

1

2 **Applications of Digital Microscopy and Densely Connected Convolutional Neural Networks for**
3 **Automated Quantitation of Babesia-Infected Erythrocytes**

4 Thomas JS Durant^{1*}; Sarah Dudgeon^{2,3}; Jacob McPadden⁴; Anisia Simpson⁵; Nathan Price⁶; Wade
5 Schulz^{1,2,6}; Richard Torres¹; Eben M Olson¹

6

7 1 – Department of Laboratory Medicine, at Yale School of Medicine, New Haven, CT.

8 2 – Center for Outcomes Research and Evaluation, at Yale New Haven Hospital, New Haven, CT.

9 3 – Biological and Biomedical Sciences, at Yale University, New Haven, CT.

10 4 – Department of Neonatology, at Yale School of Medicine, New Haven, CT.

11 5 – Department of Laboratory Medicine, at Yale New Haven Hospital, New Haven, CT.

12 6 – Center for Computational Health, at Yale New Haven Hospital, New Haven, CT.

13

14 * To whom correspondence should be addressed:

15 Thomas JS Durant

16 Department of Laboratory Medicine,

17 55 Park Street PS345D, New Haven, CT 06511.

18 Tel.: 203-688-2301; E-mail: thomas.durant@yale.edu

19

20 **Abbreviations:**

21 AUC: Area Under the Curve

22 FN: False Negative

23 FP: False Positive

24 IG: Integrated Gradient

25 MLS: Medical Laboratory Scientist

26 MLS-RS: Medical Laboratory Scientist-Reference Standard

27 RBC: Red blood cell

28 TN: True Negative

29 TP: True Positive

30 XAI: Explainable Artificial Intelligence

31

32

33 Running title: Machine Learning for Babesia Quantitation

34 Disclaimers: None

35 Support: Academy of Clinical Laboratory Physicians and Scientists: Paul E. Strandjord Young Investigator Research Grant

36 Keywords: machine learning, convolutional neural networks, peripheral blood smear, erythrocyte, red blood cells, babesia,
37 image analysis.

38

39 **Background:** Clinical babesiosis is diagnosed, and parasite burden is determined, by microscopic
40 inspection of a thick or thin Giemsa-stained peripheral blood smear. However, quantitative analysis by
41 manual microscopy is subject to observer bias, slide distribution errors, statistical sampling error,
42 recording errors, and is inherently burdensome from time management and workflow efficiency
43 standpoints. As such, methods for the automated measurement of percent parasitemia in digital
44 microscopic images of peripheral blood smears could improve clinical accuracy, relative to the predicate
45 method.

46 **Methods:** Individual erythrocyte images (shape: 70x70x3) were manually labeled as “parasite” or
47 “normal” and were used to train a model for binary image classification. The best model was then used
48 to calculate percent parasitemia from a clinical validation dataset, and values were compared to a
49 clinical reference value. Lastly, model interpretability was examined using an integrated gradient to
50 identify pixels most likely to influence classification decisions.

51 **Results:** The precision and recall of the model during development testing were 0.92 and 1.00,
52 respectively. In clinical validation, the model returned increasing positive signal with increasing mean
53 reference value. However, there were two highly erroneous false positive values returned by the model.
54 Lastly, the model incorrectly assessed three cases well above the clinical threshold of 10%. The
55 integrated gradient suggested potential sources of false positives including rouleaux formations, cell
56 boundaries, and precipitate as deterministic factors in negative erythrocyte images.

57 **Conclusions:** While the model demonstrated highly accurate single cell classification and correctly
58 assessed most slides, several false positives were highly incorrect. This project highlights the need for
59 integrated testing of ML-based models, even when models in the development phase perform well.

60

61 **INTRODUCTION:**

62 Clinical Babesiosis is a haemoprotozoan disease that is most commonly transmitted from animals to
63 humans by invertebrate vectors (e.g., *Ixodes scapularis*, the black legged deer tick)(1). In the United
64 States, 95% of cases occur in the Northeast and Upper Midwest states, occurring primarily between May
65 and October. In the state of Connecticut, the seroprevalence has been shown to range between 0.3-
66 17.8%, with the number of reported cases being approximately 44 per 100,000 (2). Disease severity can
67 range from asymptomatic to severe, the latter of which may lead to life-threatening scenarios. Severe
68 disease is more common in specific at-risk populations including those who are post-splenectomy,
69 immunocompromised, or older than 50 years of age. The all-cause mortality of babesiosis has been
70 estimated as <1% for clinical cases, and approximately 10% for iatrogenic cases (e.g., transfusion-
71 transmitted) (2).

72 The diagnostic gold standard for babesiosis is microscopic inspection of thick, or thin, Giemsa-
73 stained peripheral blood smear (1). If *Babesia* spp is identified, the degree of parasitemia is used to
74 guide patient management strategies. For mild disease, or minimal parasitemia, antimicrobials are the
75 preferred therapy. However, the American Society for Apheresis (ASFA) guidelines state that severe
76 babesiosis is a category II indication for red blood cell (RBC) exchange. Severe disease is determined
77 both by clinical and laboratory criteria including significant parasitemia (e.g., >10%), the presence of
78 comorbidities (e.g., asplenia), or severe symptoms such as, disseminated intravascular coagulation or
79 multiorgan failure (2). While there is no consensus on when to discontinue RBC exchange, it is
80 recommended that patients with severe babesiosis be monitored closely, with parasitized erythrocytes
81 quantified daily alongside continued RBC exchange until parasite burden decreases below 5% (2,3).

82 Percent parasitemia is the quotient of parasite-infected erythrocytes over the number of total
83 erythrocytes counted. To derive this in a clinical laboratory, the process commonly involves a medical
84 laboratory scientist (MLS) counting a large number of erythrocytes (e.g., 1,000) using a 100x oil-

85 immersion objective. While this process requires minimal laboratory equipment, it does require an
86 experienced MLS to ensure optimal accuracy and reproducibility for serial measurement purposes (1). In
87 addition, quantitative analysis by manual microscopy is subject to observer bias, slide distribution errors,
88 statistical sampling error and recording errors, and is inherently burdensome from time management
89 and workflow efficiency standpoints (4,5). Such limitations can mislead or delay therapeutic decision
90 making, particularly in the context of therapeutic RBC exchange. Accordingly, there remains a significant
91 need to develop automated methods to optimize the cost, efficiency, and accuracy of quantitative
92 analysis.

93 The progress made in computer vision and machine learning (ML) technology over the last
94 decade has encouraged a corresponding increase in their implementation in the clinical laboratory (6).
95 With the decreasing availability of experienced medical laboratory scientists, evaluating ML-based
96 software capabilities without expert operator review remains an important consideration in study
97 design (7,8). To this end, we sought to develop and evaluate the accuracy of a an ML-based method for
98 the automated measurement of percent parasitemia in digital microscopic images of peripheral blood
99 smears. Specifically, we sought to describe the accuracy of parasitemia measurements, as determined
100 by ML-based software, relative to an MLS-derived reference standard (MLS-RS). We hypothesized that
101 results generated by the ML-based software would show superior precision to MLS-RS while achieving
102 clinically comparable numerical results to the average MLS-RS.

103

104 **METHODS:**

105 *Hardware and Operating Systems:*

106 Computation for model training was performed on a local Linux server (NVIDIA DGX Server Version
107 4.6.0) (GNU/Linux 4.15.0-122-generic x86_64) running Ubuntu (version: 18.04.5 LTS). Processing

108 hardware included 80 CPUs (Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz) and 8 GPUs (Tesla V100-
109 SXM2-16GB) using CUDA Toolkit (version: 11.0).

110

111 *Data Set Curation:*

112 This study has been reviewed and approved by the Yale University Internal Review Board (IRB#
113 2000020244). Clinical blood samples were originally collected as part of routine clinical workflow in
114 lavender-top (EDTA) tubes for screen and quantitation of *Babesia* spp. Slides and concomitant digital
115 images of the associated peripheral blood smears, which were found by to be positive for *Babesia* spp
116 and negative for *Malaria* spp (BinaxNOW Malaria; Abbott, Chicago, IL), were flagged for inclusion using
117 previously described methods (9–11). Slides and the concomitant digital images of *Babesia*-negative
118 samples were collected from the routine clinical workflow throughout the study period and reviewed by
119 a clinical pathologist for the absence of *Babesia* spp prior to inclusion.

120 Slides and images were separated into two distinct groups, representing separate patient
121 cohorts: (1) The model development dataset and (2) the clinical validation dataset. The model
122 development dataset was used for training, validation, and preliminary evaluation of the cell
123 classification model. The clinical validation dataset was used as a second, ‘external’ validation dataset to
124 evaluate how the model would perform in a clinical implementation workflow, as compared to a
125 predicate method-based reference standard.

126 All peripheral blood smears were created and imaged on a DI-60 Integrated Slide Processing
127 System (Cellviation AB, Lund, Sweden). The DI-60 uses a 100X-objective and a 0.5X magnifier prior to
128 imaging, rendering an effective magnification of 50X. Images are 3-channel RGB, with a resolution of 5
129 pixels per micron. In the model development dataset, slide images had an average height and width of
130 2884 pixels (95% CI: 2882-2885) and 2867 pixels (95% CI: 2865-2868) (Figure 1A). Slides included in the
131 model development dataset were imaged a single time. Slides included in the clinical validation dataset

132 were imaged three times on the same scanner to compute intra-precision for quantitation of *Babesia*
133 spp during subsequent portions of the study.

134

135 *Cell Labeling for Model Development Dataset:*

136 Slide-level images from the model development dataset were uploaded to a custom-built web
137 application for labeling of individual erythrocytes using one of two labels: (1) parasite or (2) normal.
138 Using the web application, annotators marked central X-Y coordinates of infected and non-infected
139 erythrocytes (Figure 1B). X-Y coordinates of cell centers were then used to crop individual erythrocytes
140 from the slide-level parent image into 70x70 pixel, 3-channel image arrays. These 70x70x3 images were
141 then paired with their corresponding label of either 'parasite' or 'normal' (Figure 1C). The labeling
142 process was performed by a single laboratory medicine attending and author of this manuscript (TJD).
143 As a post-processing step, X-Y coordinates which were within 140 pixels of another set of X-Y
144 coordinates were removed from the dataset following completion of the annotation process. This was
145 done to ensure that there was no overlap of images in the final development dataset which, if present,
146 could have resulted in part of an image being represented in both the training and validation and test
147 datasets, leading to overfitting, or an over-optimistic estimate of model performance.

148 Ultimately, the final dataset used for model development consisted of non-overlapping,
149 individual erythrocyte images (shape: 70x70x3) with an associated label of 'parasite' or 'normal'. These
150 data were split and used to train, validate, and test the image classification model. The model
151 development dataset was divided 80:20 into train and test datasets, respectively (Figure 1D). The train
152 dataset was further subdivided 70:30 into train and validation datasets, respectively. The train and
153 validation datasets were used during the training of the image classification model (Figure 1E). The test
154 dataset was used to evaluate model performance following completion of training (Figure 1F).

155

156 *Network Implementation:*

157 For image classification, the authors implemented DenseNet121 as the base model, initialized with
158 pretrained weights from ImageNet (7). Densely connected neural networks were first described by
159 Huang et al. and are a commonly used architecture for learning image classification tasks (8). This neural
160 network was chosen based on previously published performance metrics comparable with current state
161 of the art models, and because it uses relatively fewer parameters, making it faster to train and easily
162 portable (12). Base model layers were not frozen and were configured as trainable. DenseNet121 was
163 combined with a custom set of prediction layers, specific to this image classification task. These included
164 a 2-dimensional global average pooling layer, a dropout layer, and a densely connected layer with
165 sigmoid activation function for binary classification. The Adam method was used for gradient-based
166 optimization. In total, there were 7,038,529 parameters, 6,954,881 of which were trainable. The
167 network was implemented using Tensorflow (version: 2.4.0rc0), Tensorflow-gpu (version: 2.4.0rc0) and
168 Python (version: 3.6.9).

169

170 *Model Development Protocol:*

171 Train dataset images were subjected to label preserving augmentation prior to being served as input to
172 the network. Image augmentation included random horizontal and vertical flips, random rotation,
173 random translation, random zoom, random contrast adjustments, and random brightness adjustments.
174 Lastly, due to the imbalanced nature of our training dataset the 'parasite' class was oversampled to
175 produce a 1:1 ratio of parasite and normal images during training. The network was trained for a total of
176 50 epochs (i.e., iterations) over the complete training dataset. The validation dataset was used to

177 monitor model performance during training for subsequent tuning according to the calculated binary
178 cross-entropy loss. Model parameters were saved following a reduction in the binary cross-entropy loss,
179 calculated from the validation dataset after each epoch. The initial learning rate was set to 1e-5 and
180 decreased by a factor of 10 if validation loss did not improve after 5 epochs. The total training process
181 was repeated three times using unique random seed initializers to evaluate variability in train
182 performance metrics. Performance metrics monitored during training included true positives (TP), false
183 positives (FP), true negatives (TN), false negatives (FN), binary accuracy, precision (i.e., positive
184 predictive value) ($TP / (TP + FP)$), recall (i.e., sensitivity) ($TP / (TP + FN)$), and area under the receiver
185 operator characteristic curve (AUC). These were calculated on both train and validation datasets
186 following the completion of each epoch. Following model training, the best model parameters (i.e.,
187 those which achieved the lowest validation loss) were used to evaluate individually labeled erythrocytes
188 in the test dataset. Cells with a probability score greater than or equal to 0.5 were assigned 'parasite'
189 prediction labels. Test predictions were then used to calculate the performance metrics for the test
190 dataset. Similarly, the 'best model' was used to evaluate cells in the clinical validation protocol.

191

192 *Clinical Validation Protocol:*

193 Following model development, a separate set of peripheral blood smear slides were used to assess the
194 accuracy of the model in a simulated clinical workflow. Due to the inherent variability seen with
195 quantitative analysis by microscopy, a clinical reference standard consisting of multiple measurements
196 was compiled for comparisons between the model and the predicate method. Accordingly, each glass
197 slide in the clinical validation dataset was independently evaluated by three MLS's with 26, 6, and 4
198 years of experience for MLS A, B, and C, respectively. The clinical validation slides were shuffled,
199 specimen numbers on the glass slides were covered, and a box containing the clinical validations slides

200 was given to each of the MLS' for independent evaluation. Each MLS evaluated all clinical validation
201 slides three separate times (Figure 2A). In total, this process generated 9 results of percent parasitemia
202 for each slide in the clinical validation dataset. These data were used to calculate the average percent
203 parasitemia across all 9 reads which was used as the MLS-RS for each case/sample (Figure 2B). Of note,
204 the lower limit of quantitation for percent parasitemia in the clinical laboratory at our institution is 1%
205 and results below this value are reported out as <1% in routine practice. For the purposes of this study,
206 MLSs were asked to record the precise parasitemia value, including those below 1%, to allow for a
207 completely empirical comparison against the model.

208 For the model-based method, as mentioned, each slide in the clinical validation dataset was
209 scanned three separate times by the DI-60 (Figure 2C). A custom cell-segmentation script was then used
210 to crop individual erythrocytes from the peripheral blood smear image (Figure 2D). Cell-segmentation
211 was implemented using OpenCV (version: 4.2.0.34) using contour-based (`cv.findContours()`). Individual
212 erythrocytes (shape: 70x70x3) were then provided as input to the best model, as defined in the
213 development protocol, to yield a predicted class (i.e., 'parasite' or 'normal') for each individually
214 cropped erythrocyte (Figure 2E). Following classification of individual erythrocytes, the number of cells
215 with the predicted label of 'parasite' were divided by number of total cells classified to yield the
216 quantification of percent parasitemia. This process was done one time for each image with three images
217 per specimen, yielding a total of 3 parasitemia results per slide (Figure 2F).

218 Method-to-method (i.e., accuracy) comparisons between the model and MLS-RS percent
219 parasitemia were made using a variety of approaches: (1) bar plot visualization; (2) regression and
220 Bland-Altman plots; (3) quantitative agreement of model percent parasitemia in relation to ± 2 SD of the
221 average MLS-RS percent parasitemia (n=9) for each case in the clinical validation dataset; (4) categorical
222 agreement of percent parasitemia bins; (5) categorical agreement around the clinical decision threshold

223 of 10%. Precision was assessed using the coefficient of variation, which was calculated on a case-wise
224 basis across the MLS (n=9) and model results (n=3).

225

226 *Model Interpretability:*

227 In an effort to examine the relationship between model predictions and image features, we
228 implemented an explainable artificial intelligence (XAI) technique based on axiomatic attribution for
229 deep networks and known as Integrated Gradients (IG) (13). While the methods of IG are outside the
230 scope of this report, the general purpose is to identify pixels within each image which most heavily
231 influence a model's prediction, and derived from the gradient (i.e., slope or derivative) of the prediction
232 function relative to each feature (i.e., pixel). For the purposes of this report we attempted to provide
233 representative samples of what we observed when reviewing the images derived from an IG
234 implementation. This was done on the test images in the model development dataset.

235

236 **RESULTS:**

237 *Dataset Curation:*

238 A total of 96 unique slides were included in this study. Of these, 71 slides were included in the
239 development dataset, 28 of which were found to be positive for *Babesia* spp by routine clinical
240 workflow. A total of 14,633 individual erythrocyte images were initially labeled. Of those, 2,019 images
241 that had overlapping cells were removed, yielding a final development dataset of 11,388 erythrocytes
242 labeled as normal and 1,226 with a parasite. The mean number of labeled cells per unique slide was 178
243 (SD 63; range 1-286). Of the slide-level images which were *Babesia*-positive, the mean parasitemia was
244 6.5% (SD 4.5; range 1.0-20.0). The clinical validation dataset consisted of the remaining 25 slides, of
245 which 64% (n=16) were *Babesia*-positive. The mean parasitemia among the *Babesia*-positive slides in the
246 clinical validation dataset was 8.9% (SD 9.4; range 1.0-29.2).

247

248 *Model Development:*

249 The cell classification model was trained 3 separate times. Each training replicate consisted of 50 epochs
250 (iterations). Learning rates decayed following validation loss plateau across all training replicates, with
251 the final value ranging from 1e-8 to 1e-9. Minimum validation loss was observed following completion
252 of training epoch 22, 22, and 31 for each of the training replicates, with an average binary cross-entropy
253 of 0.024 (SD 0.003). Binary cross-entropy loss was plotted and inspected for positive divergence of
254 validation loss, relative to training loss, as an empirical indicator of overfitting. This was observed
255 minimally in the later training epochs (Figure 3A). Precision, recall, and AUC for asymptotically
256 approached model performance limits which were concordant with plateaus of validation loss,
257 indicating model improvement to be unlikely to occur with additional training iterations (Figures 3B-D).
258 Training replicate 3 achieved the lowest validation loss during training (0.021) and was subsequently
259 used for evaluation of the test and clinical validation datasets. Model predictions on the test dataset
260 resulted in 20 false positives and zero false negatives. The precision and recall were 0.92 and 1.00,
261 respectively (Figure 4A). The binary classification accuracy was 0.99. The distribution of predicted
262 probabilities for erythrocytes in the test dataset was visualized and demonstrated a predominantly
263 bimodal distribution between the predicted classes (Figure 4B).

264

265 *Clinical Validation of Model-Based Method*

266 A total of 25 unique slides were identified for evaluation in the clinical validation set, 16 of which were
267 found to be positive for *Babesia* spp by routine clinical workflow. Of those 16, one (Case #15) was
268 excluded from analysis, as per the consensus recommendation of the participating MLS' due to
269 excessive artifact, Howell-Jolly bodies, and only rare, dying parasites. The remaining slides were

270 evaluated in three separate instances by each of the MLS' with an average parasitemia ranging from
271 <0.1% to 38.5% (Supplemental Table 1 and Supplemental Figure 1).

272 Model classification demonstrated an increasing positive signal (i.e., higher parasite count) with
273 respect to the MLS-RS; however, the automated model also demonstrated spurious positive signal with
274 the negative cases (Cases 16-25). In addition, the model returned highly erroneous false positive signal
275 on cases 11 and 16, relative to the MLS-RS (Figure 5). A simple linear regression was performed to
276 evaluate the concordance between the MLS-RS and the model predictions. The regression equation was
277 determined as: $4.78 + 0.55x$ with correlation coefficient (R^2) of 0.244 (Figure 6A). With cases 11 and 16
278 removed, the regression equation is calculated as: $1.68 + 0.68x$ with an R^2 of 0.916. Bland-Altman plots
279 were also assessed for bias trends, and similarly demonstrate erroneous positive signal on the low end
280 and erroneously low positive signal on the high end (Figure 6A and 6B).

281 Of the 14 positive cases included in the clinical validation dataset, 10 were within 2 SD of the
282 MLS-RS mean. However, only 7 were concordant between the model and MLS-RS with regards to the
283 percent parasitemia bins. In addition, there were three major errors by the model-based method, which
284 were defined as discordance around the clinical decision point of 10% parasitemia. Of the 14 positive
285 cases, the MLS-RS CV was less than 20% in only 3 cases, whereas the Model CV was less than 20% for 10
286 of the cases (Supplemental Table 2).

287

288 *Model Interpretability:*

289 Cells from the test dataset and the clinical validation dataset were evaluated using the IG approach to
290 visualize feature pixel-level activation patterns. Cells from the test dataset generally demonstrated
291 activation of pixels which were near the intra-erythrocytic parasite (Figure 7). Cells from case 25, a
292 negative case in the clinical validation set, were also examined and demonstrated erroneous activation
293 on non-parasitic features. Some of these features included erythrocyte abnormalities (e.g., target cell

294 contours), precipitate, and overlying platelets. In some cases, the model appeared to be focusing on
295 background pixels which may be indicative of overfitting in some aspects of the model (Figure 8).

296

297 **DISCUSSION**

298 In this report, we describe an approach to quantifying percent-parasitemia in peripheral blood smears
299 using computer vision and machine learning technology. We sought to examine the accuracy of an ML-
300 based solution without the use of expert operator-reclassification. Since the beginning of modern
301 computing, there has been considerable interest in the optimization of peripheral blood smear review,
302 with published efforts for smear image analysis dating back to the 1970's (14,15). While previous
303 attempts yielded variable results, recent improvements in computing hardware have led to significant
304 advancements in performance, particularly in the context of object classification tasks (16). Indeed,
305 there has been a resurgence over recent years investigating the application of machine learning-based
306 technologies for classification, speciation, and quantitative tasks using digital images of the peripheral
307 blood smear (17,18). Automated image analysis tools are becoming increasingly available for peripheral
308 smear analysis, however, the scope of FDA approval is limited and classification algorithms demonstrate
309 suboptimal performance without human reclassification (19,20).

310 We found that in the context of the train-test development cycle, model performance metrics
311 demonstrated highly accurate results. Train and validation loss curves demonstrated minimally
312 appreciable divergence towards the end of training iterations which would imply that there is negligible
313 overfitting with the cell classification model (Figure 3A). The sigmoid activation function used for the
314 classification layer of the model demonstrated good separation between the parasite class and the non-
315 parasite class, with only 20 false positive cells in the test dataset (Figure 4). However, when the model
316 was implemented with contour-based cell segmentation and applied to the clinical validation dataset,
317 method comparison studies with the MLS-RS demonstrated suboptimal concordance with the model-

318 based method. Simple linear regression between the two methods had a calculated correlation
319 coefficient (R^2) of 0.244 and 0.916 with and without outliers, respectively. In addition, only 7 of the 14
320 positive cases were concordant between the model and MLS-RS when grouped by percent parasitemia
321 bins. Lastly, there were three major errors by the model-based method, which were defined as
322 discordance around the clinical decision point of 10% parasitemia (Supplemental Table 2).

323 The root cause of these discrepancies is likely multifactorial and highlights the need to
324 interrogate the performance of ML-based technology beyond the train-test development cycle. In the
325 clinical validation method-to-method comparison, the model returned highly erroneous positive signal
326 with cases 11 and 16, relative to the MLS-RS (Figure 5). These errors were likely driven, in part, by the
327 quality of the blood smear which contained significant amount of precipitate and rouleaux formations.
328 For blood smear images where there was minimal to no rouleaux formation, visual inspection of
329 contour-based cell segmentation suggested adequate performance (Supplemental Figure 2). However,
330 in the context of significant rouleaux formation, cell segmentation resulted in fewer individual cells
331 identified for evaluation (Supplemental Figure 3 and 4). In combination with overlying precipitate, which
332 can be mistaken for intra-erythrocyte parasites, this can result in a high numerator (i.e., false positives)
333 and a low denominator (i.e., fewer individually segmented cells), which led to artificially elevated
334 parasitemia quantification. Future work in this area could explore the use of ML-based approaches to
335 cell segmentation. However, these approaches would theoretically encounter similar barriers when
336 initializing models with coordinates for segmentation training and would need specific considerations
337 for handling rouleaux formations. During the initial stages of this work, we had found there to be little
338 qualitative difference between computer vision and ML-based segmentation for smears when there was
339 minimal rouleaux formation to contend with (data not shown).

340 Model interpretability experiments were used to develop an intuitive sense as to what
341 effectuates the observed model behavior, a limitation being that this method only provides an

342 indication of feature importance on individual images and does not offer a mechanism to provide insight
343 across the entire dataset. It also only explains individual feature contributions, but does not examine
344 how feature interactions may contribute to predictions (21). Nonetheless, these experiments revealed
345 that model predictions of the target class, 'parasite', were generally most impacted by pixels spatially
346 related to intraerythrocytic ring-forms (Figure 7). However, there were instances wherein pixel-wise
347 activation patterns were found to be localized outside of the erythrocyte and corresponding to
348 background noise (Supplemental Figure 5). This would suggest that there is some degree of overfitting
349 which is not obviously appreciable through visual inspection of the train and validation loss curves.
350 Integrated gradients also provided some context as to model fallibility when applied to the clinical
351 validation dataset. Cells which were classified as 'parasite' from case 25 demonstrated pixel-wise
352 activation patterns which suggest that the model prediction of the target class was susceptible to
353 influence by features which share similarities to ring-form parasites. Examples of these microscopic
354 features which were associated with localized pixel activation included variations in erythrocyte
355 morphology (e.g., target cell contours) and overlying precipitate or platelets (Figure 8).

356 In general, model misclassification errors may be remedied by increasing the number of class
357 examples during training. In doing so, the model input space would be more representative of the
358 heterogeneity the model may be expected to encounter with real-world data, relative to a model
359 trained with fewer class examples. However, in the context of training classification models in
360 healthcare, particularly those which rely on cases of low prevalence diseases, increasing the number of
361 training examples can be prohibitive. There are techniques which can be implemented to artificially
362 expand the size of the training dataset (e.g., label-preserving image transformations) and improve
363 model performance and generalizability. However, these techniques are limited in terms of their
364 performance benefits and cannot portray inherent intra-class variability which is not already

365 represented in the existing training dataset. Overall, results of this study reinforce the need for
366 consistent, artifact-free, high quality data for optimal algorithm performance.

367 Most scientific literature on parasite quantitation is done in the context of Malaria diagnostics,
368 whereas approaches leveraging deep learning methods have only recently been described (22). To our
369 knowledge, this is the first published work to focus on the quantitation of *Babesia* with interpretable
370 clinical results, using images that are derived from routine clinical workflows. Further, we also evaluated
371 the utility of the model-based method using external validation datasets, not commonly done in malaria
372 quantitation studies (18,23). Similar to other published reports, we classified and quantified intracellular
373 parasites using ‘per-cell’ images (24). Other articles have also described a region-based approach,
374 wherein images containing multiples cells are evaluated for intracellular parasites, and a final
375 quantitative score is ultimately produced (18). However, while there are arguably benefits to each, there
376 is currently no clear advantage to either approach. Indeed, with the increasing breadth of machine
377 learning technologies, there are multiple avenues to pursue for parasite quantitation. Further research
378 is needed to delineate which methods are most performant, scalable, and most easily implemented into
379 clinical workflows, as well as addressing data quality for machine learning implementation in
380 microscopic image-based computer analysis.

381

382 **ACKNOWLEDGEMENTS:**

383 We would like to acknowledge Lisa Mehlin, Holly Base, and Laura Pires for volunteering their time to
384 quantitate *Babesia* parasites for the purposes of the clinical validation portion of this study. We would
385 also like to acknowledge John Errico and Cai Mayberry for their administrative support of this work.

386

387

388

389 **Bibliography**

- 390 1. Miller JM, Binnicker MJ, Campbell S, Carroll KC, Chapin KC, Gilligan PH, et al. A guide to utilization
391 of the microbiology laboratory for diagnosis of infectious diseases: 2018 update by the infectious
392 diseases society of america and the american society for microbiology. *Clin Infect Dis*. 2018;67:e1–
393 e94.
- 394 2. Padmanabhan A, Connelly-Smith L, Aqui N, Balogun RA, Klingel R, Meyer E, et al. Guidelines on the
395 Use of Therapeutic Apheresis in Clinical Practice - Evidence-Based Approach from the Writing
396 Committee of the American Society for Apheresis: The Eighth Special Issue. *J Clin Apher*.
397 2019;34:171–354.
- 398 3. Wormser GP, Dattwyler RJ, Shapiro ED, Halperin JJ, Steere AC, Klemperer MS, et al. The clinical
399 assessment, treatment, and prevention of lyme disease, human granulocytic anaplasmosis, and
400 babesiosis: clinical practice guidelines by the Infectious Diseases Society of America. *Clin Infect*
401 *Dis*. 2006;43:1089–1134.
- 402 4. Pierre RV. Peripheral blood film review. The demise of the eyecount leukocyte differential. *Clin*
403 *Lab Med*. 2002;22:279–297.
- 404 5. Rümke CL. Imprecision of ratio-derived differential leukocyte counts. *Blood Cells*. 1985;11:, 315.
- 405 6. Florin L, Maelegheer K, Muyldermans A, Van Esbroeck M, Nulens E, Emmerechts J. Evaluation of
406 the CellaVision DM96 advanced RBC application for screening and follow-up of malaria infection.
407 *Diagn Microbiol Infect Dis*. 2018;90:253–256.
- 408 7. Garcia E, Kundu I, Kelly M, Soles R. The american society for clinical pathology’s 2018 vacancy
409 survey of medical laboratories in the united states. *Am J Clin Pathol*. 2019;152:155–168.
- 410 8. Garcia E, Kundu I, Ali A, Soles R. The American Society for Clinical Pathology’s 2016-2017 Vacancy
411 Survey of Medical Laboratories in the United States. *Am J Clin Pathol*. 2018;149:387–400.
- 412 9. McPadden J, Warner F, Young HP, Hurley NC, Pulk RA, Singh A, et al. Clinical Characteristics and
413 Outcomes for 7,995 Patients with SARS-CoV-2 Infection. *medRxiv*. 2020;
- 414 10. Durant TJS, Gong G, Price N, Schulz WL. Bridging the Collaboration Gap: Real-time Identification of
415 Clinical Specimens for Biomedical Research. *J Pathol Inform*. 2020;11:14.
- 416 11. McPadden J, Durant TJ, Bunch DR, Coppi A, Price N, Rodgerson K, et al. Health care and precision
417 medicine research: analysis of a scalable data science platform. *J Med Internet Res*.
418 2019;21:e13043.
- 419 12. Keras Applications [Internet]. [cited 2021 Feb 15]. Available from:
420 <https://keras.io/api/applications/>
- 421 13. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. *arXiv*. 2017;
- 422 14. Bacus JW, Belanger MG, Aggarwal RK, Trobaugh FE. Image processing for automated erythrocyte
423 classification. *Journal of Histochemistry & Cytochemistry*. 1976;24:195–201.
- 424 15. Prewitt JM, Mendelsohn ML. The analysis of cell images. *Ann N Y Acad Sci*. 1966;128:1035–1053.

- 425 16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural
426 networks. *Commun ACM*. 2012;60:84–90.
- 427 17. Durant TJS, Olson EM, Schulz WL, Torres R. Very deep convolutional neural networks for
428 morphologic classification of erythrocytes. *Clin Chem*. 2017;63:1847–1855.
- 429 18. Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G. Image analysis and machine learning for
430 detecting malaria. *Transl Res*. 2018;194:36–55.
- 431 19. Yamamoto T, Tabe Y, Ishii K, Itoh S, Maeno I, Matsumoto K, et al. [Performance evaluation of the
432 CellaVision DM96 system in WBC differentials]. *Rinsho Byori*. 2010;58:884–890.
- 433 20. Kratz A, Bengtsson H-I, Casey JE, Keefe JM, Beatrice GH, Grzybek DY, et al. Performance evaluation
434 of the CellaVision DM96 system: WBC differentials by automated digital image analysis supported
435 by an artificial neural network. *Am J Clin Pathol*. 2005;124:770–781.
- 436 21. Integrated gradients | TensorFlow Core [Internet]. [cited 2020 Nov 12]. Available from:
437 https://www.tensorflow.org/tutorials/interpretability/integrated_gradients
- 438 22. Liang Z, Powell A, Ersoy I, Poostchi M, Silamut K, Palaniappan K, et al. CNN-based image analysis
439 for malaria diagnosis. 2016 IEEE International Conference on Bioinformatics and Biomedicine
440 (BIBM). IEEE; 2016. page 493–496.
- 441 23. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial
442 intelligence technology for medical diagnosis and prediction. *Radiology*. 2018;286:800–809.
- 443 24. Li S, Yang Q, Jiang H, Cortés-Vecino JA, Zhang Y. Parasitologist-level classification of apicomplexan
444 parasites and host cell with deep cycle transfer learning (DCTL). *Bioinformatics*. 2020;36:4498–
445 4505.

446

447

448

449 **FIGURES:**

450 **Figure 1:** Flow diagram of model development process. (A) Slides included in the model development
451 dataset were imaged a single time by the Cellavision DI-60 and uploaded to a custom-built-web
452 application for label annotation. (B) Central X-Y coordinates of infected (red) and non-infected (blue)
453 erythrocytes were marked on the slide-level images. (C) Central X-Y coordinates were used to crop
454 individual erythrocytes into 70 x 70 pixel, 3-channel arrays and paired with the corresponding label of
455 either 'parasite' (red) or 'normal' (blue). (D) Labeled erythrocyte images were collectively divided 80:20
456 into train and test datasets, respectively. The train dataset was further subdivided 70:30 into train and
457 validation datasets, respectively. (E) The train and validation datasets were used to train the image
458 classification model. (F) Following completion of training, the [best model] was used to evaluate model
459 performance using the test dataset.

460 **Figure 2:** Flow diagram of clinical validation process. (A) Each peripheral blood smear was evaluated
461 three times, in a blinded fashion, by each MLS. (B) This process yielded a total of 9 parasitemia results
462 for each slide in the clinical validation dataset. These data were used to calculate the average
463 parasitemia across all 9 reads which was used as the clinical reference standard for each case. (C) Each
464 glass slide in the clinical validation dataset was imaged three separate times by the Cellavision DI-60. (D)
465 Contour-based cell segmentation was used to extract individual erythrocytes from the DI-60 slide-level
466 images as 70x70x3 cropped images. (E) Individually cropped erythrocytes were independently evaluated
467 by the [best model] to yield a predicted class (i.e., 'parasite' or 'normal'). (F) The number of cells with
468 the predicted label of 'parasite' were divided by number of total cells classified to yield the parasitemia
469 result. This process was done one time for each DI-60 image. With three images per specimen, this
470 yielded a total of 3 parasitemia results per slide, which were used to calculate an average parasitemia
471 result for each specimen.

472 **Figure 3:** Model performance metrics plotted as a function of training epochs (iterations). (A) Train and
473 validation loss. (B) Train and validation recall (sensitivity). (C) Train and validation area under the
474 receiver operator characteristic curve. (D) Train and validation precision (positive predictive value).

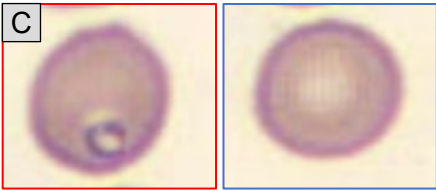
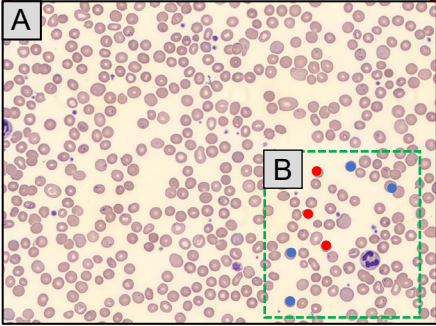
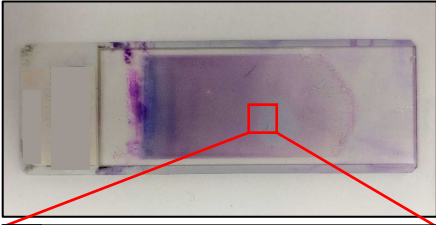
475 **Figure 4:** Model classification results on test dataset. (A) Confusion matrix of actual versus predicted
476 labels. (B) Per-cell probability distribution of model predicted class with actual labels depicted in color
477 (red = parasite) (blue = normal). X-axis: The probability of the predicted class being 'parasite'. Y-axis:
478 Random number between 0 and 1 was assigned to each cell for better visualizing data points. Green
479 dotted line: Decision threshold for prediction label of 'parasite' – i.e., cells with a predicted probability
480 of ≥ 0.5 are labeled as 'parasite'.

481 **Figure 5:** Bar plot of mean percent parasitemia for the MLS-RS (n=9) and the model-based method (n=3).
482 Error bars represent 1 standard deviation.

483 **Figure 6:** Visualizations for method-to-method comparison of MLS-RS and model-based method. (A) XY-
484 scatter plot with regression line overlay (red-dotted line represents 95% confidence interval of
485 regression). (B) Bland-Altman absolute bias plot. (C) Bland-Altman percent bias plot.

486 **Figure 7:** Integrated gradient (IG) visualizations including the original image, the pixel-wise IG
487 attribution mask, and the overlay of the two. Images are from the model development test dataset. (A
488 and B) Representative examples from the 'parasite' class. (C and D) Representative examples from the
489 'normal' class.

490 **Figure 8:** Integrated gradient (IG) visualizations including the original image and an overlay of the pixel-
491 wise IG attribution mask and the original image. Images are from Case #25 of the clinical validation
492 dataset and are those which were predicted as belonging to the 'parasite' class.

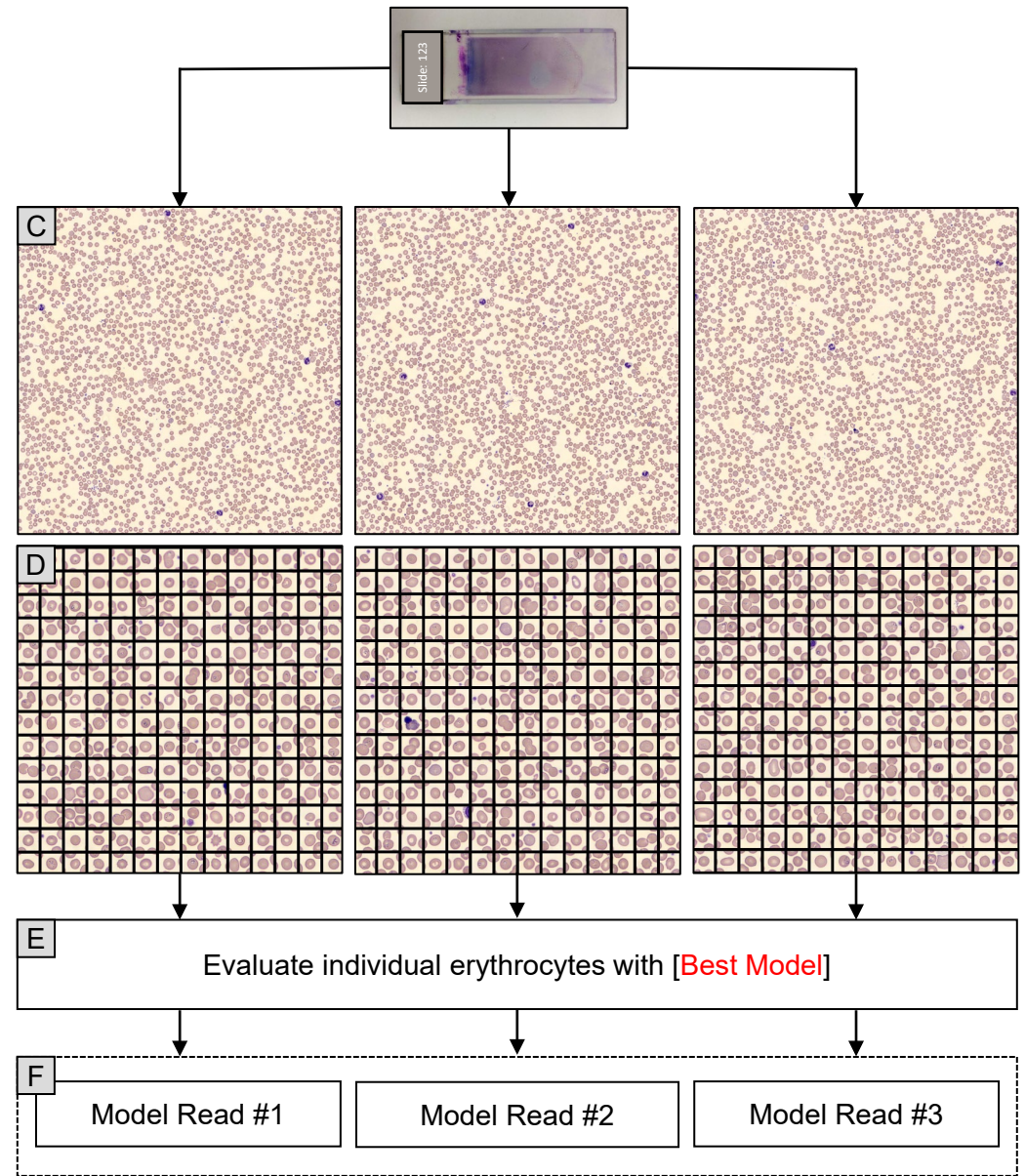
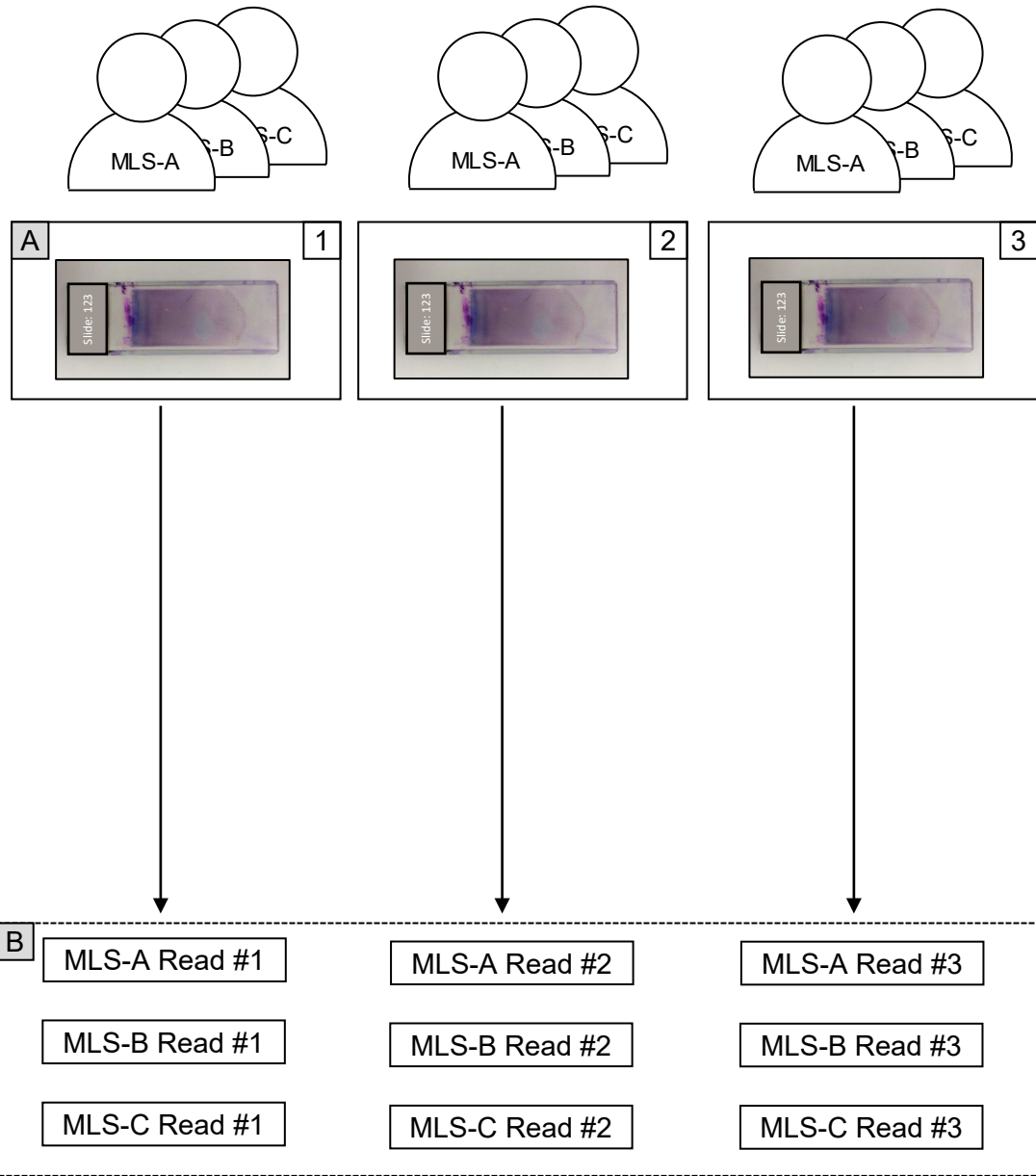


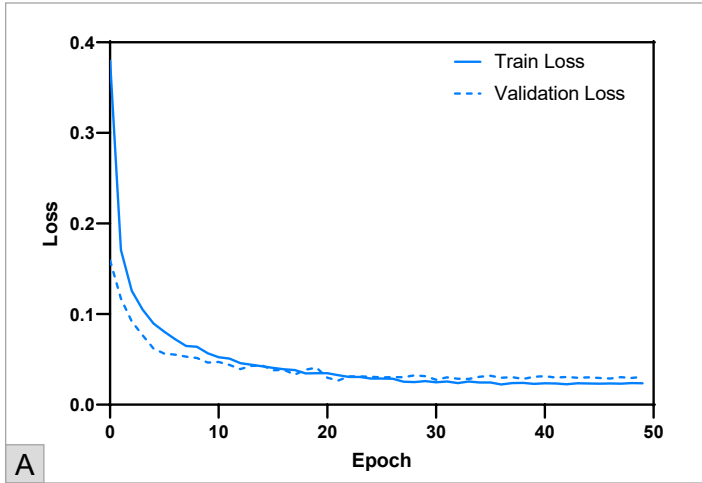
D Train Dataset Test Dataset

Train Dataset Validation Dataset

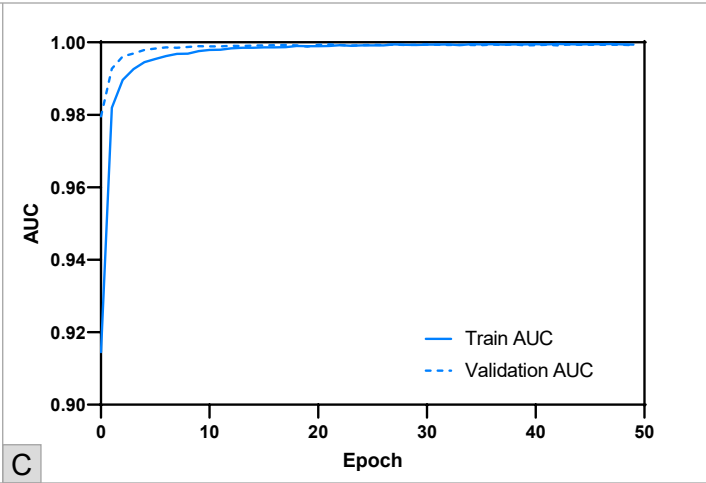
E Train Model
Save [Best Model]

F Evaluate
Test Dataset
[Best Model]

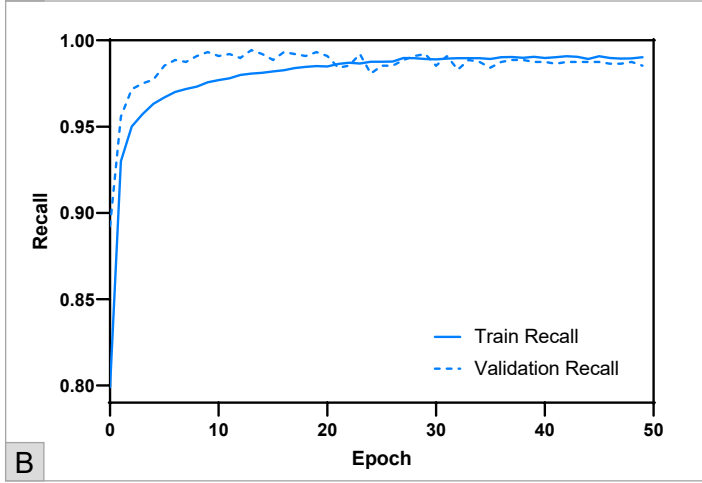




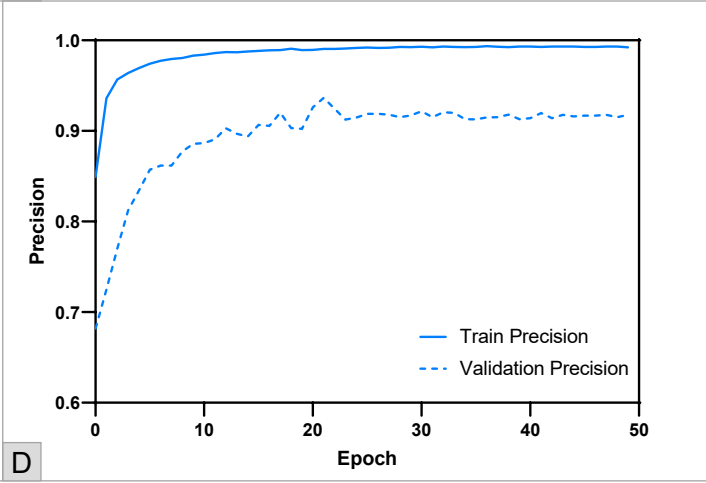
A



C



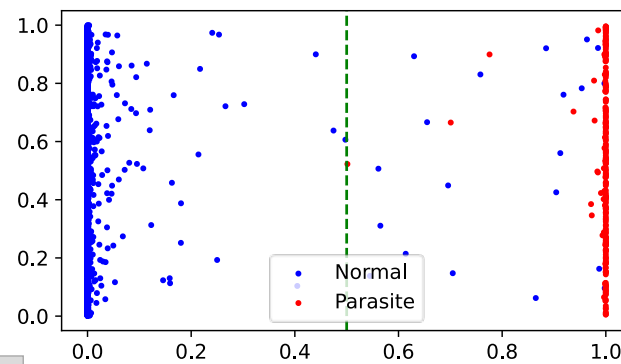
B



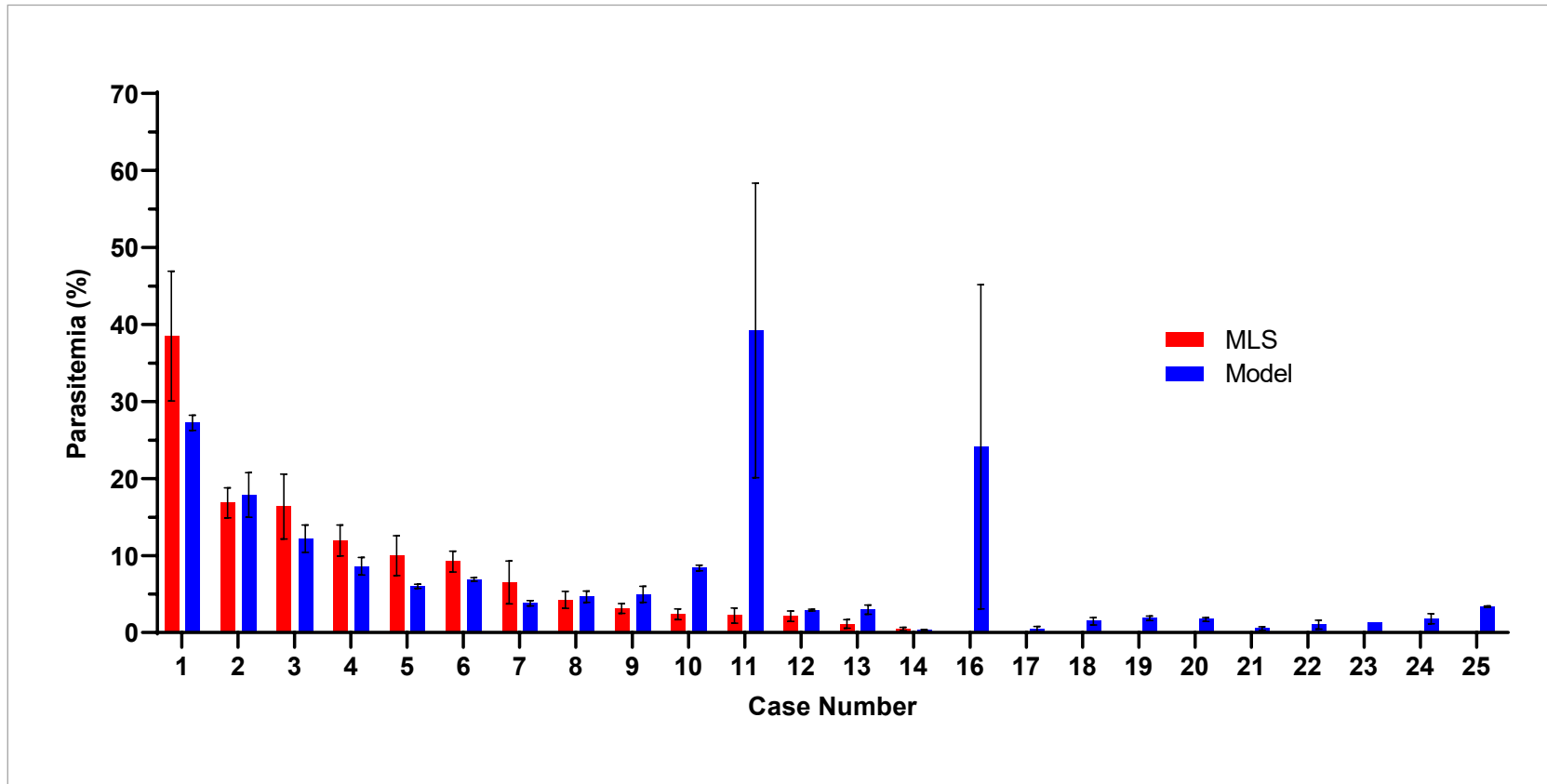
D

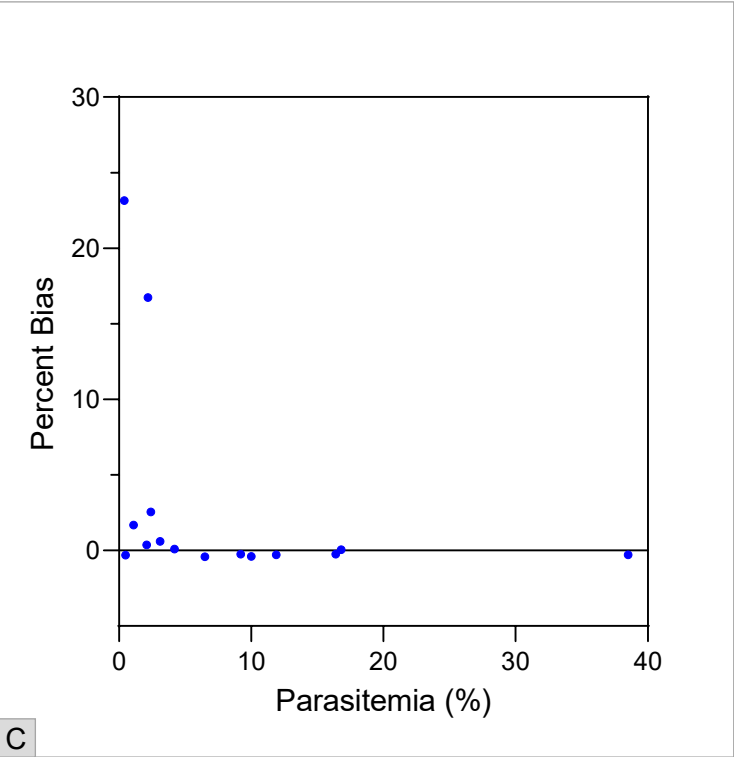
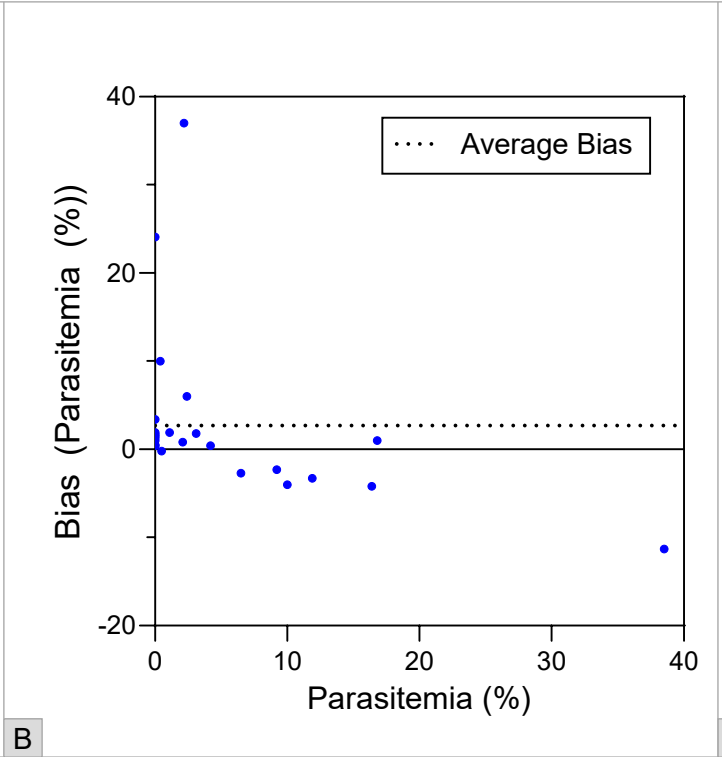
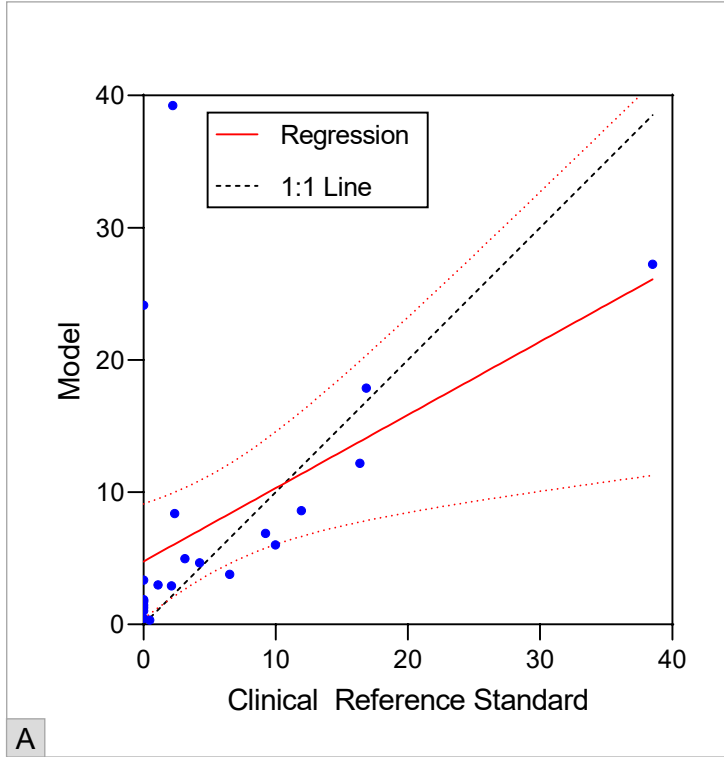
Actual Label	Predicted Label	
	Normal	Parasite
Normal	2258	20
Parasite	0	245

A



B

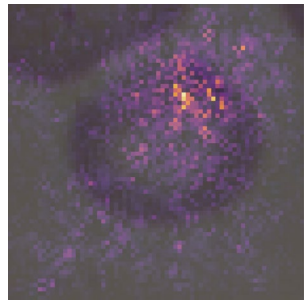
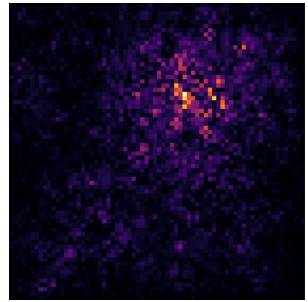
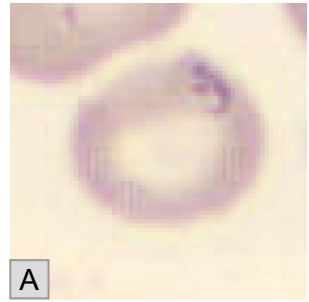




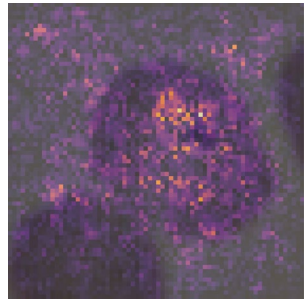
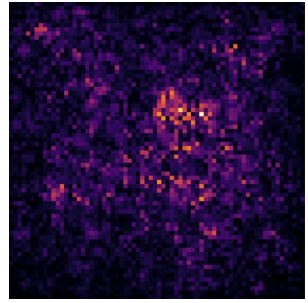
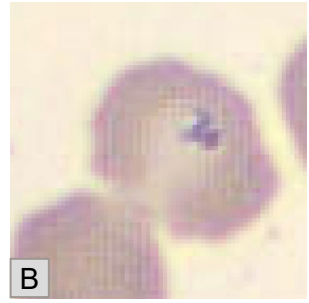
Original Image

Attribution Mask

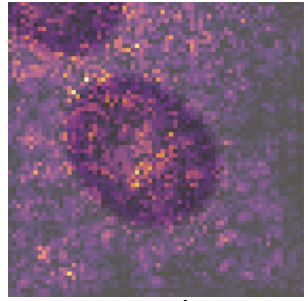
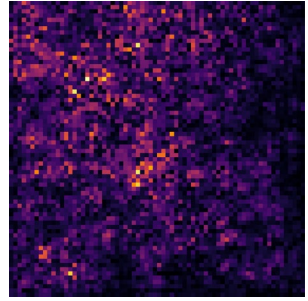
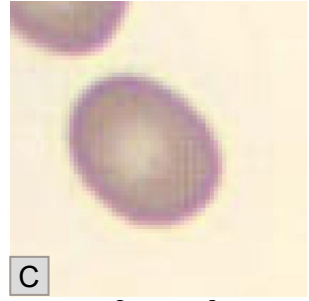
Overlay



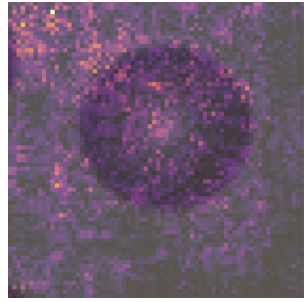
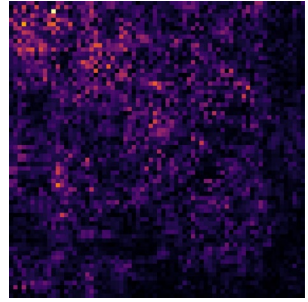
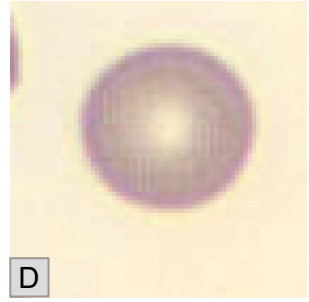
A



B



C



D

