be a good (slightly optimistic) proxy for actual LoD when assessing algorithms during development. In particular, it consistently tracked diagnostic accuracy on holdout sets of low parasitemia samples, i.e. lower estimated LoDs mapped to higher accuracy

(sensitivity and specificity at the patient level) in holdout sets and field trials. It has the practical advantage that low parasitemia samples are unnecessary, because the vector $\boldsymbol{S}$ 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 $\sigma(\mathbf{S})$ by having a denominator = $\mu(\boldsymbol{S})-\beta\,\sigma(\boldsymbol{S})$ for some β.

3.7 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, optionally set these aside also.
- For each C:
  - Calculate $\boldsymbol{F}$ over the validation negatives, and $\mu(\boldsymbol{S})$ over the validation positives if available, or (less ideal but workable) over the training set positives.
  - Determine $T = T(C, K, F)$ which hits the target specificity K on the validation negatives, as in 3.5.
  - Estimate LoD as in 3.6.
- 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.

3.8 Modified LoD and operating point formulas

The methods for setting T (Equation 2) and for estimating LoD (Equations 3, 4) both assume that the FPR vector $\boldsymbol{F}$ is gaussian. In our experience this is often not the case. Rather, the FPR distribution may be asymmetrical, with mostly low-FPR samples and a few high-FPR samples. This can be handled by modifying the methods in 3.5 and 3.6 as follows:

- For $\mu(\boldsymbol{F})$, use the median of $\boldsymbol{F}$ instead of the mean of $\boldsymbol{F}$. Similarly, if the vector $\boldsymbol{S}$ is non-gaussian, the median can be used instead of the mean for $\mu(\boldsymbol{S})$.
- For $\sigma(\boldsymbol{F})$, use one-sided std devs, which can be calculated by keeping only the points to the right (or left) of the median and reflecting them across the median as centerpoint to create a symmetric distribution. This gives, for the FPR distribution above, a large right std dev $\sigma_R(\boldsymbol{F})$ and a small left std dev $\sigma_L(\boldsymbol{F})$.
- Then the new versions of Equations 1, 3 are

$$T = median(\boldsymbol{F}) + a\sigma_R(\boldsymbol{F})$$

$$L = \frac{1.65(\sigma_L(\boldsymbol{F}) + \sigma_R(\boldsymbol{F}))}{\mu(\boldsymbol{S})}$$

Two other methods of calculating T from $\boldsymbol{F}$ may be useful:

1. Set T based on the K^{th} percentile of $\boldsymbol{F}$.
2. Manually choose T based on a scatterplot of the FP counts in the validation negative samples.

For both these methods, the detected objects are assumed to be already classified. If a threshold C on object scores was used, then first T must be calculated for each C, before choosing the best $\{C,T\}$ pair as in 3.7.

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.

3.9 Quantitation

Quantitation sometimes has clinical importance. For example, accurate quantitation is needed to monitor for drug-resistant malaria strains by calculating clearance curves (White, 2011; Ashley et al., 2014; WHO, 2016d). For helminths, quantitation targets are typically rough only (e.g., low, medium, high) (WHO, 2002). For *Loa loa*, a remarkable drug reaction necessitates accurate quantitation at certain high parasitemias only (≈ 20k to 30k worms/mL) (Gardon et al., 1997; D'Ambrosio et al., 2015).

3.9.1 Measuring quantitation accuracy

Quantitation accuracy should be reported at the patient level due to high interpatient variability. For plotting quantitation error per patient, Bland-Altman plots are preferable because relative quantitation error is generally most important (WHO, 2016b).

Reporting the R^2 value of a linear fit of estimated vs. true (i.e. $\hat{P}$ vs. P) is unsuitable when parasitemias range over orders of magnitude (common in malaria and NTDs), because effects of the L_2 norm almost guarantee that high parasitemia samples will lay on the fitted line while high relative errors on low parasitemia samples will be downplayed, giving an illusion of strong fit. Fitting the $\log(P)$ rather than P values helps to reduce this illusion.

3.9.2 Estimating parasitemia

As described in Delahunt et al. (2019), we can estimate the parasitemia $\hat{P}$ for a given patient by

$$\hat{P} = \frac{n(\frac{cV}{V}) - \hat{F}}{\hat{S}}, where \tag{5}$$

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

Three types of error affect Equation 5: irreducible Poisson, estimates of examined volume, and counts of alleged parasites.

3.9.2.1 Irreducible Poisson error

This is discussed below in 3.10.

3.9.2.2 Examined volume error

Error in estimating V impacts quantitation accuracy via the $\frac{cV}{V}$ term of Equation 5. For example, thick film blood volume V is typically estimated by counting WBCs (WHO, 2016a). Any error in the WBC count causes proportional quantitation error. This error type can be compartmentalized, for performance evaluation purposes only, as follows:

- Manually count WBCs on a test set to ensure oracle V estimates and use these counts to calculate V, ensuring zero error of this type.
- Separately report the patient-level error statistics of the WBC counter.

3.9.2.3 Parasite counting errors

Errors in parasite count stem from patient-level variations in sensitivity and FPR, as follows:

- The number of alleged parasites per cV in the sample is $(tp + fp)\frac{cV}{V} = TP + FP$.
- Let P be the true parasite count per cV. Then $\hat{S}P$ is the expected number of correctly labeled true parasites per cV, and the difference between TP and $\hat{S}P$ 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. $\sigma(\boldsymbol{S})$ and $\sigma(\boldsymbol{F})$ quantify these deviations over the population.
- A figure of merit to assess parasite counting error, derived and discussed in Delahunt et al. (2019), is thus

$$\frac{\sigma(\boldsymbol{S})}{\mu(\boldsymbol{S})} + \frac{\sigma(\boldsymbol{F})}{\mu(\boldsymbol{S})}\frac{1}{P}$$

While the FPR term is usually hardest to control, it also shrinks as $1/P$, so for large P the sensitivity term dominates. This effect can be leveraged by using different operating points according to whether initial estimated parasitemia is low or high, to favor FPR or sensitivity. In particular, different operating points are indicated for diagnosis (since the hard cases have low parasitemia, where FPR dominates) and for quantitation (high parasitemias, where sensitivity dominates).

We note that parasitemia estimates based on manual microscopy are also subject to these three error types. This complicates assessment of a model's quantitation accuracy against microscopy ground truth.

3.10 Effect of poisson statistics

Poisson statistics for rare events give variation in the actual number of parasites in a particular sample with volume V, given a fixed true parasitemia P over the whole sample. The variation is most visible at low parasitemias, e.g., at 100 p/μL, where each RBC has a 1/50,000 chance of containing a parasite in thin film, or each WBC has a 1/80 chance of corresponding to a nearby parasite in thick film.

This variability has two main 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 at all. Otherwise, for a statistically predictable subset of positive patients the examined volume will contain 0 parasites, reducing patient-level sensitivity from the start. For malaria, to attain LoD of 100 p/μL requires that at least ≈0.05 μL of blood should be examined, equivalent to 400 WBCs in thick film, or 250,000 RBCs in thin film (see Figure 5). The difficulty of finding this many acceptable RBCs, and the long processing time required, are two reasons why thin films are not standard protocol for manual field diagnosis; however, see progress by (Noul, 2023; Nowak et al., 2023). Indeed, the low sensitivity (relative to PCR) of manual microscopy at parasitemias < 50 p/μL (see, e.g., Torres et al., 2018; Das et al., 2022; Rees-Channer et al., 2023) is due largely to Poisson variability: expert microscopists can certainly recognize even a single parasite, but (following protocols) they do not examine sufficient blood to ensure that such a parasite is present when parasitemias are very low. The systems used in our group's studies examine >0.1 μL of thick film (>800 WBCs, the red curve in Figure 5).

(ii) For quantitation, a sufficiently high volume V (depending on P) must be examined to control irreducible error. For more detail and plots see S.I. of Delahunt et al. (2019). Poisson error affects manual microscopy also, and when possible is mitigated by combining multiple manual reads (WWARN, 2023).

In both cases, automated systems hold a strong advantage because they can scan higher volumes than human technicians, who often by necessity work in a high Poisson error regime (S.I. of

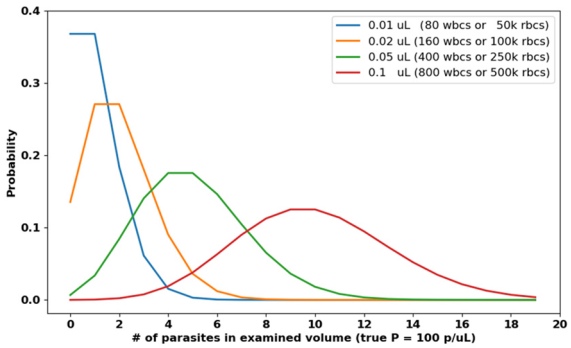

FIGURE 5
Poisson distributions of parasite counts for various examined volumes V , assuming 100 p/μL. Low parasitemia samples may present as negative (i.e. zero parasites) if V is too small.

Delahunt et al., 2019). Manual microscopy protocols average multiple readers' estimates (when available) to reduce quantitation error (WHO, 2016b; WWARN, 2023).

When reporting results on datasets of small size, authors should understand how Poisson variability limits their estimates of algorithm performance.

3.11 Malaria species identification metrics

Identification of malaria species is one of the three tasks assessed by the WHO 56 evaluation system (WHO, 2016c). Correct species ID matters clinically because (i) *falciparum* infections are much more likely to be fatal; and (ii) treatment plans differ by species (CDC, 2023b), since (for example) the hypnozoites of *vivax* and *ovale* species require special care.

Because not all species ID errors are equal from a clinical perspective, reported results should preferably include a confusion matrix as in (Delahunt et al., 2019).

Aside: In our experience, it is relatively straightforward to distinguish *falciparum* vs. non-*falciparum* on thick film alone (Torres et al., 2018; Vongpromek et al., 2019; Das et al., 2022), also (Kassim et al., 2021), and even mixed species infections that include *falciparum* can often be identified on thick film by comparing the ring stage and late stage parasite counts (Horning et al., 2021). However, thin films are still typically needed to distinguish between the various non-*falciparum* species, unless the clinical use case allows geographical priors to be leveraged. A method to distinguish non-*falciparum* species on thick film would yield clinical benefit by eliminating the need for thin films, due to (i) the ease of thick-only workflows (Carter, pers. comm.; Proux, pers. comm.), and (ii) thin film problems with quality (Long, pers. comm.) and difficulty of species ID at low parasitemias (Lilley, pers. comm.).

Staging parasites (as ring, later trophozoite, schizont, or gametocyte) is not part of the WHO evaluation, and is not generally useful clinically, except as used during species identification or when quantitating asexual forms in non-*falciparum* species. In *falciparum* (the main target of quantitation, for drug resistance studies) the difference between ring and gametocyte is glaring.

The methods described here were developed to address the exigencies of field-prepared blood films. They apply equally well to analysis of particular field isolates of any *Plasmodium* species, since the core issues (inter-sample variability, importance of FP objects, etc) apply to field isolates. The caveats connected to analysis of *in vitro* cultures (3.4.4) apply to field isolates as well.

4 Discussion

Malaria and NTDs are amenable though difficult targets for ML methods, and successful development of translatable ML solutions would yield tremendous health care benefits for currently underserved populations by enabling automated malaria diagnosis to augment the throughput capacities of hard-pressed clinicians.

Unfortunately communal ML progress, in which researchers build on each others' work to reach a performance goal, is handicapped for malaria by lack of attention to clinical needs, and by widespread use of ill-suited evaluation metrics. As a result, the synergistic power of the ML community is not being applied with full force to this important task, since many papers present methods that cannot be usefully extended.

Individual ML research teams can radically improve the situation by grounding their ML work in an understanding of the use case, and by 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. We have also listed some essential technical background reading from the WHO and others.

Peer reviewers play a special role in determining the success or failure of the communal ML effort: (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 counterproductive status quo.

With attention to the clinical use case and deliberate choice of metrics, the ML community can better equip itself to successfully address automated malaria and NTD diagnosis, and thus deliver concrete benefit to the populations suffering the dire effects of these illnesses.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

CD, NG, and MH researched methods. CD wrote the manuscript. NG and MH reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Funding provided by Global Health Labs, Inc. (www.ghlabs.org).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Armstrong, M., Coulibaly, J., Essien-Baidoo, S., Bogoch, I., Ephraim, R. K. D., Fletcher, D., et al. (2022). Point-of-care sample preparation and automated quantitative detection of *Schistosoma haematobium* using mobile phone microscopy. *Am. J. Trop. Med. Hyg.* 106 (5). doi: 10.4269/ajtmh.21-1071

Arndt, L., Koleala, T., Orban, A., Ibam, C., Kezsmarki, I., Karl, S., et al. (2021). Magneto-optical diagnosis of symptomatic malaria in Papua New Guinea. *Nat. Commun.* 12 (1). doi: 10.1038/s41467-021-21110-w

Ashley, E., Dhorda, M., Fairhurst, R., Amaratunga, C., Lim, P., White, J., et al. (2014). Spread of artemisinin resistance in *Plasmodium falciparum* malaria. *N. Engl. J. Med* 371 (5), 411–423. doi: 10.1056/NEJMoa1314981

BMGF (2023) *Neglected tropical diseases*. Available online at: https://www.gatesfoundation.org/our-work/programs/global-health/neglected-tropical-diseases.

CDC (2023) *Malaria website*. Available online at: https://www.cdc.gov/malaria/diagnosistreatment/diagnosis.html.

CDC (2023a). *Malaria website* (USA: Centers for Disease Control). Available at: https://www.cdc.gov/malaria/index.html.

CDC (2023b). *Algorithm for diagnosis and treatment of malaria in the United States* (Centers for Disease Control). Available at: https://www.cdc.gov/malaria/resources/pdf/Malaria_Managment_Algorithm_202208.pdf.

D'Ambrosio, M. V., Bakalar, M., Bennure, S., Reber, C., Sakndarajah, A., Fletcher, D. A., et al. (2015). Point-of-care quantification of blood-borne filarial parasites with a mobile phone microscope. *Sci. Trans. Med* 7 (286). doi: 10.1126/scitranslmed.aaa3480

Das, D., Vongpromek, R., Assawariyathipat, T., Srinamon, K., Kennon, K., Dhorda, M., et al. (2022). Field evaluation of the diagnostic performance of EasyScan GO: a digital malaria microscopy device based on machine-learning. *Malaria J.* 21, 122. doi: 10.1186/s12936-022-04146-1

Delahunt, C., Horning, M., Wilson, B., Proctor, J., and Hegg, M. (2014b). Limitations of haemozoin-based diagnosis of Plasmodium falciparum using dark-field microscopy. *Malaria J.*, 393–399. doi: 10.1186/1475-2875-13-147

Delahunt, C. B., Jaiswal, M. S., Horning, M. P., Janko, S., Thompson, C. M., Mehanian, C., Kulhare, S., et al. (2019). Fully-automated patient-level malaria assessment on field-prepared thin blood film microscopy images. *arXiv*. doi: 10.48550/arXiv.1908.01901

Delahunt, C. B., Mehanian, C., Hu, L., McGuire, S. K., Thompson, C., Wilson, B. K., et al. (2014a). Automated microscopy and machine learning for expert-level malaria field diagnosis. *IEEE GHTC Proc*. doi: 10.1109/GHTC.2015.7344002

Gardon, J., Gardon-Wendel, N., Demanga-Ngangue, , Demanga-Ngangue Kamgno, J., Chippaux, J.P., and Boussinesq, M. (1997). Serious reactions after mass treatment of onchocerciasis with ivermectin in an area endemic for loa loa infection. *Lancet* 350 (9070), 18–22. doi: 10.1016/S0140-6736(96)11094-1

Garnham, P. (1966). *Malaria parasites and other haemosporidia* (Oxford, UK: Blackwell Scientific Publications Ltd).

Horning, M.P., Delahunt, C.B., Bachman, C.M., Luchavez, J., Luna, C., Mehanian, C, et al. (2021). Performance of a fully-automated system on a WHO malaria microscopy evaluation slide set. *Malaria J* 20, 110. doi: 10.1186/s12936-021-03631-3

Jamjoom, G. (1988). Patterns of pigment accumulation in *Plasmodium falciparum* trophozoites in peripheral blood samples. *Am. J. Trop. Med. Hyg* 39 (1), 21–25. doi: 10.4269/ajtmh.1988.39.21

Kassim, Y., Yang, F., Yu, H., Maude, R., and Jaeger, S. (2021). Diagnosing malaria patients with *Plasmodium falciparum* and *vivax* using deep learning for thick smears. *Diagnostics* 11 (11). doi: 10.3390/diagnostics11111994

Koller, D., and Bengio, Y. (2018) *A fireside chat with Daphne Koller at ICLR*. Available online at: https://www.youtube.com/watch?v=N4mdV1CIpvI.

Linder, N., Turkki, R., Walliander, M., Martensson, A., Diwan, V., Lundin, J., et al. (2014). A malaria diagnostic tool based on computer vision screening and visualization of *Plasmodium falciparum* candidate areas in digitized blood smears. *PloS One* 9 (8). doi: 10.1371/journal.pone.0104855

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Jager, P. F., et al. (2022). Metrics reloaded: Pitfalls and recommendations for image analysis validation. *arXiv*. https://arxiv.org/abs/2206.01653. doi: 10.48550/arXiv.2206.01653

Manescu, P., Bendkowski, C., Claveau, R., Elmi, M., Brown, B. J., Fernandez-Reyes, D., et al. (2020b). A weakly supervised deep learning approach for detecting malaria and sickle cells in blood films. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 226–235. doi: 10.1007/978-3-030-59722-1_22

Manescu, P., Shaw, M. J., Elmi, M., Neary-Zajiczek, L., Claveau, R., Fernandez-Reyes, D., et al. (2020a). Expert-level automated malaria diagnosis on routine blood films with deep neural networks. *Am. J. Hematol.* 95 (8). doi: 10.1002/ajh.25827

Mehanian, C., Jaiswal, M., Delahunt, C. B., Thompson, C., Horning, M. P., Bell, D., et al. (2017). Computer-automated malaria diagnosis and quantitation using convolutional neural networks. *ICCV* pp. 116–125. doi: 10.1109/ICCVW.2017.22

Ministerio de Salud (2003). *Manual de Procedimientos de Laboratoria Para el Diagnostico de Malaria* (Lima, Peru: Instituto Nacional de Salud).

Mischlinger, J., Manego, R., Mombo-Ngoma, G., Ekoka Mbassi, D., Hackbarth, N., Ekoka Mbassi, F. A., et al. (2021). Diagnostic performance of capillary and venous blood samples in the detection of *Loa loa* and *Mansonella perstans* microfilaraemia using light microscopy. *PloS Negl. Trop. Dis.* 15 (8). doi: 10.1371/journal.pntd.0009623

Noul (2023). *miLab platform* (S. Korea: Noul). Available at: https://noul.kr/en/milab.

Nowak, J., Kothari, A., Li, H., Pannu, J., Algazi, D., and Prakash, M. (2023). Inkwell: Design and validation of a low-cost open electricity-free 3d printed device for automated thin smearing of whole blood. *arXiv*. doi: 10.48550/arXiv.2304.10200

Oyibo, P., Meulah, B., Bengtson, M., van Lieshout, L., Oyibo, W., Agbana, T., et al. (2023). Two-stage automated diagnosis framework for urogenital schistosomiasis in microscopy images from low-resource settings. *J. Med. Imaging* 10 (4). doi: 10.1117/1.JMI.10.4.044005

Oyibo, P., Jujjavarapu, S., Meulah, B., Agbana, T., Braakman, I., Diehl, J. -C., et al. (2022). Schistoscope: An automated microscope with artificial intelligence for detection of *Schistosoma haematobium* eggs in resource-limited settings. *Micromachines* 13 (643). doi: 10.3390/mi13050643

Poostchi, M., Ersoy, I., McMenamin, K., Gordon, E., Palaniappan, N., Jaeger, S., et al. (2018a). Malaria parasite detection and cell counting for human and mouse using thin blood smear microscopy. *J. Med. Imaging* 5 (4). doi: 10.1117/1.JMI.5.4.044506

Poostchi, M., Silamut, K., Maude, R., Jaeger, S., and Thoma, G. (2018b). Image analysis and machine learning for detecting malaria. *Trans. Res.* doi: 10.1016/j.trsl.2017.12.004

Rebelo, M., Shapiro, H., Amaral, T., Melo-Cristino, J., and Hanscheid, T. (2011). Haemozoin detection in infected erythrocytes for *Plasmodium falciparum* malaria diagnosis-prospects and limitations. *Acta Tropica* 123 (1), 58–61. doi: 10.1016/j.actatropica.2012.03.005

Rees-Channer, R. R., Bachman, C. M., Grignard, L., Gatton, M. L., Burkot, S., Chiodini, P. L., et al. (2023). Evaluation of an automated microscope using machine learning for the detection of malaria in travelers returned to the UK. *Front. Malaria* 1. doi: 10.3389/fmala.2023.1148115

Reinke, A., and Tizabi, M. (2024). Understanding metric-related pitfalls in image analysis validation. *Nat. Methods* 21, 182–194. doi: 10.1038/s41592-023-02150-0

Torres, K., Bachman, C.M., Delahunt, C.B., Baldeon, J.A., Alava, F., Bell, D., et al. (2018). Automated microscopy for routine malaria diagnosis: a field comparison on Giemsa-stained blood films in Peru. *Malaria J* 17, 33. doi: 10.1186/s12936-018-2493-0

Vongpromek, R., Proux, S., Ekawati, L., Archasuksan, L., Bachman, C., Bell, D., et al. (2019). Field evaluation of automated digital malaria microscopy: EasyScan GO. *Trans. R Soc. Trop. Med. Hyg.*

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Goldenberg, A., et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* 25 (9). doi: 10.1038/s41591-019-0548-6

White, N. (2011). The parasite clearance curve. *Malaria J.* 10, 278. doi: 10.1186/1475-2875-10-278

WHO (2002). *Prevention and control of schistosomiasis and soil-transmitted helminthiasis* (Geneva, Switzerland: World Health Organization).

WHO (2009). *Malaria Microscopy Quality Assurance Manual V1* (Geneva, Switzerland: World Health Organization).

WHO (2010). *Basic malaria microscopy. Part I. Learner's guide. 2nd ed* (Geneva, Switzerland: World Health Organization).

WHO (2011). *Serum retinol concentrations for determining the prevalence of vitamin A deficiency in populations* (Geneva, Switzerland: World Health Organization).

WHO (2016a). *Microscopy examination of thick and thin blood films for identification of malaria parasites (esp SOPs 8 and 9)* (Geneva, Switzerland: World Health Organization).

WHO (2016b). *Microscopy for the detection, identification and quantification of malaria parasites on stained thick and thin blood films in research settings, ver 1* (Geneva, Switzerland: World Health Organization).

WHO (2016c). *Malaria microscopy quality assurance manual v2* (Geneva, Switzerland: World Health Organization).

WHO (2016d). *Malaria Microscopy Standard Operating Procedure MM-SOP-09: Malaria Parasite Counting* (Geneva, Switzerland: World Health Organization).

WHO (2019). *World malaria report 2019* (Geneva, Switzerland: World Health Organization).

WHO (2021a). *Generating evidence for artificial intelligence-based medical devices: a framework for training, validation and evaluation* (Geneva, Switzerland: World Health Organization).

WHO (2021b). *Diagnostic target product profiles for monitoring, evaluation and surveillance of schistosomiasis control programmes* (Geneva, Switzerland: World Health Organization).

WHO (2023). *Global report on neglected tropical diseases 2023* (Geneva, Switzerland: World Health Organization).

WWARN (2023) *Obare method calculator*. Available online at: https://www.wwarn.org/obare-method-calculator.

Yang, F., Poostchi, M., Yu, H., Zhou, Z., Silamut, K., Yu, J., et al. (2020). Deep learning for smartphone-based malaria parasite detection in thick blood smears. *IEEE J. BioMed. Health Inform* 24, 1428–1438. doi: 10.1109/JBHI.6221020

Yu, H., Mohammed, F. O., Hamid, M. A., Yang, F., Kassim, Y. M., Jaeger, S., et al. (2023). Patient-level performance evaluation of a smartphone-based malaria diagnostic application. *Malaria J.* 22, 33. doi: 10.1186/s12936-023-04446-0