Pre-trained convolutional neural networks as feature extractors toward improved Malaria parasite detection in thin blood smear images

Sivaramakrishnan Rajaraman[1], Sameer K Antani[1], Mahdieh Poostchi[1], Kamolrat Silamut[2], Md. A Hossain[3], Richard J Maude[2], Stefan Jaeger[1], George R Thoma[1]

[1] Communications Engineering Branch, National Library of Medicine, Bethesda, MD 20894, USA

[2] Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, Bangkok 10400, Thailand

[3] Department of Medicine, Chittagong Medical Hospital, Chittagong 4203, Bangladesh

Corresponding Author:

Sivaramakrishnan Rajaraman[1]

Email address: sivaramakrishnan.rajaraman@nih.gov

32 **Pre-trained convolutional neural networks as feature extractors toward**

33 **improved Malaria parasite detection in thin blood smear images**

34 Sivaramakrishnan Rajaraman[1], Sameer K Antani[1], Mahdieh Poostchi[1], Kamolrat Silamut[2], Md. A

35 Hossain[3], Richard J Maude[2], Stefan Jaeger[1], George R Thoma[1]

36 [1] Communications Engineering Branch, National Library of Medicine, Bethesda, MD 20894, USA

37 [2] Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, Bangkok 10400,

38 Thailand

39 [3] Department of Medicine, Chittagong Medical Hospital, Chittagong 4203, Bangladesh

40

Abstract

41

42 Malaria is a blood disease caused by the *Plasmodium* parasites transmitted through the bite of

43 female Anopheles mosquito. Microscopists commonly examine thick and thin blood smears to

44 diagnose disease and compute parasitemia. However, their accuracy depends on smear quality and

45 expertise in classifying and counting parasitized and uninfected cells. Such an examination could

46 be arduous for large-scale diagnoses resulting in poor quality. State-of-the-art image-analysis

47 based computer-aided diagnosis (CADx) methods using machine learning (ML) techniques,

48 applied to microscopic images of the smears using hand-engineered features demand expertise in

49 analyzing morphological, textural, and positional variations of the region of interest (ROI). In

50 contrast, Convolutional Neural Networks (CNN), a class of deep learning (DL) models promise

51 highly scalable and superior results with end-to-end feature extraction and classification.

52 Automated malaria screening using DL techniques could, therefore, serve as an effective

53 diagnostic aid. In this study, we evaluate the performance of pre-trained CNN based DL models

54 as feature extractors toward classifying parasitized and uninfected cells to aid in improved disease

55 screening. We experimentally determine the optimal model layers for feature extraction from the

56 underlying data. Statistical validation of the results demonstrates the use of pre-trained CNNs as a

57 promising tool for feature extraction for this purpose.

58

Introduction

59

60 Malaria is a mosquito-borne blood disease caused by the *Plasmodium* parasites transmitted

61 through the bite of the female Anopheles mosquito. Different kinds of parasites including *P. ovale,*

62 *P. malariae, P. vivax and P. falciparum* infect the humans, however, the effects of *P. falciparum*

63 can be lethal. In 2016, World Health Organization (WHO) reported 212 million instances of the

64 disease across the world (WHO, 2016). Microscopic thick and thin blood smear examinations are

65 the most reliable and commonly used method for disease diagnosis. Thick blood smears assist in

66 detecting the presence of parasites while thin blood smears assist in identifying the species of the

67 parasite causing the infection (Centers for Disease Control and Prevention, 2012). The diagnostic

68 accuracy heavily depends on the human expertise and can be adversely impacted by the inter-

69 observer variability and the liability imposed by large-scale diagnoses in disease-

70 endemic/resource-constrained regions (Mitiku, Mengistu & Gelaw, 2003). Alternative techniques

71 such as polymerase chain reaction (PCR) and rapid diagnostic tests (RDT) are used, however, PCR

72 analysis is limited in its performance (Hommelsheim et al., 2015) and RDTs are less cost-effective
73 in disease-endemic regions (Hawkes, Katsuva & Masumbuko, 2009).
74 In the process of applying machine learning (ML) methods to medical data analysis, meaningful
75 feature representation lies at the core of their success to accomplish desired results. A majority of
76 image analysis-based computer-aided diagnosis (CADx) software use ML techniques with hand-
77 engineered features for decision-making (Ross et al., 2006; Das et al., 2013; Poostchi et al., 2018).
78 However, the process demands expertise in analyzing the variability in size, background, angle,
79 and position of the region of interest (ROI) on the images. To overcome challenges of devising
80 hand-engineered features that capture variations in the underlying data, Deep Learning (DL), also
81 known as deep hierarchical learning, is used with significant success (LeCun, Yoshua & Geoffrey,
82 2015). DL models use a cascade of layers of non-linear processing units to self-discover
83 hierarchical feature representations in the raw data. Higher-level features are abstracted from
84 lower-level features to aid in learning complex, non-linear decision-making functions, resulting in
85 end-to-end feature extraction and classification (Schmidhuber, 2015). Unlike kernel-based
86 algorithms like Support Vector Machines (SVMs), DL models exhibit improved performance with
87 an increase in data size and computational resources, making them highly scalable (Srivastava et
88 al., 2014).
89 For images, an important source of information lies in the spatial local correlation among the
90 neighboring pixels/voxels. Convolutional Neural Networks (CNN), a class of DL models are
91 designed to exploit this information through the mechanisms of local receptive fields, shared
92 weights and pooling (Krizhevsky, Sutskever & Hinton, 2012). In 2012, Alex Krizhevsky proposed
93 AlexNet, a CNN based DL model that won the ImageNet Large Scale Visual Recognition
94 Challenge (ILSVRC) and substantially boosted the performance of CNNs toward classifying
95 natural images (Krizhevsky, Sutskever & Hinton, 2012). Several representative CNNs like
96 VGGNet (Simonyan & Zisserman, 2015), GoogLeNet (Szegedy et al., 2014), and ResNet (He et
97 al., 2016) demonstrated significant improvements in succeeding ILSVRC annual challenges. A
98 model named Xception was proposed that uses depth-wise separable convolutions (Chollet, 2016)
99 to outperform the Inception-V3 model (Szegedy et al., 2016) on the ImageNet (Deng et al., 2009)
100 data classification task. A CNN variant called Densely Connected Convolutional Networks
101 (DenseNet) was proposed (Huang et al., 2016) that utilizes a network architecture in which each
102 layer is directly connected to every later layer. The model has achieved noteworthy improvements
103 over the state-of-the-art while using significantly fewer parameters and computations.
104 The promising performance of CNNs is accompanied by the availability of a huge amount of
105 annotated data. With scarcity for annotated medical imagery, Transfer Learning (TL) methods are
106 used where pre-trained DL models are either fine-tuned on the underlying data or used as feature
107 extractors to aid in visual recognition tasks (Razavian et al., 2014). These models transfer their
108 knowledge gained while learning the generic features from large-scale datasets like ImageNet to
109 the underlying task. The transfer of previously-learned skills to a new situation is generalized,
110 rather than unique to the situation. Since the results published in (Razavian et al., 2014), it is
111 recognized that CNNs trained on large-scale datasets could serve as feature extractors for a wide

112   range of computer vision tasks to aid in improved performance, as compared to state-of-the-art
113   methods (Bousetouane & Morris, 2015).
114   At the present, researchers across the world have begun to apply DL tools and obtain promising
115   results in a wide variety of medical image analyses/understanding tasks (Rajaraman et al., 2017;
116   Suzuki, 2017). Literature also reveals studies pertaining to applying DL methods to the task of
117   malaria parasite detection. Dong *et al.* compared the performance of SVM and pre-trained DL
118   models including LeNet (LeCun et al., 1998), AlexNet, and GoogLeNet toward classifying
119   parasitized and uninfected cells (Dong et al., 2017). Red Blood cells (RBCs) were segmented from
120   thin blood smear images and randomly split into train and test sets. 25% of the training images
121   were randomly selected to validate the models. Liang *et al.* proposed a 16-layer CNN toward
122   classifying the uninfected and parasitized cells. Features were extracted using the pre-trained
123   AlexNet and an SVM classifier was trained on the extracted features (Liang et al., 2017). The
124   performance of the proposed model was compared to that of the pre-trained CNN. The study
125   reported that the custom model was more accurate, sensitive and specific than the pre-trained
126   model. Images were resampled to 44×44 pixel resolution to compensate for the lack of
127   computational resources that would have led to the loss of image resolution. Bibin *et al.* proposed
128   a 6-layer deep belief network toward malaria parasite detection in peripheral blood smear images
129   (Bibin, Nair & Punitha, 2017). The authors reported 96.4% accuracy in the task of classifying a
130   dataset of 4100 cells with randomized train/test splits. Gopakumar *et al.* employed a customized
131   CNN model for analyzing videos containing a focus stack of the field of views of Leishman stained
132   slide images toward the process of automated parasite detection (Gopakumar et al., 2017). The
133   authors used a customized portable slide scanner and off-the shelf components for data acquisition
134   and demonstrated sensitivity and specificity of 97.06% and 98.50% respectively. In summary,
135   existing DL studies have been evaluated on relatively small image sets and/or randomized train/test
136   splits. None of the studies have reported the performance of the predictive models at the patient
137   level. Although the reported outcomes are promising, existing approaches need to substantiate
138   their robustness on a larger set of images with cross-validation studies at the patient level.
139   Evaluation on patient-level provides a more realistic performance evaluation of the predictive
140   models as the images in the independent test set represent truly unseen images for the training
141   process, with no information about staining variations or other artifacts leaking into the training
142   data. This would help to reduce bias and generalization errors. Tests for statistically significant
143   differences in performance would further assist in the process of optimal model selection prior to
144   deployment. It is reasonable to mention that the state-of-the-art still leaves much room for progress
145   in this regard.
146   In this work, we evaluate the performance of pre-trained CNN based DL models as feature
147   extractors toward classifying the parasitized and uninfected cells to aid in improved disease
148   screening. The important contributions of this work are as follows: (a) presentation of a
149   comparative analysis of the performance of customized and pre-trained DL models as feature
150   extractors toward classifying parasitized and uninfected cells, (b) cross-validating the performance
151   of the predictive models at the patient level to reduce bias and generalization errors, (c) analysis

152 and selection of the optimal layer in the pre-trained models to extract features from the underlying
153 data, and (d) testing for the presence/absence of a statistically significant difference in the
154 performance of customized and pre-trained CNN models under study. The following paper is
155 organized as follows: Section 2 elaborates on the materials and methods, Section 3 presents the
156 results, and Section 4 discusses the results and concludes the paper.
157

158 **Materials and Methods**

159 **Data collection**
160 To reduce the burden for microscopists in resource-constrained regions and improve diagnostic
161 accuracy, researchers at the Lister Hill National Center for Biomedical Communications
162 (LHNCBC), part of National Library of Medicine (NLM) have developed a mobile application
163 that runs on a standard Android® smartphone attached to a conventional light microscope (Poostchi
164 et al., 2018). Giemsa-stained thin blood smear slides from 150 *P. falciparum*-infected and 50
165 healthy patients were collected and photographed at Chittagong Medical College Hospital,
166 Bangladesh. The smartphone's built-in camera acquired images of slides for each microscopic
167 field of view. The images were manually annotated by an expert slide reader at the Mahidol-Oxford
168 Tropical Medicine Research Unit in Bangkok, Thailand. The de-identified images and annotations
169 are archived at NLM (IRB#12972). We applied a level-set based algorithm to detect and segment
170 the red blood cells (Poostchi et al., 2018).
171

172 **Cross-validation studies**
173 The dataset consists of 27,558 cell images with equal instances of parasitized and uninfected cells.
174 Positive samples contained the *Plasmodium* and negative samples contained no *Plasmodium* but
175 other types of objects including staining artifacts/impurities. We evaluated the predictive models
176 through five-fold cross-validation. Cross-validation has been performed at the patient level to
177 ensure alleviating model biasing and generalization errors. The count of cells for the different folds
178 is shown in Table 1.
179

180 **Table 1:**
181 **Data for cross-validation studies.**
182

183 The images were re-sampled to 100×100, 224×224, 227×227 and 299×299 pixel resolutions to
184 suit the input requirements of customized and pre-trained CNNs and normalized to assist in faster
185 convergence. The models were trained and tested on a Windows® system with Intel® Xeon® CPU
186 E5-2640v3 2.60-GHz processor, 1 TB HDD, 16 GB RAM, a CUDA-enabled Nvidia® GTX 1080
187 Ti 11GB graphical processing unit (GPU), Matlab® R2017b, Python® 3.6.3, Keras® 2.1.1 with
188 Tensorflow® 1.4.0 backend, and CUDA 8.0/cuDNN 5.1 dependencies for GPU acceleration.
189

190 **Customized model configuration**

191 We also evaluated the performance of a customized, sequential CNN in the task of classifying
192 parasitized and uninfected cells toward disease screening. We propose a sequential CNN as shown
193 in Fig. 1, similar to the architecture that LeCun *et al.* advocated for image classification (LeCun &
194 Bengio, 1995).
195
196 **Figure 1: Architecture of the customized model.**
197
198 The proposed CNN has three convolutional layers and two fully connected layers. The input to the
199 model constitutes segmented cells of 100×100×3 pixel resolution. The convolutional layers use
200 3×3 filters with 2 pixel strides. The first and second convolutional layers have 32 filters and the
201 third convolutional layer has 64 filters. The sandwich design of convolutional/rectified linear units
202 (ReLU) and proper weight initialization enhances the learning process (Shang et al., 2016). Max-
203 pooling layers with a pooling window of 2×2 and 2 pixel strides follow the convolutional layers
204 for summarizing the outputs of neighboring neuronal groups in the feature maps. The pooled
205 output of the third convolutional layer is fed to the first fully-connected layer that has 64 neurons,
206 and the second fully connected layer feeds into the Softmax classifier. Dropout regularization
207 (Srivastava et al., 2014) with a dropout ratio of 0.5 is applied to outputs of the first fully connected
208 layer. The model is trained by optimizing the multinomial logistic regression objective using
209 stochastic gradient descent (SGD) (LeCun, Bengio & Hinton, 2015) and Nesterov's momentum
210 (Botev, Lever & Barber, 2017). The customized model is optimized for hyper-parameters by a
211 randomized grid search method (Bergstra & Bengio, 2012). We initialized search ranges to be [1e-
212 7 5e-2], [0.8 0.99] and [1e-10 1e-2] for the learning rate, stochastic gradient descent (SGD) and
213 L2-regularization parameters, respectively. We evaluated the performance of the customized
214 model in terms of accuracy, AUC, sensitivity, specificity, F1-score (Lipton, Elkan &
215 Naryanaswamy, 2014) and Matthews correlation coefficient (MCC) (Matthews, 1975).
216
217 **Feature extraction using pre-trained models**
218 We evaluated the performance of pre-trained CNNs including AlexNet (winner of ILSVRC 2012),
219 VGG-16 (winner of ILSVRC's localization task in 2014), Xception, ResNet-50 (winner of
220 ILSVRC 2015) and DenseNet-121 (winner of the best paper award in CVPR 2017) toward
221 extracting the features from the parasitized and uninfected cells. The models were optimized for
222 hyper-parameters by the randomized grid search method. We initialized search ranges to be [1e-5
223 5e-2], [0.8 0.99] and [1e-10 1e-2] for the learning rate, Nesterov's accelerated stochastic gradient
224 descent (SGD) and L2-regularization parameters, respectively. We instantiated the convolutional
225 part of the pre-trained CNNs and trained a fully-connected model with dropout (dropout ratio of
226 0.5) on top of the extracted features. We also empirically determined the optimal layer for feature
227 extraction to aid in improved classification. We evaluated the performance of the pre-trained
228 CNNs in terms of accuracy, AUC, sensitivity, specificity, F1-score, and MCC. The model
229 architecture and weights for the pre-trained CNNs were downloaded from GitHub repositories
230 (Chollet; Yu, 2016).

231
232 **Statistical analysis**
233 We performed statistical analyses to choose the best model for deployment. Statistical methods
234 like one-way analysis of variance (ANOVA) are used to determine the presence or absence of a
235 statistically significant difference between the means of three or more individual, unrelated groups
236 (Rossi, 1987). One-way ANOVA tests the null hypothesis (H0) given by H0: $\mu_1 = \mu_2 = \ldots = \mu_k$
237 where μ = mean of parameters for the individual groups and k = total number of groups. If a
238 statistically significant result is returned by the test, H0 is rejected and the alternative hypothesis
239 (H1) is accepted to infer that a statistically significant difference exists between the means of at
240 least two groups under study. However, it would be appropriate to use this parametric test only
241 when the underlying data satisfies the assumptions of independence of observations, absence of
242 significant outliers, normality of data and homogeneity of variances (Daya, 2003). When the
243 conditions are violated, a non-parametric alternative like Kruskal-Wallis H test (also called the
244 one-way ANOVA on ranks) could be used (Vargha, Delaney & Vargha, 1998). This is an omnibus
245 test that couldn't identify the specific groups that demonstrate statistically significant differences
246 in their mean values. A post-hoc analysis is needed to identify these groups that demonstrate
247 statistically significant differences (Kucuk et al., 2016). We performed Shapiro-Wilk test
248 (Royston, 1992) to check for data normality and Levene's statistic test (Gastwirth, Gel & Miao,
249 2009) to study the homogeneity of variances for the performance metrics for the different models
250 under study. Statistical analyses were performed using IBM® SPSS® statistical package (IBM
251 Corp. Released, 2015).
252
253 **Results**
254 **Cell segmentation and detection**
255 We applied a level-set based algorithm to detect and segment the red blood cells as shown in Fig.
256 2. The first step is the cell detection where we applied a multi-scale Laplacian of Gaussian (LoG)
257 filter to detect centroids of individual RBCs.
258
259 **Figure 2: RBC detection and segmentation using iterative voting and level sets.**
260 (A) Input image. (B) Initial cell detection using iterative voting. (C) Final RBC segmentation mask.
261 (D) Segmentation results superimposed on the original image.
262
263 The generated markers are used to segment the cells within a level set active contour framework
264 to confine the evolving contour to the cell boundary. Morphology opening operation is applied as
265 post-processing to remove false detected objects such as staining artifacts using average cell size.
266 White blood cells (WBCs) are filtered out using a one-to-one correspondence based on cell
267 ground-truth annotations since WBCs are not ROIs for this work. We have evaluated our cell
268 detection based on the manual point-wise annotation of infected and uninfected cells. To do so,
269 we applied a one-to-one point matching scheme: For each segmented cell, we checked the number
270 of manual ground-truth points in the segmented cell region. If there is exactly one point in the

271   region, we counted this as a true positive (TP). If there is no point in the region, we counted this
272   as a false positive (FP). If there is more than one point in the region, we considered this as an
273   under-segmentation or false negative (FN). These counts then allowed us to compute the presented
274   values for positive predictive value (PPV), sensitivity and F1-score. For cell detection, we obtained
275   a PPV of 0.944, sensitivity of 0.962 and F1-score of 0.952.

276

277   **Performance metrics evaluation**
278   For the customized and pre-trained models, we empirically determined the optimum value to be
279   0.9 and 1e-6 for the SGD momentum and L2-regularization, respectively. For the learning rate, we
280   determined the optimum value to be 1e-5 and 1e-6 for the customized and pre-trained CNNs
281   respectively. The second fully connected layer from AlexNet, VGG-16 and the last layer before
282   the final classification layer from Xception, ResNet-50, and DenseNet-121 were selected for
283   feature extraction. Table 2 lists the performance metrics achieved by the models in the process of
284   classifying parasitized and uninfected cells. We also evaluated the performance of pre-trained
285   CNNs by extracting features from different layers in the process of identifying the optimal layer
286   for feature extraction from the underlying data. The naming conventions for these layers are based
287   on the models obtained from Keras® DL library.

288

289   **Table 2:**
290   **Performance metrics.**

291

292   Layers that gave the best values for the performance metrics are listed in Table 3. Table 4 shows
293   the results obtained by extracting the features from the optimal layers toward classifying the
294   parasitized and uninfected cells.

295

296   **Table 3:**
297   **Candidate layers giving the best performance.**

298

299   **Table 4:**
300   **Performance metrics achieved with feature extraction from optimal layers.**

301

302   While performing statistical analyses, we observed that the results of Shapiro-Wilk test were
303   statistically significant for all the performance metrics ($p < 0.05$) to signify that the normality of
304   data has been violated. For this reason, we opted to use the non-parametric Kruskal-Wallis H test.
305   The consolidated results of Kruskal-Wallis H and post-hoc analyses are given in Table 5. We
306   observed that, in terms of accuracy, there existed a statistically significant difference in
307   performance between the different CNNs (χ^2 (5) = 15.508, p = 0.008). Post-hoc tests further
308   revealed that the statistically significant difference existed between the pre-trained Xception,
309   VGG-16, ResNet-50, and customized model. In terms of AUC, a statistically significant difference
310   was observed (χ^2 (5) = 18.958, p = 0.002) in the performance between Xception, ResNet-50, VGG-

311 16, and DenseNet-121. Similar results were observed for the F1-score (χ^2 (5) = 14.798, p = 0.011)
312 and MCC (χ^2 (5) = 14.487, p = 0.013). No statistically significant difference was observed across
313 the models in terms of sensitivity (χ^2 (5) = 5.518, p = 0.356) and specificity (χ^2 (5) = 6.639, p =
314 0.249). However, ResNet-50 obtained the highest mean ranks for accuracy, specificity, F1-score,
315 and MCC.

316

317 **Table 5:**
318 **Consolidated results of Kruskal-Wallis H and post-hoc tests.**

319

320 **Discussions and conclusion**
321 The customized model converged to an optimal solution due to hyper-parameter optimization,
322 implicit regularization imposed by smaller convolutional filter sizes and aggressive dropouts in
323 the fully connected layers. Usage of L2 regularization reduced the effect of model overfitting and
324 converging to a better solution (Simonyan & Zisserman, 2015).
325 Each layer of the CNNs produces an activation for the given image. Earlier layers capture primitive
326 features like blobs, edges, and colors that are abstracted by the deeper layers to form higher level
327 features to present a more affluent image representation (Zeiler & Fergus, 2014). Studies from the
328 literature reveal that while using pre-trained CNNs for feature extraction, the features are extracted
329 from the layer, right before the classification layer (Razavian et al., 2014). For this reason, we
330 extracted the features from the second fully connected layer for AlexNet and VGG-16 and the last
331 layer before the final classification layer from Xception, ResNet-50, and DenseNet-121 models.
332 We observed from the patient-level cross-validation studies (Table 2) that ResNet-50
333 outperformed the customized and other pre-trained CNNs in all performance metrics toward the
334 task of classifying parasitized and uninfected cells. Literature studies reveal that DenseNet-121
335 outperformed ResNets and other pre-trained CNNs in the ImageNet data classification task (Huang
336 et al., 2016). In our case, for the binary task of classifying parasitized and uninfected cells, the
337 variability in data is several orders of magnitude smaller, the top layers of deep CNNs like
338 DenseNet-121 are probably too specialized, progressively more complex and not the best
339 candidate to re-use for the task of our interest. For this reason, we evaluated the performance of
340 pre-trained CNNs by extracting features from different layers in the process of identifying the
341 optimal layer for feature extraction from the underlying data (Table 3). In the process, we observed
342 that for the pre-trained CNNs, the performance of the layer before the classification layer was
343 degraded compared to the other layers. In contrast to the results shown in Table 2, DenseNet-121
344 achieved the best values for sensitivity but demonstrated similar AUC values as ResNet-50 and
345 VGG-16 (Table 4). Both VGG-16 and ResNet-50 were equally accurate and demonstrated equal
346 values for AUC and F1-score. However, ResNet-50 was highly specific, demonstrated high MCC
347 and performed relatively better than the other models under study. These results demonstrate that
348 the final layer of pre-trained CNNs is not always optimal for extracting the features from the
349 underlying data. In our study, features from shallow layers performed better than deep features to
350 aid in improved classification of parasitized and uninfected cells. Literature studies reveal that

351   MCC is an informative single score to evaluate the performance of a binary classifier in a confusion
352   matrix context (Chicco, 2017). In this regard, ResNet-50 demonstrated statistically significant
353   MCC metrics as compared to the other models. The consolidated results demonstrated that the pre-
354   trained ResNet-50 relatively outperformed the other models under study toward classifying the
355   parasitized and uninfected cells.
356   While performing Kruskal-Wallis H and post-hoc analyses, we observed that the pre-trained
357   ResNet-50 obtained the highest mean ranks for accuracy, specificity, F1-score, and MCC. If we
358   were looking to select a model based on a balance between PPV and sensitivity as demonstrated
359   by the F1-score, we could observe that the pre-trained ResNet-50 outperformed the other models
360   under study. We have demonstrated the performance of the models in terms of mean (μ) and
361   standard deviation (σ) to present a measure of the dispersion in the performance metrics. The pre-
362   trained ResNet-50 outperformed the other models by achieving 0.947±0.015 sensitivity and
363   0.972±0.10 specificity. Statistical analyses show that the predictive model could capture all
364   observations within three standard deviations from the mean [-3σ 3σ], i.e., the model could exhibit
365   sensitivity and specificity in the range [0.902 0.992] and [0.942 1.00] respectively. However, our
366   study is focused on disease screening, therefore, the sensitivity metric carries significance. We also
367   determined the number of RBCs be analyzed by the proposed model to confidentially return a
368   positive test result. We used the epiR tools for the Analysis of Epidemiological Data (Stevenson
369   et al., 2015) for these computations. The number of cells needed to diagnose (NND) is defined as
370   the number of RBCs to be tested to yield a correct positive test. Youden's index gives a measure
371   of the performance of the model, the value ranges from -1 to +1 with values closer to 1 for higher
372   values of sensitivity and specificity. With a confidence level (CI) of 0.95 ($p < 0.05$), we found that
373   11 RBCs need to be tested to return 10 positive results.
374   To our knowledge, we could find no comparable literature that performed cross-validation studies
375   at the patient level, with a large-scale clinical dataset for the underlying task. For this reason, we
376   also performed cross-validation studies at the cell level and compared with the state-of-the-art
377   (Table 6).
378
379   **Table 6:**
380   **Comparison with the state-of-the-art literature.**
381
382   In the process, we found that the pre-trained ResNet-50 outperformed the state-of-the-art in all
383   performance metrics. Das *et al.* achieved similar values for sensitivity with a small-scale dataset
384   but demonstrated sub-optimal specificity. The lack of performance at the patient level is attributed
385   to the staining variations between patients. We observed that it is harder for the classifier to learn
386   the different stains, which indicates that we may need to acquire more images with different
387   staining colors for training or apply color normalization techniques. However, by validating the
388   predictive models at the patient-level, which we believe simulate real-world conditions, we ensure
389   getting rid of bias, reduce overfitting and generalization errors toward optimal model deployment.

390   We are currently performing pilot studies in deploying the customized and pre-trained DL models
391   into mobile devices and analyzing the performances. From the literature studies, we observed that
392   we could either opt to both train/predict on the mobile device or to train the model offline and
393   import to the mobile device to predict on the independent test data (Howard et al., 2017). Currently,
394   Android and IOS ML libraries (like CoreMLStudio) offer the flexibility for dynamic allocation of
395   CPU and GPU based on the computational cost, thus, memory allocation probably doesn't seem
396   to an issue while deploying deep CNN models. From our pilot studies, we observed that the
397   proposed model occupied only 96 MB and took less RAM to do prediction on the test data. The
398   deployed model could serve as triage, minimize delays in disease-endemic/resource-constrained
399   settings.
400

401   **References**

402   Bergstra J., Bengio Y. 2012. Random Search for Hyper-Parameter Optimization. Journal of
403   Machine Learning Research 13:281–305. DOI: 10.1162/153244303322533223.
404   Bibin D., Nair MS., Punitha P. 2017. Malaria Parasite Detection from Peripheral Blood Smear
405   Images   Using   Deep   Belief   Networks.   IEEE   Access   5:9099–9108.   DOI:
406   10.1109/ACCESS.2017.2705642.
407   Botev A., Lever G., Barber D. 2017. Nesterov's accelerated gradient and momentum as
408   approximations to regularised update descent. In: Proceedings of the International Joint
409   Conference on Neural Networks. 1899–1903. DOI: 10.1109/IJCNN.2017.7966082.
410   Bousetouane F., Morris B. 2015. Off-the-shelf CNN features for fine-grained classification of
411   vessels in a maritime environment. In: Lecture Notes in Computer Science (including subseries
412   Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 379–388. DOI:
413   10.1007/978-3-319-27863-6_35.
414   Centers for Disease Control and Prevention. 2012. CDC - Malaria. Available at
415   http://www.cdc.gov/malaria/about/biology/. DOI: 10.1371/journal.pmed.0030473.
416   Chicco D. 2017. Ten quick tips for machine learning in computational biology. BioData Mining
417   10. DOI: 10.1186/s13040-017-0155-3.
418   Chollet F. Deep Learning Models. Available at https://github.com/fchollet/deep-learning-models
419   (accessed February 2, 2017).
420   Chollet F. 2016. Xception: Deep Learning with Separable Convolutions. arXiv preprint
421   arXiv:1610.02357:1–14.
422   Das DK., Ghosh M., Pal M., Maiti AK., Chakraborty C. 2013. Machine learning approach for
423   automated screening of malaria parasite using light microscopic images. Micron 45:97–106. DOI:
424   10.1016/j.micron.2012.11.002.
425   Daya S. 2003. One-way analysis of variance. Evidence-based Obstetrics and Gynecology 5:153–
426   155. DOI: 10.1016/j.ebobgyn.2003.11.001.
427   Deng J., Dong W., Socher R., Li LJ., Li K., Fei-Fei L. 2009. ImageNet: A large-scale hierarchical
428   image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. 248–
429   255. DOI: 10.1109/CVPRW.2009.5206848.

430 Dong Y., Jiang Z., Shen H., David Pan W., Williams LA., Reddy VVB., Benjamin WH., Bryan
431 AW. 2017. Evaluations of deep convolutional neural networks for automatic identification of
432 malaria infected cells. In: 2017 IEEE EMBS International Conference on Biomedical and Health
433 Informatics, BHI 2017. 101–104. DOI: 10.1109/BHI.2017.7897215.

434 Gastwirth JL., Gel YR., Miao W. 2009. The Impact of Levene's Test of Equality of Variances on
435 Statistical Theory and Practice. Statistical Science 24:343–360. DOI: 10.1214/09-STS301.

436 Gopakumar GP., Swetha M., Sai Siva G., Sai Subrahmanyam GRK. 2017. Convolutional neural
437 network-based malaria diagnosis from focus stack of blood smear images acquired using custom-
438 built slide scanner. Journal of Biophotonics:e201700003. DOI: 10.1002/jbio.201700003.

439 Hawkes M., Katsuva J., Masumbuko CK. 2009. Use and limitations of malaria rapid diagnostic
440 testing by community health workers in war-torn Democratic Republic of Congo. Malaria Journal
441 8:308. DOI: 10.1186/1475-2875-8-308.

442 He K., Zhang X., Ren S., Sun J. 2016. Deep Residual Learning for Image Recognition. In: 2016
443 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 770–778. DOI:
444 10.1109/CVPR.2016.90.

445 Hommelsheim CM., Frantzeskakis L., Huang M., Ülker B. 2015. PCR amplification of repetitive
446 DNA: a limitation to genome editing technologies and many other applications. Scientific Reports
447 4:5052. DOI: 10.1038/srep05052.

448 Howard AG., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H.
449 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.
450 ArXiv:9. DOI: arXiv:1704.04861.

451 Huang G., Liu Z., Weinberger KQ., van der Maaten L. 2016. Densely Connected Convolutional
452 Networks. DOI: 10.1109/CVPR.2017.243.

453 Krizhevsky A., Sutskever I., Hinton GE. 2012. ImageNet Classification with Deep Convolutional
454 Neural Networks. In: Advances In Neural Information Processing Systems. 1–9. DOI:
455 http://dx.doi.org/10.1016/j.protcy.2014.09.007.

456 Kucuk U., Eyuboglu M., Kucuk HO., Degirmencioglu G. 2016. Importance of using proper post
457 hoc test with ANOVA. International Journal of Cardiology 209:346. DOI:
458 10.1016/j.ijcard.2015.11.061.

459 LeCun Y., Bengio Y. 1995. Convolutional networks for images, speech, and time series. The
460 handbook of brain theory and neural networks 3361:255–258. DOI:
461 10.1109/IJCNN.2004.1381049.

462 LeCun Y., Bottou L., Bengio Y., Haffner P. 1998. Gradient-based learning applied to document
463 recognition. Proceedings of the IEEE 86:2278–2323. DOI: 10.1109/5.726791.

464 LeCun Y., Yoshua B., Geoffrey H. 2015. Deep learning. Nature 521:436–444. DOI:
465 10.1038/nature14539.

466 Liang Z., Powell A., Ersoy I., Poostchi M., Silamut K., Palaniappan K., Guo P., Hossain MA.,
467 Sameer A., Maude RJ., Huang JX., Jaeger S., Thoma G. 2017. CNN-based image analysis for
468 malaria diagnosis. In: Proceedings - 2016 IEEE International Conference on Bioinformatics and
469 Biomedicine, BIBM 2016. 493–496. DOI: 10.1109/BIBM.2016.7822567.

470   Lipton ZC., Elkan C., Naryanaswamy B. 2014. Optimal thresholding of classifiers to maximize
471   F1 measure. In: Lecture Notes in Computer Science (including subseries Lecture Notes in
472   Artificial Intelligence and Lecture Notes in Bioinformatics). 225–239. DOI: 10.1007/978-3-662-
473   44851-9_15.
474   Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage
475   lysozyme. BBA - Protein Structure 405:442–451. DOI: 10.1016/0005-2795(75)90109-9.
476   Mitiku K., Mengistu G., Gelaw B. 2003. The reliability of blood film examination for malaria at
477   the peripheral health unit. Ethiop.J.Health Dev. 17:197–204.
478   Poostchi M., Silamut K., Maude R., Jaeger S., Thoma G. 2018. Image analysis and machine
479   learning for detecting malaria. Translational Research. DOI: 10.1016/j.trsl.2017.12.004.
480   Razavian AS., Azizpour H., Sullivan J., Carlsson S. 2014. CNN features off-the-shelf: An
481   astounding baseline for recognition. In: IEEE Computer Society Conference on Computer Vision
482   and Pattern Recognition Workshops. 512–519. DOI: 10.1109/CVPRW.2014.131.
483   Rajaraman S., Antani S., Xue Z., Candemir S., Jaeger S. 2017. Visualizing abnormalities in chest
484   radiographs through salient network activations in Deep Learning. In: *Life Sciences Conference*
485   *(LSC), 2017 IEEE*. Sydney, NSW, Australia: IEEE, 71–74. DOI: 10.1109/LSC.2017.8268146.
486   Ross NE., Pritchard CJ., Rubin DM., Dusé AG. 2006. Automated image processing method for
487   the diagnosis and classification of malaria on thin blood smears. Medical & biological engineering
488   & computing 44:427–436. DOI: 10.1007/s11517-006-0044-2.
489   Rossi JS. 1987. One-Way Anova from Summary Statistics. Educational and Psychological
490   Measurement 47:37–38. DOI: 10.1177/0013164487471004.
491   Royston P. 1992. Approximating the Shapiro-Wilk W-test for non-normality. Statistics and
492   Computing 2:117–119. DOI: 10.1007/BF01891203.
493   Schmidhuber J. 2015. Deep Learning in neural networks: An overview. Neural Networks 61:85–
494   117. DOI: 10.1016/j.neunet.2014.09.003.
495   Shang W., Sohn K., Almeida D., Lee H. 2016. Understanding and Improving Convolutional
496   Neural Networks via Concatenated Rectified Linear Units. In: Proceedings of 33rd International
497   Conference on Machine Learning (ICML2016).
498   Simonyan K., Zisserman A. 2015. Very Deep Convolutional Networks for Large-Scale Image
499   Recognition. International Conference on Learning Representations (ICRL). DOI:
500   10.1016/j.infsof.2008.09.005.
501   Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. 2014. Dropout: A Simple
502   Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research
503   15:1929–1958. DOI: 10.1214/12-AOS1000.
504   Stevenson M., Nunes T., Heuer C., Marshall J., Sanchez J., Thornton R., Reiczigel J., Robison-
505   Cox J., Sebastiani P., Solymos P., Yoshida K., Firestone S. 2015. Tools for the Analysis of
506   Epidemiological Data R package version 0.9-62. CRAN.R-project.org.
507   Suzuki K. 2017. Overview of deep learning in medical imaging. *Radiological Physics and*
508   *Technology* 10:257–273. DOI: 10.1007/s12194-017-0406-5.

509   Szegedy C., Liu W., Jia Y., Sermanet P. 2014. Going deeper with convolutions. arXiv preprint
510   arXiv: 1409.4842:1–9. DOI: 10.1109/CVPR.2015.7298594.
511   Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. 2016. Rethinking the Inception
512   Architecture for Computer Vision. Proceedings of the IEEE Computer Society Conference on
513   Computer Vision and Pattern Recognition (CVPR):2818–2826. DOI: 10.1002/2014GB005021.
514   Vargha A., Delaney HD., Vargha A. 1998. The Kruskal-Wallis Test and Stochastic Homogeneity.
515   Journal of Educational and Behavioral Statistics 23:170. DOI: 10.2307/1165320.
516   WHO.   2016.   World   Malaria   Report.   Available   at
517   http://apps.who.int/iris/bitstream/10665/252038/1/9789241511711-eng.pdf?ua=1   (accessed
518   January 4, 2017). DOI: 10.4135/9781452276151.n221.
519   Yu F. 2016.Fine-tune CNN in Keras. Available at https://github.com/flyyufelix/cnn_finetune
520   (accessed October 2, 2017).
521   Zeiler MD., Fergus R. 2014. Visualizing and understanding convolutional networks. In: Lecture
522   Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and
523   Lecture Notes in Bioinformatics). 818–833. DOI: 10.1007/978-3-319-10590-1_53.

Figure 1

Architecture of the customized model.

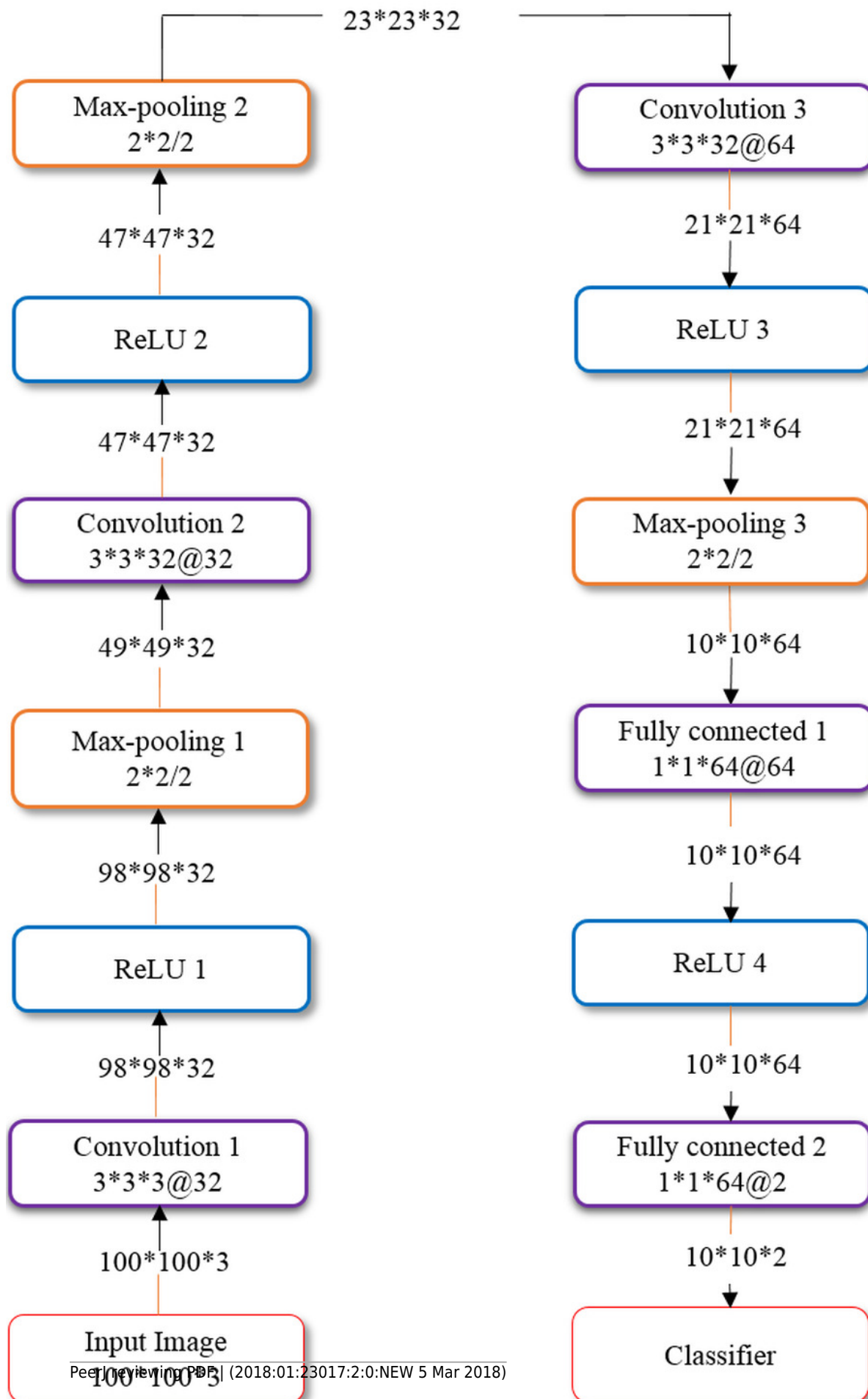
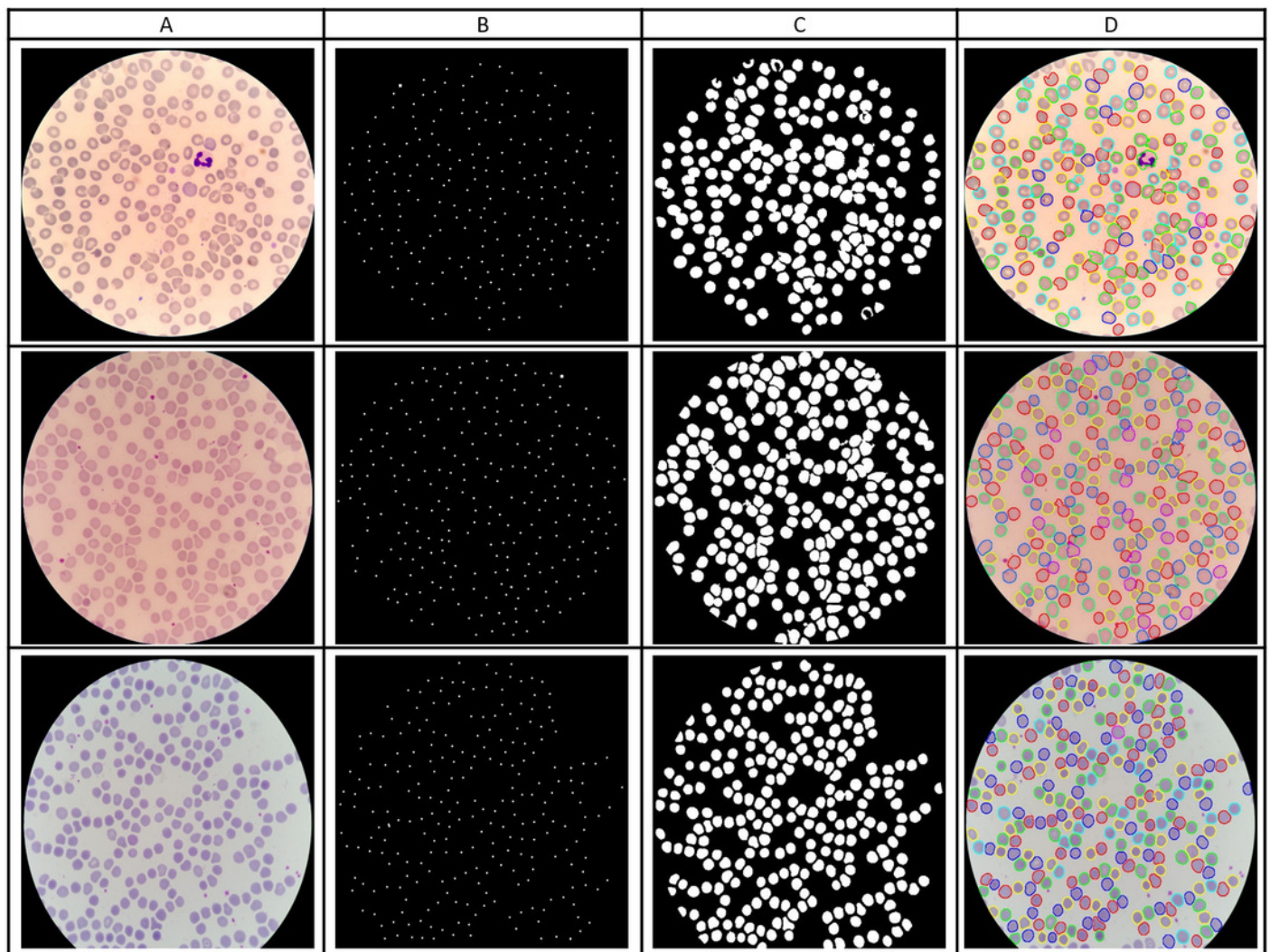



Figure 2

RBC detection and segmentation using iterative voting and level sets.

(A) Input image. (B) Initial cell detection using iterative voting. (C) Final RBC segmentation mask. (D) Segmentation results superimposed on the original image.

Data for cross-validation studies.

1 **Table 1:**

2 **Data for cross-validation studies.**

Folds	Parasitized	Uninfected
1	2756	2757
2	2758	2758
3	2776	2762
4	2832	2760
5	2657	2742
Total	**13779**	**13779**

Performance metrics.

1 **Table 2:**

2 **Performance metrics.**

Models	Accuracy	AUC	Sensitivity	Specificity	F1-score	MCC
AlexNet	0.937±0.012	0.981±0.007	0.940±0.017	0.933±0.034	0.937±0.011	0.872±0.024
VGG-16	0.945±0.015	0.981±0.007	0.939±0.022	0.951±0.019	0.945±0.016	0.887±0.030
ResNet-50	**0.957±0.007**	**0.990±0.004**	**0.945±0.020**	**0.969±0.009**	**0.957±0.008**	**0.912±0.014**
Xception	0.890±0.107	0.948±0.062	0.931±0.039	0.835±0.218	0.895±0.100	0.772±0.233
DenseNet-121	0.931±0.018	0.976±0.023	0.942±0.023	0.926±0.032	0.931±0.017	0.894±0.036
Customized	0.940±0.010	0.979±0.009	0.931±0.026	0.951±0.030	0.941±0.010	0.880±0.020

Candidate layers giving the best performance.

1 **Table 3:**

2 **Candidate layers giving the best performance.**

Model	Optimal layer
AlexNet	fc6
VGG-16	block5_conv2
ResNet-50	res5c_branch2c
Xception	block14_sepconv1
DenseNet-121	Conv5_16_x2

Performance metrics achieved with feature extraction from optimal layers.

1 **Table 4:**

2 **Performance metrics achieved with feature extraction from optimal layers.**

Models	Accuracy	AUC	Sensitivity	Specificity	F1-score	MCC
AlexNet	0.944±0.010	0.983±0.006	0.947±0.016	0.941±0.025	0.944±0.010	0.886±0.020
VGG-16	**0.959±0.009**	**0.991±0.004**	0.949±0.020	0.969±0.016	**0.959±0.009**	0.916±0.017
ResNet-50	**0.959±0.008**	**0.991±0.005**	0.947±0.015	**0.972±0.010**	**0.959±0.009**	**0.917±0.017**
Xception	0.915±0.005	0.965±0.019	0.925±0.039	0.907±0.120	0.918±0.042	0.836±0.088
DenseNet-121	0.952±0.022	**0.991±0.004**	**0.960±0.009**	0.944±0.048	0.953±0.020	0.902±0.041
Customized	0.927±0.026	0.978±0.012	0.905±0.074	0.951±0.031	0.928±0.041	0.884±0.002

Consolidated results of Kruskal-Wallis H and post-hoc tests.

1 **Table 5:**

2 **Consolidated results of Kruskal-Wallis H and post-hoc tests.**

Metric	Kruskal-Wallis H summary	Mean ranks		Post-hoc
Accuracy	$\chi^2(5) = 15.508$, $p = 0.008$	AlexNet	11.20	Xception & ResNet-50 ($p = 0.005$)
		VGG-16	22.30	Xception & VGG-16 ($p = 0.007$)
		ResNet-50	**23.00**	
		Xception	7.20	
		DenseNet-121	19.60	Customized & ResNet-50 ($p = 0.017$)
		Customized	9.70	
AUC	$\chi^2(5) = 18.958$, $p = 0.002$	AlexNet	13.00	Xception & ResNet-50 ($p = 0.034$)
		VGG-16	21.70	
		ResNet-50	21.50	Xception & VGG-16 ($p = 0.030$)
		Xception	4.50	
		DenseNet-121	**22.90**	Xception & DenseNet-121 ($p = 0.014$)
		Customized	9.40	
Sensitivity	$\chi^2(5) = 5.518$, $p = 0.356$	AlexNet	16.20	-
		VGG-16	17.30	
		ResNet-50	15.80	
		Xception	11.40	
		DenseNet-121	**21.80**	
		Customized	10.50	
Specificity	$\chi^2(5) = 6.639$, $p = 0.249$	AlexNet	9.80	-
		VGG-16	20.70	
		ResNet-50	**21.30**	
		Xception	13.30	
		DenseNet-121	14.10	
		Customized	13.80	
F1-score	$\chi^2(5) = 14.798$, $p = 0.011$	AlexNet	11.70	Xception & ResNet-50 ($p = 0.005$)
		VGG-16	22.20	
		ResNet-50	**22.60**	Xception & VGG-16 ($p = 0.006$)
		Xception	6.90	
		DenseNet-121	19.50	Xception & DenseNet-121 ($p = 0.023$)
		Customized	10.10	
MCC	$\chi^2(5) = 14.487$, $p = 0.013$	AlexNet	11.30	Xception & ResNet-50 ($p = 0.007$)
		VGG-16	22.30	
		ResNet-50	**22.60**	Xception & VGG-16 ($p =$

		Xception	7.60	0.008)
		DenseNet-121	19.40	Xception & DenseNet-121 (p
		Customized	9.80	= 0.034)
				Customized & ResNet-50 (p
				= 0.021

Comparison with the state-of-the-art literature.

1 **Table 6:**

2 **Comparison with the state-of-the-art literature.**

Method	Accuracy	Sensitivity	Specificity	AUC	F1-score	MCC
Proposed Model (cell Level)	**0.986**	**0.981**	**0.992**	**0.999**	**0.987**	**0.972**
Proposed Model (patient level)	0.959	0.947	0.972	0.991	0.959	0.917
Gopakumar *et al.* (Gopakumar et al., 2017)	0.977	0.971	0.985	-	-	0.731
Bibin *et al.* (Bibin, Nair & Punitha, 2017)	0.963	0.976	0.959	-		
Dong *et al.* (Dong et al., 2017)	0.981	-	-	-		
Liang *et al.* (Liang et al., 2017)	0.973	0.969	0.977	-		
Das *et al.* (Das et al., 2013)	0.840	**0.981**	0.689	-		
Ross *et al.* (Ross et al., 2006)	0.730	0.850	-	-		

5
6
7
8