THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Automated detection and staging of malaria parasites from cytological smears using convolutional neural networks

OPEN ACCESS

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

Automated detection and staging of malaria parasites from cytological smears using convolutional neural networks

Mira S. Davidson[1] ⓘ, Clare Andradi-Brown[1,2], Sabrina Yahiya[1], Jill Chmielewski[3],
Aidan J. O'Donnell[4] ⓘ, Pratima Gurung[5,6] ⓘ, Myriam D. Jeninga[7], Parichat Prommana[8],
Dean W. Andrew[9], Michaela Petter[7], Chairat Uthaipibull[8,9], Michelle J. Boyle[10],
George W. Ashdown[1] ⓘ, Jeffrey D. Dvorin[5,6] ⓘ, Sarah E. Reece[4] ⓘ, Danny W. Wilson[3,11] ⓘ,
Kane A. Cunningham[12], D. Michael. Ando[13] ⓘ, Michelle Dimon[13] ⓘ and Jake Baum[1]* ⓘ

[1]Department of Life Sciences, Imperial College London, London, United Kingdom
[2]Department of Infectious Disease, Imperial College London, London, United Kingdom
[3]Research Centre for Infectious Diseases, School of Biological Sciences, University of Adelaide, Adelaide, South Australia, Australia
[4]Institute of Evolutionary Biology, and Institute of Immunology and Infection Research, School of Biological Sciences, The University of Edinburgh, Edinburgh, United Kingdom
[5]Division of Infectious Diseases, Boston Children's Hospital, Boston, Massachusetts, USA
[6]Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA
[7]Mikrobiologisches Institut–Klinische Mikrobiologie, Immunologie und Hygiene, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Erlangen, Germany
[8]Medical Molecular Biotechnology Research Group National Center for Genetic Engineering and Biotechnology (BIOTEC), Pathum Thani, Thailand
[9]Thailand Center of Excellence for Life Sciences, Bangkok, Thailand
[10]QIMR Berghofer Medical Research Institute, Herston, Queensland, Australia
[11]Burnet Institute, Melbourne, Victoria, Australia
[12]Google Cloud AI, Sunnyvale, California, USA
[13]Applied Science Team, Google Research, Mountain View, California, USA
*Corresponding author. Email: jake.baum@imperial.ac.uk

Abstract

Microscopic examination of blood smears remains the gold standard for laboratory inspection and diagnosis of malaria. Smear inspection is, however, time-consuming and dependent on trained microscopists with results varying in accuracy. We sought to develop an automated image analysis method to improve accuracy and standardization of smear inspection that retains capacity for expert confirmation and image archiving. Here, we present a machine learning method that achieves red blood cell (RBC) detection, differentiation between infected/uninfected cells, and parasite life stage categorization from unprocessed, heterogeneous smear images. Based on a pretrained Faster Region-Based Convolutional Neural Networks (R-CNN) model for RBC detection, our model performs accurately, with an average precision of 0.99 at an intersection-over-union threshold of 0.5. Application of a residual neural network-50 model to infected cells also performs accurately, with an area under the receiver operating characteristic curve of 0.98. Finally, combining our method with a regression model successfully recapitulates intraerythrocytic developmental cycle with accurate lifecycle stage categorization. Combined with a mobile-friendly web-based interface, called PlasmoCount, our method permits rapid navigation through and review of results for quality assurance. By standardizing assessment of Giemsa smears, our method markedly improves inspection reproducibility and presents a realistic route to both routine lab and future field-based automated malaria diagnosis.

Impact Statement

Microscopy inspection of Giemsa-stained thin blood smears on glass slides has been used in the diagnosis of malaria and monitoring of malaria cultures in laboratory settings for >100 years. Manual evaluation is, however, time-consuming, error-prone, and subjective with no currently available tool that permits reliable automated counting and archiving of Giemsa-stained images. Here, we present a machine learning method for automated detection and staging of parasite infected red cells from heterogeneous smears. Our method calculates parasitemia and frequency data on the malaria parasite intraerythrocytic developmental cycle directly from raw images, standardizing smear assessment, and providing reproducible and archivable results. Developed into a web tool, PlasmoCount, this method provides improved standardization of smear inspection for malaria research and potentially field diagnosis.

1. Introduction

Malaria is an infectious disease caused by protozoan parasites from the genus *Plasmodium*, of which *Plasmodium falciparum* is the most common and lethal to humans[1]. Despite advances in rapid point-of-care diagnostics, the most widely used method for diagnosing malaria remains the manual counting of parasites within infected red blood cells (RBCs) from microscopic inspection of Giemsa-stained blood films[2]; a method that has remained largely unchanged in nearly 120 years[3]. Besides diagnosing malaria, Giemsa staining is also the cornerstone of laboratory research that involves parasite tissue culture[4], including drug and vaccine efficacy trials. However, the identification and counting of parasites is a time-consuming process that requires trained microscopy technicians[5]. Moreover, manual evaluation can be erroneous and varies between slide readers[6–8].

Recent advances in machine learning (ML) have provided an opportunity to explore automating the detection of parasites from cytological smears[9]. In general, these efforts focus on some or all of a sequence of computational tasks: (a) cell segmentation (partitioning a digital image into multiple segments such as pixels) and the detection of individual RBCs, (b) parasite identification and discrimination between infected and uninfected cells, and (c) subclassification of the different stages of parasite development. Early efforts mostly applied methods based on histogram thresholding and morphological operations, extracting hand-crafted features, and classification using traditional ML methods[10–15]. Recently, there has been a general trend toward using deep learning methods for feature computation[16–20] as well as cell segmentation[21–23].

Despite significant improvements in model performance, however, there is currently no widely used tool for the automated detection and staging of malaria parasites. Making models available to the wider community would enable their use in routine malaria research, where standards of parasite nomenclature and parasitemia count vary between users and groups and provide a route to field testing to advance the automated clinical diagnosis of malaria from smears. A major caveat that has held back automated approaches to date is the absence of standardization in the preparation of thin blood films, which results in extensive staining and lighting variations between laboratories[24]. Moreover, introducing a usable method requires a processing pipeline from the raw image to result, with only a few studies focused on all aspects of the pipeline. For example, most studies have developed methods for datasets of images of single segmented RBCs[20]. This method requires manual cell segmentation by a microscopist and neglects the effect of dust, debris, staining artefacts, or neighboring cells on parasite detection. Only a few studies have also combined parasite detection with the classification of the different stages of the intraerythrocytic development cycle (IDC). Furthermore, treating parasite development as a classification problem disregards information on progression within and between the individual stages; progression through the IDC is a continuous process, and experts disagree on the boundaries between the different life stages[10,21,25,26]. Automation has the potential to save time and add to the number of RBCs sampled for both diagnosis and lab usage; however, its usage in diagnosis would require confidence in the analysis process; if results are accessible for review post-analysis by a microscopist, such a system is more likely to be implemented as a robust decision support tool.

Here, we present an ML method that combines cell segmentation, parasite detection, and staging in *P. falciparum*. We use a combination of residual neural networks (ResNets) with transfer learning to quantify parasitemia and categorize IDC stages in both an accurate and standardized manner. Moreover, we present the first application of a regression model to order life stage categorization that accurately recapitulates the progression of the IDC. Bringing these approaches together, we have built a web tool prototype, called PlasmoCount, which allows for versatile detection and stage categorization of malaria parasites from Giemsa-stained images with an interactive assessment of the results. We test the performance of the tool with independent test sets and observe high performance across all models. We demonstrate that the web tool can be used in conjunction with mobile devices from image capture to assessment, opening up new opportunities for rapid, low-cost automated diagnosis from gold-standard smears. We believe this approach can bring new rigor to *Plasmodium* biology studies based on the inspection of smears and potentially as a future platform for on-the-go support in clinical settings where Giemsa-based cytological smears are used as the primary tool for confirming diagnosis of malaria.

2. Methods

2.1. Ethical approval

The work described requires the use of human RBCs for the propagation of *P. falciparum* in in vitro culture. Baum laboratory research grade RBCs are purchased from the National Health Service Blood and Transplant (NHSBT). These RBCs are given anonymously with no clinical history. Sourcing of RBCs in this manner, including informed consent, is entirely handled by NHSBT. Imperial College London has determined that this does not represent human subjects research, so no independent ethical approval is required. Use of RBCs sourced from the NHSBT is approved by the Imperial College London BioSafety Committee. Human RBCs for use by the Wilson and Boyle Laboratories were provided by the Australian Red Cross Blood Bank with ethics approval for use of the cells obtained from the University of Adelaide Human Ethics Committee and the Human Research and Ethics Committee of the QIMR-Berghofer Institute of Medical Research, respectively. Human RBCs for use by the Uthaipibull Laboratory were obtained from donors after providing informed written consent, following a protocol approved by the Ethics Committee, National Science and Technology Development Agency, Pathum Thani, Thailand. Anonymous human RBCs for *Plasmodium* culture in the Petter Laboratory were purchased from the German Red Cross Service (Blutspendedienst des Bayerischen Roten Kreuzes) while anonymous human RBCs used by the Dvorin Laboratory were obtained from a commercial vendor (Valley Biomedical, Winchester, Virginia, USA). Commercial provision is not deemed as human subjects research and therefore does not require ethical approval. All procedures involving *Plasmodium chabaudi* in the Reece Laboratory were carried out in accordance with the UK Home Office regulations (Animals Scientific Procedures Act 1986; project license number 70/8546) and approved by the University of Edinburgh.

2.2. Sample preparation

An image dataset was collected by each research center according to the protocol as follows without further instructions to allow for experimental variation. Varied *P. falciparum* parasites (strains 3D7, NF54, DD2, and D10 depending on the research group) were cultured with a parasitemia of around 5% according to standard protocols in the individual laboratories. Thin blood films were prepared and air-dried for 1–2 min. Smears were then fixed in methanol for 30 s and stained by flooding the slide for 15 min in a fresh Giemsa solution of 10% Giemsa stain in a phosphate-buffered solution. The smear was then washed in water and air-dried. Slides were imaged under 100% oil immersion with a $100\times$ objective and saved in TIFF format. For phone capture, images were taken as JPEG files with $2\times$ optical zoom on an iPhone 8 by manually aligning the phone camera with the microscope eyepiece. Details on sample and imaging specifications from each dataset can be found in Supplementary Table 1 (Supplementary Figure S1 [example images]).

2.3. Labeling pipeline

We used a model-assisted approach for labeling using the LabelBox platform[27], as this greatly improved labeling speed and performance. Each labeling round contained ~100 raw images of Giemsa smears. To aid the first labeling round, we trained our object detection model on a dataset of *Plasmodium vivax*-infected blood smears from the Broad Bioimage Benchmark Collection[28]. Predictions on our *P. falciparum* dataset were then uploaded as prelabels using the LabelBox Python SDK. For each of the following labeling rounds, the RBC detection model and malaria identification classifier were trained on all the previous labeled datasets to generate new labels. Annotators could correct, add, and delete bounding boxes around each RBC, and choose from three labels: *infected*, *uninfected*, and *unsure.* At this stage, each image was corrected only by one annotator. RBCs labeled as either *infected* or *unsure* were automatically uploaded to a second labeling round as crops alongside the original image. Annotators could choose one label from *ring*, *trophozoite*, and *schizont* to mark infections approximately 0–12, 12–36, and 36–48 hr post-infection for asexual development; *gametocyte* to mark sexual development; *multiple infections* to mark more than one parasite per infected RBC; *uninfected cell*, *not a cell (e.g., debris and merozoites)* to flag mislabeled cases from the object detection labeling round; or *unsure*. RBCs labeled as *ring*, *trophozoite*, or *schizont* were resubmitted for a third labeling round where annotators could choose multiple labels: *early ring, late ring, early trophozoite, late trophozoite, early schizont, late schizont, multiple infections,* and *unsure.* For both labeling rounds, each RBC was labeled by three different annotators using random allocation. All images were labeled five times by five designated annotators from three different research centers for the test set.

2.4. Data analysis

A ResNet-50 architecture[29] pretrained on ImageNet[30] was used for the backbone of the Faster R-CNN object detection model and all classification models. For regression of the life stage development, we adopted a smaller ResNet-34 model[29] pretrained on ImageNet to avoid overfitting. We pooled and flattened the final convolutional layer using average and maximum pooling and extended the model with two fully connected (FC) layers (1024, 512; ReLu activated) preceded with batch normalization and dropout ($p = .5$) and a final FC layer[16]. The output of the 512-neuron FC layer feeds into a final FC layer of 16 neurons to measure the life stage. Training data were split into training and validation sets by 80/20 with random sampling. The validation set was used during the training phase to set the learning rate and determine early stopping. Data augmentation methods were used to increase the variety of our training data. For the object detection model, we applied random horizontal flipping ($p = .5$), vertical flipping ($p = .5$), Gaussian blurring (sigma = (0,15), $p = .5$), contrast changes (pixel value (v) scaling by $127 + \text{alpha}*(v\text{-}127)$ with alpha = (0.75, 1.25)), brightness changes (image multiplication using range = (0.75, 1.25), once per channel for 20% of images), and saturation multiplication (range = (0.5, 1.25)). For the other models, we applied random horizontal flipping ($p = .5$), vertical flipping ($p = .5$), rotation (range = $(-10,10$ degrees), $p = .75$), contrast changes (pixel multiplication by loguniform(0.8, 1.2), $p = .75$), and brightness changes (image addition using logit(uniform(0.8, 1.2)), $p = .75$). For life stage assessment, we increased the brightness changes (image addition using logit(uniform(0.25, 1.75))), and randomized RGB channels (scale below threshold = 0.3, $p = .1$). For the Faster R-CNN model, we used stochastic gradient descent for optimization and a learning rate of 0.01 with stepwise learning rate decay. For all other models, we used the Adam optimization algorithm[31] and adopted a one-cycle learning rate policy[32].

2.5. Web application architecture

PlasmoCount is a beta version, available on www.plasmocount.org (access details are available at http://www.baumlab.com/plasmocount). The web server was built on a Flask framework and provides application programming interface endpoints to connect the front-end interface to the ML models. The front end was developed using the React web application framework and the Plotly interactive graphing library.

The results are stored in cloud storage using Google Storage. When a user uploads their images, a unique job gets created, data are submitted to the backend, and the user is redirected to the job-associated results page. The user also receives an email at the email address provided with a nonexpiring link to this page for future reference. The Flask backend listens to updates in the cloud storage; once the analysis is completed, the results get returned to the front end where they are visualized with interactive components using React and Plotly.

3. Results

3.1. Data

Given the inherent inconsistencies of manual inspection, counting, and staging of malaria parasites by Giemsa-stained cytological smear, we sought to develop an ML-based approach that could automate analysis of standard laboratory-acquired images of *Plasmodium* cultures. Six datasets of microscopy images of thin blood films of *P. falciparum* cultures were generated using a standardized Giemsa protocol described in the Methods section. Each dataset was collected by a different research center with one dataset left out for testing (Figure 1a, Supplementary Table 1, and Supplementary Figure S1). Datasets were labeled in five labeling rounds by 10 expert malaria researchers, familiar with Giemsa-smears, across two different laboratories. Labeling was split into two steps: marking out bounding boxes around uninfected and infected RBCs and subsequent life stages classification.

To facilitate bounding box demarcation, we used a model-assisted labeling workflow via the LabelBox platform[27]. A new dataset was made available with bounding boxes drawn on uninfected and infected RBCs in images by the model after every training round. Annotators were then asked to correct model predictions made on the *P. falciparum* dataset (Figure 1b) and annotate RBCs according to standard IDC asexual stage nomenclature (ring/immature trophozoite, trophozoite/mature trophozoite, and schizont) and gametocytes. Moreover, annotators could label RBCs as "*unsure.*" While comprehensive, this classification task resulted in only 60% unanimous agreement between *all* annotators when labeling
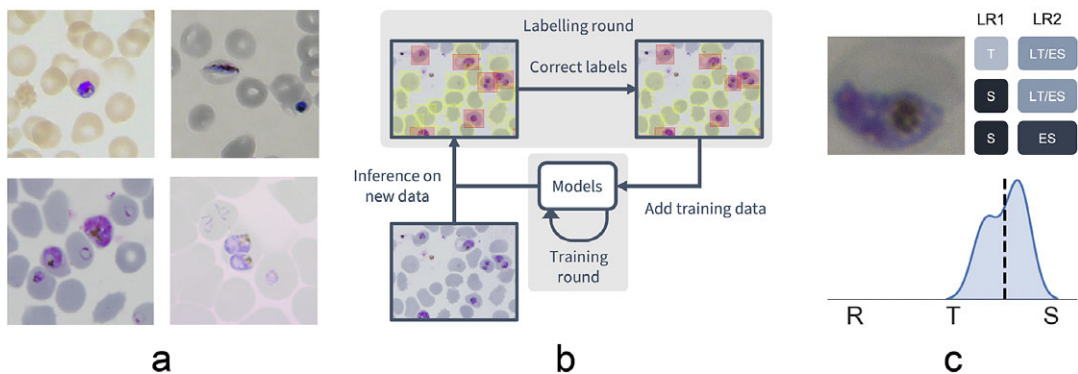
Figure 1. *Data collection workflow. (a) Example images of the* P. falciparum *dataset. Datasets were collected by six different research centers resulting in staining and lightning variations. (b) Model-assisted labeling workflow. At each labeling round, model predictions were imported and loaded as editable annotations on an image. These annotations were corrected by annotators and added to the training data for a new training round. This process was then repeated with the retrained object detection and malaria classification models. (c) Example of disagreement between annotators on parasite stage classification (R, ring; T, trophozoite; S, schizont). Whereas the first labeling round (LR1) only captures the disagreement between two stages, the second round (LR2) reveals that this is due to the parasite existing between these two stages (LT, late trophozoite; ES, early schizont). The final value for the image is calculated by averaging labels across all annotators.*

the training set, with most disagreements occurring at the boundaries between IDC stages, particularly between trophozoite and schizont stages. Annotators never labeled the same image twice, so no data were collected on intra-annotator consistency; however, 9% of the collected labels were "*unsure*" unveiling intra-annotator uncertainty. To address the disagreements caused by the ambiguity of the life cycle stage boundaries, life stage assessment was repeated a second time for IDC stages but this time allowing annotators to select multiple values to account for uncertainty at boundaries. In addition, each IDC stage was subdivided into early and late stages (Figure 1c). For example, a parasite between the ring and trophozoite stages could be annotated by selecting both "*late ring*" and "*early trophozoite.*" To provide a numerical score associated with labels, classifications were converted to a numeric scale (ring = 1, trophozoite = 2, and schizont = 3) and averaged across all annotators (Supplementary Figure S2). In total, this led to a dataset with 38,000 different RBCs, of which 6% were infected.

3.2. Red blood cell detection

Having developed a standardized dataset, we next trained a Faster R-CNN object detection model[33] to detect both uninfected and infected RBCs in microscopy images of thin blood smears. Faster R-CNN is a state-of-the-art object detection network that consists of a region proposal network and a Faster R-CNN network that classifies the proposed regions[34]. This approach has recently shown some promising results in cell segmentation in blood smears of *P. vivax*[23]. We used a Faster R-CNN model pretrained on the COCO dataset[35], fine-tuned on a concatenation of our *P. falciparum* dataset and the aforementioned *P. vivax* dataset resulting in a dataset of 1491/241 images with 108573/14989 RBCs for training/testing with an 80/20 training/validation split. The classification task was not included at this first stage, instead focusing only on cell segmentation, aiming to make the model robust to variations between laboratories. False positives in the form of debris, merozoites (generally smaller than RBCs), or leukocytes (generally larger than RBCs) were excluded by applying a cutoff based on the area distribution of the detected bounding boxes with those four standard deviations above the mean area across the image discarded. Finally, cells that touched the border of the crop were excluded. A nonmaximum suppression was applied using an intersection over union (IOU) threshold of 0.5 to exclude overlapping bounding boxes. The average precision (AP) at an IOU threshold of 0.5 was 0.98 on the *P. vivax* test set and 0.99 on our *P. falciparum* test set (Figure 2). Thus, we can extract individual RBCs from the background effectively as an input for further classification.
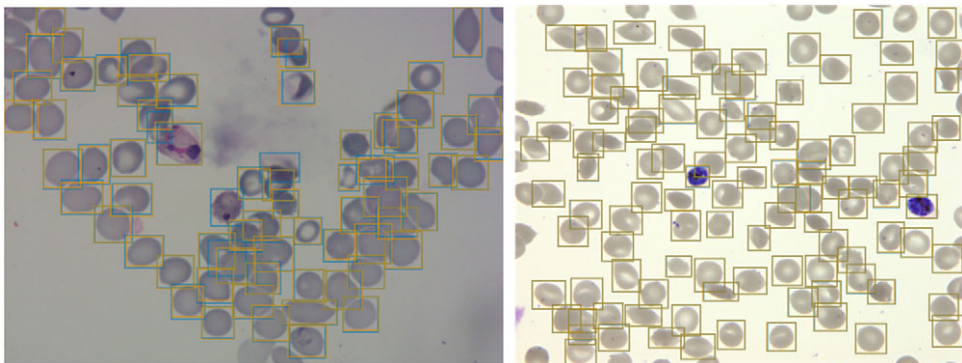
Figure 2. *Object detection of red blood cells (RBCs) on thin smear images. Ground truth labels are shown in orange, and predictions are shown in blue. Examples shown for* P. vivax *(left) and* P. falciparum *(right) test data. The object detection model detects individual RBCs with high precision despite the presence of dust, residual stain, and overlap with other cells.*

3.3. Infected and uninfected red blood cell differentiation

Having shown our ability to isolate RBCs with high precision, we next turned our attention to the delineation of infected versus uninfected cells. From this point onward, we only trained on our *P. falciparum* dataset and removed cases labeled as *unsure.* This resulted in a dataset of 27529/9006 RBCs for training/testing with an 80/20 training/validation split. RBCs were cropped and scaled to 70×70-pixel resolution. We trained a ResNet-50 architecture[29] pretrained on the ImageNet dataset[30] to classify infected RBCs. To address the heavy class imbalance, training data were undersampled at each epoch (the number of times that the learning algorithm works through the entire training dataset) to generate a 1:1 ratio of infected:uninfected RBCs. Using this approach, we obtained an accuracy of 0.998 with an area under the receiver operating characteristic curve of 0.979 on the test set with only RBCs that all annotators agreed on included (99% of RBCs; Table 1 and Supplementary Figure S3). Most misclassifications observed were false positives due to debris on top of an uninfected RBC or the presence of neighboring infected cells. The model was also more likely to call an RBC infected than uninfected when annotators disagreed, and this likelihood generally increased as more annotators agree on infection (Supplementary Figure S4). Moreover, compared to cases where all other annotators agreed, three annotators achieved accuracies of 0.998, two of 0.999, and one of 1.000, demonstrating that our model's error is indistinguishable from the error of an annotator.

3.4. Life cycle stage classification

Having established a model for the classification of infected versus uninfected RBCs in cropped images, we next sought to further classify infected cells into the different parasite IDC stages and gametocytes. Rather than expanding our existing classifier to include the stages as discrete classes, we instead approached progression through the IDC as a continuous process where the stage exists on a numeric scale (ring = 1, trophozoite = 2, and schizont = 3). To achieve this, we adapted a pretrained ResNet-34 classifier[28] to a regression model as described in the Methods section. We then trained on a dataset of 1022/356 RBCs (80/20 training/validation) with cases marked as *unsure* or *multiple infections* by one of the annotators discarded. Using this model, we achieved a root-mean-square error of 0.23 in the test set, with most errors observed between late trophozoite stages and early schizont stages (Figure 3a). High levels of disagreement were found between annotators (Figure 3a and Supplementary Figure S5), and most of the model predictions generally fell within these label boundaries. Moreover, the model provided an ordering of the infected RBCs following the IDC (Figure 3b). Thus, we are able to describe the IDC with more detail than with a three-class classification approach.

3.5. PlasmoCount: A web tool for automated parasite detection and staging

Having demonstrated high performance across models on an independent test set of Giemsa-stained thin blood films, we next sought to generate a user interface that would allow our ML approach to be applied in different lab settings. Toward this, we developed PlasmoCount: an online web tool that can take as input multiple raw images of a Giemsa-stained thin blood film and outputs measures of parasitemia and parasite life stage development (https://plasmocount.org; Figure 4). The images are run through the object detection model; the resulting RBCs are run through the malaria detection and staging models sequentially. The output contains two interactive sections: a summary section that reports parasitemia and the

Table 1. *Performance of ResNet-50 model on malaria detection in images of segmented red blood cells. Metrics were calculated only for images on which all annotators agreed.*

Accuracy	Sensitivity	Specificity	AUC	F1
0.998	0.995	0.998	0.979	0.976

Abbreviation: AUC, area under the receiver operating characteristic curve.
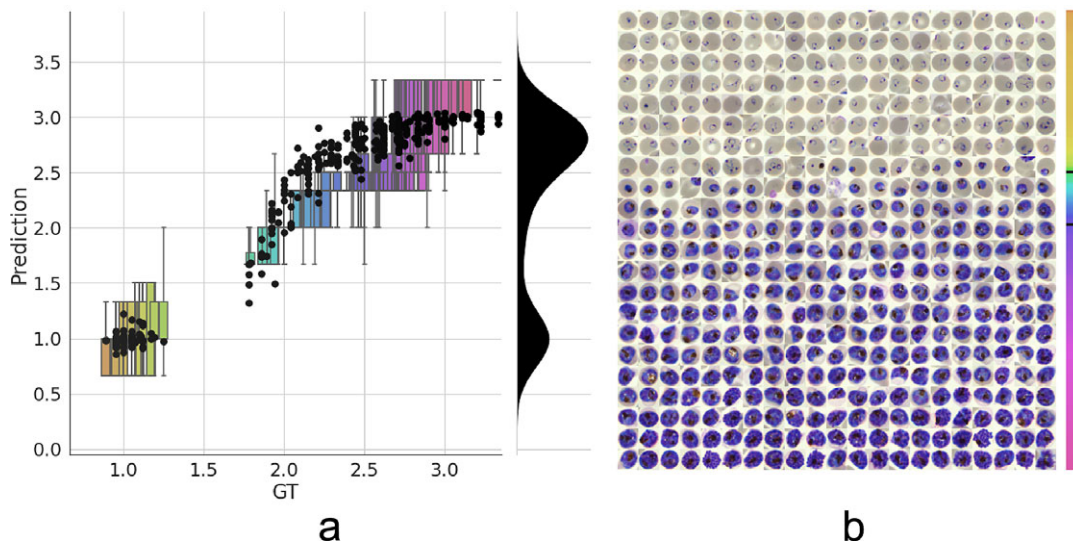
**Figure 3.** *Model prediction on parasite intraerythrocytic cycle (IDC) development. (a) Model predictions are marked with a black dot. Labels were converted to a numeric scale (ring = 1, trophozoite = 2, and schizont = 3) and averaged across all annotators to set a ground truth (GT) label. Boxplots show the label distribution across annotators with error bars determined by the outermost data values. Density plot shows the predicted life stage distribution within the sample. Colors represent progression through the IDC as defined by the GT. (b) After learning from the averaged GT labels, the model successfully orders all detected infected RBCs in the independent test set based on its intraerythrocytic cycle (IDC) life stage predictions (left to right; top to bottom). Color bar represents progression through the IDC as predicted by the model. Top and bottom black lines represent the arbitrary cutoff points used by PlasmoCount between the ring and trophozoite (cutoff = 1.5) and trophozoite and schizont (cutoff = 2.5) stages, respectively.*

IDC stage distribution in the sample and a table with results on the individual images. The interactive component within the summary section allows users to change the histogram bin size of the histogram and click on the individual bins to display the corresponding RBCs for interpretation of the life stage discrimination. The table describes, for each image, the number of RBCs, parasitemia percentage, and the number and fractions of the different IDC and gametocyte stages of the parasite, and data can be exported in CSV format. Moreover, the user can click on the individual rows to display the model predictions for each image as a measure of quality assurance. The web tool has been optimized for mobile phones in conjunction with images directly taken with the phone camera or a cell phone microscope system. Beta-testing with users from different labs, testing for accuracy and utility, was found to perform well using conventional images (Supplementary Figure S6) or using a mobile phone looking down the microscope even without an adaptor (Supplementary Figure S7). Combining the ML model approach with the utility of a mobile interface-based web tool demonstrates the power of ML for identifying and classifying parasites and its ability to rapidly assist in the assessment of parasite cultures via cytological smears.

4. Discussion

Here, we present an ML approach and a user-orientated web tool for the detection and staging of *P. falciparum* from Giemsa-smeared cytological smears. Using this simple approach, we obtain state-of-the-art performance in cell segmentation and detect parasites with high performance on an independent test set. Our object detection model can detect RBCs with an AP of 0.99 on our *P. falciparum* test set and
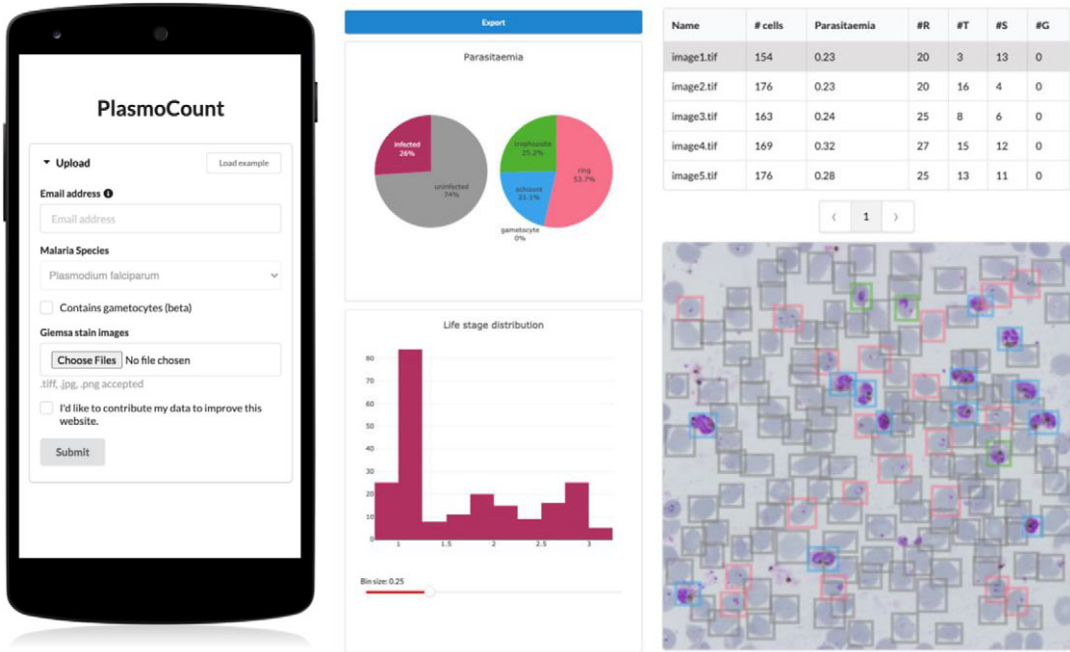
Figure 4. *PlasmoCount, a mobile friendly tool for automated assessment of Giemsa-stained cytological smears. In the upload form, users can attach their images of Giemsa-stained thin blood films. The client sends the data to the server to be passed to the models; the results then get sent back and are displayed in a summary section and a table. The summary section is divided into parasitemia pie charts and a histogram of the IDC life stage distribution. The rows in the table correspond to individual images and are selectable; this will display the corresponding image with overlaying model predictions. The user interface is optimized for mobile phone dimensions, allowing users to use a mobile device for both image acquisition and data analysis. Every job has a unique ID associated with it; this allows users to come back to their results from multiple devices at any time.*

an AP of 0.98 on a *P. vivax* dataset previously tested with the novel Keras R-CNN model (mAP = 0.78)[23] and a CellProfiler pipeline (mAP = 0.61)[23,36] at an IOU threshold of 0.5. Moreover, our classification model of infected RBCs reaches a near-perfect accuracy with a sensitivity of 0.995 and a specificity of 0.998, respectively; this is comparable to state-of-the-art methods for the classification of segmented RBCs[20,37,38]. Indeed, we could further improve our method by replacing our object detection method with a Mask R-CNN model where each cell is overlaid with segmentation masks rather than bounding boxes. This approach has been successfully applied in nucleus segmentation[39,40] and would reduce the confounding of neighboring cells on parasite detection. Finally, we have introduced a staging model that recapitulates IDC progression of the parasite and allows for more refined interpretation when reading slides. Recent studies have aimed to enhance IDC life stage classification by adding early, mid, and late substages[25]. Using a regression model, we are able to differentiate the IDC stage of the parasite more precisely and prevent penalizing predictions that have been classified as a different substage of the parasite but are developmentally not far removed.

We have shown that whereas experts disagree on parasite stages, using an automated approach to malaria detection enables robust standardization in the evaluation of Giemsa smears, providing quantitative measures, for example, when applied to high-throughput experiments. We have defined the ground truth (GT) as a summary of labels from multiple annotators and laboratories. Using this method, our model is able to generally order the developmental stages of individual cells and recapitulate the IDC life

cycle of *P. falciparum.* However, the GT could be further refined by including more annotators and allowing annotators to select from a wider range of stages. Moreover, our pipeline could be supplemented by asking annotators to order cells as part of the labeling process as this has been shown to improve consensus among annotators in fluorescent imaging data of *P. falciparum*[26]. It should be noted, however, that there is as yet no method of validating this GT; ultimately, it is defined by the experts who evaluate the slides, and this phenomenon has led groups to correct standard datasets[41]. Methods will have to establish "gold standard" slide/image sets, not only for the assessment of reader competency, but also to ensure a better benchmark for the design of automated methods.

In this work, we have focused on discriminating between the IDC stages of the parasites as these asexually replicating stages cause symptoms in patients and are the main focus of clinical studies[42] as well as providing the source population for gametocyte investment[43]. As a proof of concept, we have shown that we are also able to discriminate between asexual and mature sexual blood-stage parasites using ML methods. In the future, we would like to extend our regression approach to the sexual stages of the parasite, including progression through the five distinct stages of gametocyte development and differentiation into males and females. We have also not addressed cases where multiple parasites invade RBCs, which is likely to confound the IDC stage prediction. Possible solutions include adding a local classifier for multiplicity, multilabel classification with the IDC stages, and object detection for detecting individual parasites.

Although we have developed PlasmoCount to detect *P. falciparum*, it is by no means limited to one species of *Plasmodium.* Preliminary testing suggests that the tool can also be applied to the study of round-the-clock blood samples from infections with rodent malaria parasites (*P. chabaudi adami,* clone DK), although this will need to be validated against a GT based on expert annotations (Supplementary Figure S8). Leukocytes are not typically present in our laboratory smears of *P. falciparum,* which generally use only processed RBCs from donated whole blood. We have, however, aimed to address the possibility of leukocyte presence by including a threshold based on the size distribution of cells in the sample, but a more sensitive ML-based method could be added in the future. Such an approach will be fundamental, as we aim to extend our method to the clinical diagnosis of human malaria where leukocytes, other components of whole blood (e.g., platelets), and even other pathogens may add further classes of non-RBC material to smears.

Automated methods can speed up the evaluation of thin blood smears; however, the bottleneck of image acquisition remains. In this study, we sought to provide a solution to the speed of acquisition by developing a tool that works in conjunction with mobile devices from image acquisition to assessment. If coupled with low-cost and portable microscope systems[44–46], this has the potential to increase screening throughput markedly and opens up opportunities for thin blood smears as a tool for qualitative diagnosis in the field where thick smears are preferred due to their higher sensitivity. Speed of image acquisition via a mobile device will generally be a lot faster than manual counting and recording of thin smears. There is also the possibility that PlasmoCount could be incorporated into higher throughput image acquisition system workflows (e.g., slide readers) to speed up accurate parasite quantification in a hospital/laboratory-based setting. It can also be downloaded and locally installed for regions where Internet stability is an issue (Supplementary Figure S9, Supplementary Dataset S10 [test dataset], and Supplementary Table 2).

There are limitations to be considered especially in the use of ML methods in clinical practice. Although we have tested a dataset from independent research centers, measuring generalization performance remains challenging due to technical differences between laboratories, and performance will heavily depend on data quality. Moreover, our method does not estimate uncertainty with its predictions which would improve the reliability of the results. To this end, we have developed a web tool that lets the user check their results for quality assurance and save them for traceability in the decision-making process. This also opens up an opportunity for continuous improvement of the models; by collecting feedback from the community, a human-in-the-loop labeling workflow could be implemented similar to our model-assisted approach.

Recent efforts have applied ML techniques to the detection of malaria parasites in microscopy images of cytological smears. The aim of this work was not only to further these developments, but also to provide a tool that can be used by the malaria research community. Deep learning typically requires large amounts of data, which are not always widely available for medical applications due to the expert knowledge required. In this study, we have included data from six research centers and tested the tool's functionality with various others. In this way, we enhance robustness across laboratories and aim to drive a community-wide effort to accelerate malaria research and ultimately adopt automated methods in a clinical diagnostic context.

Data Availability Statement.   Original imaging data, experimental metadata, and pseudolabels are available from a publicly accessible Data Repository (http://dx.doi.org/10.17632/j55fyhtxn4.1). Source code is available at https://github.com/thebaumlaboratory/PlasmoCount. Login details to access PlasmoCount are available at https://www.baumlab.com/plasmocount. To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S2633903X21000015.

Authorship Contributions.   M.S.D. and J.B. designed the experiments and wrote the initial manuscript; S.Y., J.C., D.W.W., P.P., C.U., M.D.J., M.P., D.W.A., M.J.B., P.G., J.D.D., A.J.O., and S.E.R. cultured parasites, and stained and acquired images; M.S.D., G. W.A., M.D., and D.M.A. developed the imaging workflow; M.S.D., M.D., and D.M.A. optimized and carried out the ML workflow and data analysis. C.A.B. and K.A.C. undertook deployment of the web tool and migration of software to cloud-based services. All authors contributed to writing and editing of the manuscript.

References

1. WHO (2019) *World Malaria Report 2019.* World Health Organization.
2. Tangpukdee N, Duangdee C, Wilairatana P & Krudsood S (2009) Malaria diagnosis: a brief review. *Korean J Parasitol* **47**, 93–102.
3. Barcia JJ (2007) The Giemsa stain: its history and applications. *Int J Surg Pathol* **15**, 292–296.
4. Trager W & Jensen JB (1976) Human malaria parasites in continuous culture. *Science* **193**, 673–675.
5. WHO (2016) Malaria Microscopy Quality Assurance Manual—Version 2. Geneva, Switzerland: World Health Organization.
6. O'Meara WP., McKenzie, F. E., Magill, A. J. *et al.* (2005) Sources of variability in determining malaria parasite density by microscopy. *Am J Trop Med Hyg* **73**, 593–598.
7. Durrheim DN, Becker PJ & Billinghurst K (1997) Diagnostic disagreement—the lessons learnt from malaria diagnosis in Mpumalanga. *S Afr Med J* **87**, 1016.
8. McKenzie FE, Sirichaisinthop J, Miller RS, Gasser Jr RA & Wongsrichanalai C (2003) Dependence of malaria detection and species diagnosis by microscopy on parasite density. *Am J Trop Med Hyg* **69**, 372–376.
9. Poostchi M, Silamut K, Maude RJ, Jaeger S & Thoma G (2018) Image analysis and machine learning for detecting malaria. *Transl Res* **194**, 36–55.
10. Ruberto CD, Di Ruberto C, Dempster A, Khan S & Jarra B (2002) Analysis of infected blood cell images using morphological operators. *Image Vis Comput* **20**, 133–146.
11. Tek FB, Dempster AG & Kale I (2006) Malaria parasite detection in peripheral blood images. In *Proceedings of the 2006 British Machine Vision Conference*, pp. 36.1–36.10 [M Chantler, B Fisher & M Trucco, editors]. Edinburgh, UK: BMVA Press. https://doi.org/10.5244/c.20.36.
12. Ross NE, Pritchard CJ, Rubin DM & Dusé AG (2006) Automated image processing method for the diagnosis and classification of malaria on thin blood smears. *Med Biol Eng Comput* **44**, 427–436.
13. Díaz G, González FA & Romero E (2009) A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images. *J Biomed Inform* **42**, 296–307.
14. Tek FB, Dempster AG & Kale İ (2010) Parasite detection and identification for automated thin blood film malaria diagnosis. *Comput Vis Image Underst* **114**, 21–32.

15. Linder N, *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**, e104855.

16. Liang Z, *et al.* (2016) CNN-based image analysis for malaria diagnosis. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 493–496. https://doi.org/10.1109/bibm.2016.7822567.

17. Dong Y, *et al.* (2017) Evaluations of deep convolutional neural networks for automatic identification of malaria infected cells. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp, 101–104. https://doi.org/10.1109/bhi.2017.7897215.

18. Gopakumar GP, Swetha M, Sai Siva G & Sai Subrahmanyam GRK (2018) Convolutional neural network-based malaria diagnosis from focus stack of blood smear images acquired using custom-built slide scanner. *J Biophotonics* **11**. https://doi.org/10.1002/jbio.201700003

19. Bibin D, Nair MS & Punitha P (2017) Malaria parasite detection from peripheral blood smear images using deep belief networks. *IEEE Access* **5**, 9099–9108.

20. Rajaraman S, *et al.* (2018) Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ* **6**, e4568.

21. Hung J & Carpenter A (2017) Applying faster R-CNN for object detection on malaria images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 56–61. openaccess.thecvf.com

22. Yang F, *et al.* (2020) Deep learning for smartphone-based malaria parasite detection in thick blood smears. *IEEE J Biomed Health Inform* **24**, 1427–1438.

23. Hung J, *et al.* (2020) Keras R-CNN: library for cell detection in biological images using deep neural networks. *BMC Bioinformatics* **21**, 300.

24. Dhorda M, *et al.* (2020) Towards harmonization of microscopy methods for malaria clinical research studies. *Malar J* **19**, 324.

25. Abbas SS & Dijkstra TMH (2020) Detection and stage classification of *Plasmodium falciparum* from images of Giemsa stained thin blood films using random forest classifiers. *Diagn Pathol* **15**, 130.

26. Ashdown GW, *et al.* (2020) A machine learning approach to define antimalarial drug action from heterogeneous cell-based screens. *Sci Adv* **6**, eaba9338.

27. Labelbox (2020) Labelbox: the leading training data platform for data labeling. https://labelbox.com/ (accessed October 26, 2020).

28. Ljosa V, Sokolnicki KL & Carpenter AE (2012) Annotated high-throughput microscopy image sets for validation. *Nat Methods* **9**, 637.

29. He K, Zhang X, Ren S & Sun J (2016) Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

30. Deng J, *et al.* (2009) ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

31. Kingma DP & Ba J (2014) Adam: a method for stochastic optimization. *Preprint*, 2017, arXiv:1412.6980.

32. Smith LN (2018) A disciplined approach to neural network hyper-parameters: Part 1—learning rate, batch size, momentum, and weight decay. *Preprint*, 2018, arXiv:1803.09820.

33. Ren S, He K, Girshick R & Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. **39**, no. 6, pp. 1137–1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

34. Girshick R (2015) Fast R-CNN. *Preprint*, 2015, arXiv:1504.08083.

35. Lin T-Y, *et al.* (2014) Microsoft COCO: common objects in context. In *Computer Vision—ECCV 2014*. Lecture Notes in Computer Science, 8693, pp. 740–755 [D Fleet, T Pajdla, B Schiele & T Tuytelaars, editors]. Cham: Springer.

36. McQuin C, *et al.* (2018) CellProfiler 3.0: next-generation image processing for biology. *PLoS Biol* **16**, e2005970.

37. Rahman A, *et al.* (2019) Improving malaria parasite detection from red blood cell using deep convolutional neural networks. *Preprint*, 2019, arXiv:1907.10418.

38. Rajaraman S, Jaeger S & Antani SK (2019) Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. *PeerJ* **7**, e6977.

39. Jung H, Lodhi B & Kang J (2019) An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images. *BMC Biomed Eng* **1**, 24.

40. Johnson JW (2020) Automatic nucleus segmentation with Mask-RCNN. In *Advances in Computer Vision*. Advances in Intelligent Systems and Computing, 944, pp. 399–407 [K Arai & S Kapoor, editors]. Cham: Springer.

41. Fuhad KMF, *et al.* (2020) Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application. *Diagnostics (Basel)* **10**, 329.

42. Venugopal K, Hentzschel F, Valkiūnas G & Marti M (2020) *Plasmodium* asexual growth and sexual development in the haematopoietic niche of the host. *Nat Rev Microbiol* **18**, 177–189.

43. Carter LM, *et al.* (2013) Stress and sex in malaria parasites: why does commitment vary? *Evol Med Public Health* **2013**, 135–147.

44. Diederich B, *et al.* (2020) A versatile and customizable low-cost 3D-printed open standard for microscopic imaging. *Nat Commun* **11**, 5979.

45. Orth A, Wilson ER, Thompson JG & Gibson BC (2018) A dual-mode mobile phone microscope using the onboard camera flash and ambient light. *Sci Rep* **8**, 3298.

46. Diederich B, Then P, Jügler A, Förster R & Heintzmann R (2019) cellSTORM—cost-effective super-resolution on a cellphone using dSTORM. *PLoS One* **14**, e0209827.